ISSN 2395-1621

Android based Optical Character Recognition System

[#1]Divya Maheshkumar, [#2]Adhav Pratik, [#3]Ghodechor Sanket

[1]divya.mk1094@gmail.com
[2]pratikadhav231@gmail.com
[3]ghodechorsanket@gmail.com

[#1234]Department of Computer Engineering,
All India Shree Shivaji Memorial Society's College Of Engineering,
Savitribai Phule Pune University, Maharashtra, India.

ABSTRACT

In today's world, all aspects of our lives have a digital presence. With the prominent and daily use and integration of technology to a student's daily routine, utilization of their mobile gadgets for educative purposes can be advantageous both to students and teachers alike. Android O.C.R is an application which uses this fact to attempt to digitize the activity of making notes. This is an Optical character recognition system which converts handwritten scripts on a touch screen device into text format, which can be stored for later retrieval or editing using any text-editor software. Input for the system is an image; the OCR gets text from that image and then converts it into speech. This system can be useful in various applications like banking, legal industry, other industries, home and office automation. It is mainly designed for partially blind people. The proposed OCR system provides many features that require no typing, editing raw data, quick translation, and memory utilization. In the end it also highlights the major emerging trends in the field of OCR and how OCR as a technology is evolving with every passing day.

ARTICLE INFO

Article History

Received :5th February, 2016

Received in revised form :

7th February, 2016

Accepted : 9th February, 2016

Published online :

16th February, 2016

Keywords:Android application, Optical Character Recognition, Template Generation, Template matching.

I. INTRODUCTION

More people than ever before people are using personal computers, laptop, tablets, and e-readers to read books and documents. This means that old print media must be scanned and converted to a digital format in order to be accessed from these devices. Optical Character Recognition (OCR) systems are used to read scanned images and convert them into a digital character-based format. The first major use of OCR was in processing petroleum credit card sales drafts. The early devices were combined with punch units which made small holes that could be read by the computer. As computers and OCR devices became more sophisticated, direct access was provided into the CPU by scanners. This quickly led to the payment processing of credit card purchases, known as "remittance processing". Traditional OCR systems require large dataset and complex processes, thus increasing the computation time, increasing the storage space and thus in turn the performance of the system was affected.

Android is a software stack for mobile devices that includes an operating system, middleware and key applications. Android is a software platform and operating system for mobile devices based on the Linux operating system which was developed by Google and the Open Handset Alliance. Thus, a mobile OCR will pave way to common usage of the system wherein the public can utilize the features of an OCR system.

II. LITERATURE SURVEY

The authors of [1], proposed a system language-independent optical character recognition (OCR) system that is capable, in principle, of recognizing printed text from most of the world's languages. For each new language or script the system requires sample training data along with ground truth at the text-line level; there is no need to specify the location of either the lines or the words and characters. The system uses hidden Markov modeling (HMM) technology to

model each character. In addition to language independence, the technology enhances performance for degraded data, such as fax, by using unsupervised adaptation techniques. Thus far, we have demonstrated the language-independence of this approach for Arabic, English, and Chinese. Recognition results are presented in this paper, including results on faxed data aspects, including language-independent training and recognition methodology; automatic training on non-segmented data; and simultaneous segmentation and recognition. Our approach is different from other OCR approaches in three ways; firstly this approach is based on language-independent recognition. For making an OCR language independent it was made script-independent in terms of feature extraction, training, and recognition. Secondly, training and recognition are performed by doing preprocessing and feature extraction. Thirdly, this system doesn't pre-segment at the character or at the word level, in contrast with other OCR systems. This OCR system works in two phases training system and the recognition system. Requirement of the system is to find character models, lexicon and grammar, from training data. We derive lexicon from data which may be text corpus. The character-modeling makes use of the feature vectors and calculate the corresponding character models. The training phase uses orthographic rules that depend on the type of script. Orthographic rules tell whether the text lines go horizontally or vertically for e.g. in traditional Chinese. If the text is read vertically, is it read from left-to-right for e.g. in Roman script, or right-to-left for e.g. in Arabic script.

The Improved Offline Connected Script Recognition Based on Hybrid Strategy[2], presented hybrid strategy for recognition of strings of characters, a project that aims at recognizing cursive handwritten words and numeral strings. Previously there are two main approaches: explicit segmentation based and implicit segmentation based. However, both approaches have their own shortcomings. To overcome individual weaknesses, this paper presents a hybrid strategy for recognition of strings of characters. This system works in two steps: first an explicit segmentation is applied to segment either cursive handwritten words or numeric strings. However, at this stage, segmentation points are not finalized.While second step is verification stage in which statistical features are extracted from each segmented area to recognize characters using a trained neural network.
The paper [3] explains how an Optical Character Recognition system (OCR) works and how this system enables us in capturing an image of a text document. It also explains how OCR is more efficient and easier alternative to scanning a document using a scanner as the image captured using OCR is of exactly the same quality like its scanned copy, the only difference being that OCR is done with the help of a simple mobile phone camera whereas scanning is done using a bulky scanner. To perform the character recognition, this application has to go through two important steps which are as follows:-

1. Segmentation, i.e., given a binary input image, to identify the individual glyphs.
2. Feature extraction, i.e., to compute from each glyph a vector of numbers that will serve as input features for an ANN.

III.PROPOSED ARCHITECTURE

OCR is the acronym for Optical Character Recognition. Optical Character Recognition (OCR) programs are used to read scanned images and convert them into a digital character-based format. This technology allows a machine to automatically recognize characters through an optical mechanism:

- When first time any user uses the system, user will have to register.
- Afterwards, when system starts user needs to log in to the system using username and password provided at the time of registration. After verifying credentials entered by user system allows user to select images for optical character recognition.
- At this stage, preprocessing operations are applied on an input image that includes RGB separation, gray scale, threshold, blurring, thinning.
- RGB separation is nothing but separation of red, green and blue color of each pixel by using AND operation and shift operation. RGB separated image is then given as input to gray scale processing stage where each bit of the pixel is filled with the gray scale value obtained by taking average of RGB separated values. Here we are going to include two types of threshold methods. Blurring is also called as smoothing, which smooth's out portions of image or symbols. In short blurring averages out rapid changes in intensity. After blurring, thinning is applied which in turn removes unusable portions of image from that of the blurred image. All symbol data obtained by preprocessing phase is used by template extraction phase. In template extraction process, scaling is applied on obtained symbols. Here we are going to use paper salted technique, in which if given symbol has dimension like 100x80 matrix then divided into 2x2 matrix recursively, until we get 8x8 matrix. If such a 2x2 matrix does not contain text pixel then remove that matrix.
- In training phase we provide this extracted template to the storage for further use in template matching function. Template matching is based on score calculation. Template matching system compares generated template with existing templates pixel by pixel then gives matched score of each symbol. Symbol having maximum score is the symbol output by the system.

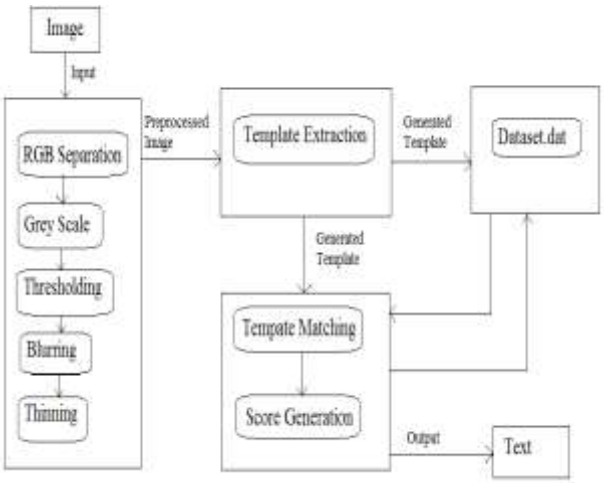

Fig 1: Proposed system architecture

IV. MATHEMATICAL MODEL

Let's S be system
S = {U, Ss, F, Q, D}
Where,
U is finite set of users
U = {U1, U2, U3…….UN}
N is finite number of users.

Ss is success state
Ss (Success)=
1. OCR system good enough to do segmentation of text.
2. All the characters from digital images are successfully recognised by the system.

F(Failure) =
1. OCR fails to recognize cursive characters.
2. Image contains noise.

D is the set of data used by system.
Q is the set of states
Q = {q0, q1, q2, q3, q4, q5, q6, q7, q8}
Where,
q0 - User log in into the system.

q1 - Input Image is given to the system.

q2 - Preprocessing applied on input image.

q3 - Template Generation.

q4 - Generated template storing into database.

q5 - Template Matching.

q6 - Character Recognition.

Q7 - Final output text that user wants to edit.

Q8 - User logout from system.

V. FEASIBILITY

1. System is designed for education purpose so it is economically feasible.
2. Also complexity of our system is O(n), where n is number of pixels in image. As space and time complexity of system is able to find.

VI. IMPLEMENTATION CONSTRAINS

1. If input characters are cursive characters then system is unable to recognize those characters or may provide unexpected result.
2. Also input image can sometime contain large amount of noise thus affecting further processing.

VII. APPLICATIONS

1. Visually impaired persons are unable to recognize characters, in such cases Android OCR helps a lot in identification of characters.
2. Android OCR also useful in language translation. Language translation is the process of transferring language from one form to other. Instead of manually translating language, which is very costly and time consuming, OCR helps in language translation.
3. Android OCR plays role in applications of digital conversion.

VIII. CONCLUSION

We are developing a mobile application named Android OCR for smartphones. We aim to provide a mobile application for recognizing characters for not only commercial but also personal use. We aim to provide a system which can recognize and convert handwritten text into a format which can be editable even on any basic text editor software.

REFRENCES

[1] A Robust, Language-Independent OCR System , Zhidong Lu, Issam Bazzi, Andras Kornai, John Makhoul,Premkumar Natarajan, and Richard Schwartz
BBN Technologies, GTE Internetworking, Cambridge, MA 02138 2015.

[2] Improved Offline Connected Script Recognition Based on Hybrid Strategy Ghazali Sulong , Amjad Rehman and Tanzila Saba University of Technology Malaysia (UTM) Skudai, Malaysia.

[3] Optical Character Recognition on Handheld Devices Sravan Ch, ShivankuMahna, NirbhayKashyap
On 22, April 2015.

