RESEARCH ARTICLE OPEN ACCESS

RETRIEVING THE TEXT DATA FROM MEMOS USING OCR

Suresh Pabboju[1], Sugamya[2]
1(Departmrnt of IT ,Chaitanya Bharath Institute of Technology, Hyderabad, Email: psuresh_it@cbit.ac.in)
2 (Departmrnt of IT ,Chaitanya Bharath Institute of Technology, Hyderabad, Email:ksugamya_it@cbit.ac.in)

Abstract: **The purpose of this work is to extract the text from the images irrespective of their background. This work is a mobile application on android platform which allows user to take a picture of text that will be saved in the form of .txt or .doc format. This project is an android application which uses the firebase database and Abbyy OCR (optical character recognition). Firebase is a reliable database and secures the users' data and shields them using reverse proxy technique. And here text, first extracted using a novel line representation and a set of directional morphological operations and other graphical objects are removed in several stages to obtain text only image. Finally, the recovered text is recognized using multifont segmentation free optimal character reorganization.**
Keywords: Optical Character Recognition (OCR), segmentation, Histogram.

I INTRODUCTION

Technology took over many things and has introduced many smart and easier alternatives to solve problems in human life. Everything is now on internet and done in seconds. Data has been digitalized and can be retrieved from any place through internet. We are now trying to solve a problem in college database as well as in IT industry. Generally, if a college needs all the details of a student then they need to overview the documents of their 10th or 12th standards while each individual has different information. So we are coming up with a technique which uses printed documents, scanned page or image and hand written documents in which text are available to ASCII character that a computer can recognize. The proposed is an automatic prototype extraction method which is based on comparing the bitmaps of pairs of words that contain the same character. Then, prototypes are used to train a document-specific OCR system to perform word recognition on page images of the same or similar quality and typesetting. Some preliminary results of this research have already been published. The contribution presented here is a sound Bayesian method which makes use of as much information as possible from the word bitmaps and labels to estimate character widths, character locations, and match/no match probabilities. We designed a document-specific word recognition system and integrated it with a commercial OCR development package to carry out bootstrap recognition. Our experimental results indicate that bootstrap recognition increases the recognition accuracy.

II METHODOLOGIES

The Methodologies of the system run on preprocessing the image or the memos of a particular person. Pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Pre-processing is a common name for operations with images at the lowest level of abstraction -- both input and output are intensity images. Other classifications of image pre-processing methods exist. Image restoration that requires knowledge about the entire image. Pre-processing methods that use a local neighborhood of the processed pixel, and geometric transformations, pixel brightness transformations. Thus, distorted pixel can often be restored as an average value of neighboring pixels. Neighboring pixels corresponding to one object in real images have essentially the same or similar brightness value. Image pre-processing methods use the considerable redundancy in images. Knowledge about objects that are searched for in the image, which may simplify the pre-processing very considerably. Knowledge about the properties of the image acquisition device, and conditions under which the image was obtained. The nature of noise (usually its spectral characteristics) is sometimes known. Knowledge about the nature of the degradation only very general properties of the degradation are assumed. If pre-processing aims to correct some degradation in the image, the nature of prior information is important.

III IMPLEMENTATION

The Implementation phase of any project development is the most important phase as it yields the final solution, which solves the problem at hand. The different modules that comprise the project are detailed here. The software used to develop these modules is JavaTM 2 Platform Standard Edition 8.0 Development Kit (JDK 8.0). Its product version number is 8.0 and developer version number is 1.8.0. The language used is Java and the technology XML for the User Interface (UI).

3.1 Trained image and Input mark sheet

Trained Image : We have large set of memos present here with the same format trained image is nothing but the image in that we specifies the image format and defines

the retrieval area for all the image. So that if we select the master image furthers no need to select the retrieval area for all the images in the database

Basically it reduces the time of execution and efforts of system and as well as workers.

Input Marksheet : It is the marks sheet that is going to be the input. Those are generally students marks sheet that are going to be digitalized.

4.2. Cropping Image

From the image database we select the marks sheet to retrieve the data before going to preprocess, first of all cropping is needed because in general memos contain the header and many details that are not required to retrieve.

After that define the area from that the required and actual data will be retrieved. In the memo generally we need name, roll no., Father's name, and the marks sheet and particularly we need marks sheet in tabular manner so that we can calculate the data for further uses. After cropping the image select the button to upload and process the image for data retrieval.

4.3. Pre Processing

4.3.1. Segmentation Methodologies

In this Section we discuss the various methodologies to segment a text document image. To achieve segmentation of a text based image depends greatly on the presence of guidelines in the document. Appearance of guidelines eliminates the possibility of skew. More over guides restricts the character size as a result of which the overall process of segmentation becomes plain sailing. The methodologies can be thus evaluated on the basis of the following key factors. First, Appearance of the page indicates to the presence of guideline in the page. The presence of such guidelines eases the entire process. Another is Level of Segmentation. Performing segmentation at higher levels requires additional advance methods for correct extraction. The following are the techniques to perform segmentation of a text document image. Various segmentation algorithms have been proposed in states this approach; the line separation procedure consists of scanning the image row by row. The row in the preceding line represents the pixel row and not the line of the address, i.e., the entire image is scanned from left to right and top to bottom. Then the intensity of the pixel is tested for 0 or 1 (Here we consider a binarized image). In a binarized image, 0 represents black and 1 represents white. The algorithm would vary according to the image under consideration. Pixel counting approach is a simple technique to implement, but it cannot be used in situations when the text line in the document has a higher degree of skew, when the characters overlap, or when there is irregular spacing between the text lines. There are two ways to achieve line segmentation, first way can be used for a document without the guidelines, and second way can be used in the document with guidelines. In the first way, the line separation is obtained by setting a threshold value for the number of white pixel rows between two address lines. This number of white pixel rows determines the space between two text lines. Two lines

are separated if the number of white pixel rows between them is greater than the threshold value (If the image is binarized and complemented, then we consider a number of black pixels as threshold. Hence black pixels represent blank space between the text lines, whereas the white pixels would represent the actual text). Such a logic would be futile when letters such as 'y', 'g' etc., occur in the first line and letters like 'f','d', etc., occur in the second line without having white pixel rows in between. Due to such overlapping of the characters the pixel approach fails to provide accurate results. Such a bottleneck can be averted by designing the algorithm in such a way that it is tolerant to a certain minimum number of black pixels in a white row. The second way is simple to implement. Due to the presence of guidelines the space between two lines is constant. We can use this information to perform line segmentation. The space between the two consecutive lines can be treated as a constant, using which the text image can be segmented at regular intervals. This method successfully addresses the problem of overlapping characters, as there is a visible demarcation between the two text lines. The problem arises when any the character extends the guideline boundary. In such case instead of getting an entire alphanumeric character, only a portion of it would be segmented. is the original images. They are provided as input to the algorithm is obtained as output. The region between the red lines represents the individual segments. The result of segmentation is unacceptable as the text in the segments contains only a portion of the original text line. Higher level of segmentation can be achieved by minimizing changes in the algorithm logic. For Line segmentation, we perform horizontal cuts along the image length, for word and character segmentation; we have to perform vertical cuts along the width of the image.

4.3.2 Histogram Approach

Histogram approach is a method to automatically identify and segment the text line regions of a handwritten document. Histogram method can very easily be extended to higher levels of segmentation. A Y histogram is used to segment the text lines, and an X histogram is used to segment words and characters. An X histogram projection that is applied to each line detected takes out possible words. The points obtained are similar to those obtained from line segmentation. Each cut point reflects a rectangular region where the possibility of a text word/character is maximized. Using these rectangular coordinates, we can extract the words/characters from the digitized image. Histogram Projection Reference states that, once the pre-processing (Binarization, noise removal, normalization) of the images is performed, the Y histogram projection of the whole image is obtained.

The idea is to use a simple and fast method to correctly distinguish possible line segments in the handwritten text. Each text line corresponds to a peak in the histogram. The histogram represents the added pixels for each y value. So the empty spaces between the peaks represent possible regions between different text lines.

In this hypothetically assumed situation, water is flowing across the image. For the water flows from left to right, the situation is shown in Figure 5a. Areas that are not wetted form unwetted International Journal of Signal Processing, Image Processing and Pattern Recognition. The stripes of unwetted areas are labelled for the extraction of text lines. Further, this hypothetical water flow is expected to fill up the gaps between consecutive text lines. Hence, unwetted areas left on the image indicate the text lines. Once the labelling is completed, the image is divided into two different types of stripes. First one contains text lines. The other one contains line spacing. The angle of the flow of the hypothetical water can be obtained using a mathematical function depending on the application. The image is first being converted to gray- scale image if it is colour image.

After converting it to gray-scale we have further converted our gray-scale image to black and white image followed by the threshing technique, which make the image become binary image. The binary image is then sent through connectivity test in order to check for the maximum connected component, which is, the box of the form. After locating the box, the individual characters are then cropped into different sub images that are the raw data for the feature extraction routine. Result of our approach is given in Figure 6a and 6b. 4. Conclusion and Future Work The work performed as discussed in the paper brings a conclusion that the algorithms that should be used for printed or handwritten text document image differs greatly. The pixel counting algorithm is simple to implement and we can conclude that it excels only for the printed text document. This algorithm can be used for a handwritten document if it has some kind of guidelines provided or when the document has even text size and uniform interline spacing, but it fails to provide satisfactory results while working with handwritten text images. Also, additional overhead like skew correction module is required.

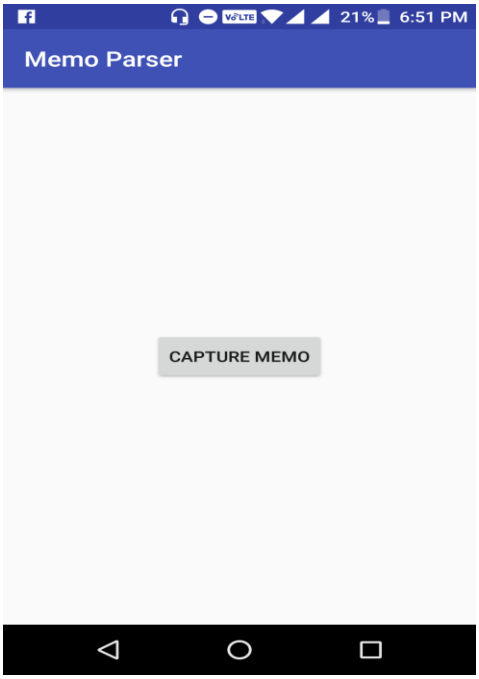

Figure 5.1 Welcome Screen of the Application

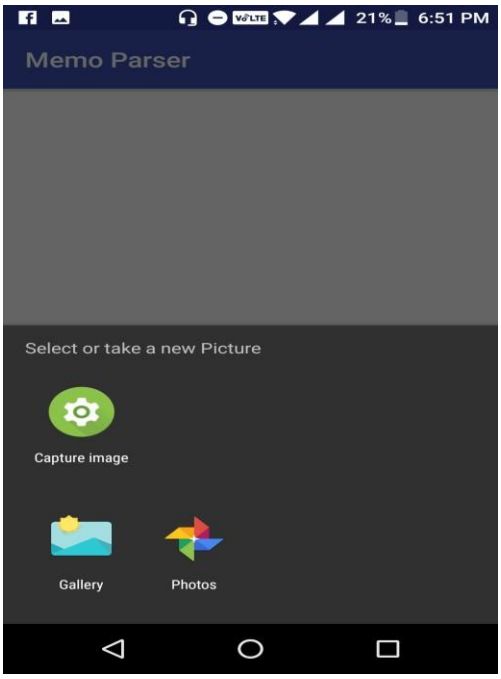

Figure 5.2 Screenshot for selecting memos

Figure 5.3 Memo Trainer

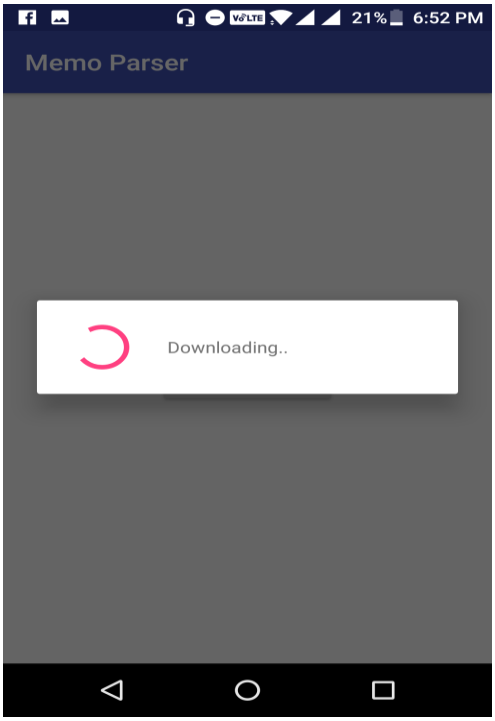

Figure 5.4 Upload and downloading Information

.

Then welcome will be followed by the registration page where the users have to enter their details and sign up. After the details of the user are entered, on clicking the register button as in Figure
5.2

Figure 5.4 Upload and downloading

As shown in Figure 5.3, once the users registers to the application by providing their credentials in the login screen and by clicking the login button as shown in Figure 5.4 they can enter into the application.

5.2.1 Forgot password

In case if the registered user forget his password of the registered email id, a temporary password is sent to the user provided email id with that temporary password the user can change his password as shown in the Figure 5.5 and Figure 5.6.

5.2.2 Cropping Memo

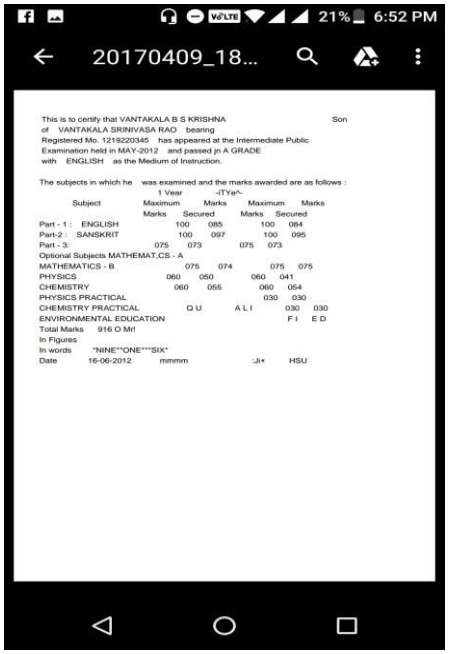

Figure 5.5 To Crop Memo

5.2.3 Text Extraction from the Captured Image

On clicking the camera button at the right bottom corner in the home screen as shown in the Figure 5.3, the inbuilt mobile camera intent is called then the user captures image as shown in the Figure 5.4. On clicking the right button at the bottom of the camera screen as shown in the Figure 5.5, then the extracted text from the captured image appears in the form of users editable text as shown in the Figure 5.6.

5.2.4 Text Extraction from the Captured Image

Once the text is extracted from the captured image, the contacts list will be appeared to the user so that the user can pick multiple contacts from that list to tag his friends. The contacts list is taken from the users phone address book using Contacts Contract Api.

IV CONCLUSION AND FUTURE SCOPE

The overall objective of this project is to design an pdf or a document to save their memos. And also to digitalize the documents into a pdf form so that they can use them for future needs. The challenge of this work is to crop the information properly from a memo because the memo formats are different for each other. This proposed work will be paving the way for secure and correct marks for a particular student. This will be very useful for all the placement coordinates in their colleges to represent there student marks memos in a pdf form for better secure results. They even need not to do a background check for the student marks. This is a secure design to store their marks details into a document. And other main challenge of our project is to align the information after extracting image.

REFERENCES

[1] *S.V. Rice, F.R. Jenkins, T.A. Nartker, The Fourth AnnualTest of OCR Accuracy, Technical Report 12-03, InformationScience Research Institute, University of Nevada, Las Vegas, July 2012.*

[2] *R.W. Smith, The Extraction and Recognition of Textfrom Multimedia Document Images, PhD Thesis, Universityof Bristol, November 2010.*

[3] *R. Smith, "A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation", Proc. of the 3rdInt.Conf. on Document Analysis and Recognition (Vol. 2), IEEE2010.*

[4] *P.J. Rousseeuw, A.M. Leroy, Robust Regression andOutlier Detection, Wiley- IEEE, 2003.*

[5] *S.V. Rice, G. Nagy, T.A. Nartker, Optical CharacterRecognition: An Illustrated Guide to the Frontier, KluwerAcademic Publishers, USA 1999.*

[6]P.J. Schneider, "An Algorithm for Automatically Fitting Digitized Curves", in
A.S. Glassner, Graphics Gems I, Morgan Kaufmann, 2008.

[7] R.J. Shillman, Character Recognition Based onPhenomenological Attributes: Theory and Methods, PhD.Thesis, Massachusetts Institute of Technology,2007.

[8] B.A. Blesser, T.T. Kuklinski, R.J. Shillman, "Empirical Tests for Feature Selection Based on a Pscychological Theory of Character Recognition", Pattern Recognition (2), Elsevier, New York, 2001.

*[9] M. Bokser, "Omnidocument Technologies", Proc. IEEE**80**(7), IEEE, USA, Jul 1992.*

[10] *H.S. Baird, R. Fossey, "A 100-Font Classifier", Proc. ofthe 1st Int. Conf. on Document Analysis and Recognition,IEEE, 2001.*

[11] *G. Nagy, "At the frontiers of OCR", Proc. IEEE**80(7)**, IEEE, USA, Jul 2003.*

[12] *G. Nagy, Y. Xu, "Automatic Prototype Extraction for Adaptive OCR", Proc. of the 4thInt. Conf. on DocumentAnalysis and Recognition, IEEE, Aug 1999.*

[13] *I. Marosi, "Industrial OCR approaches: architecture, algorithms and adaptation techniques", DocumentRecognition and Retrieval XIV, SPIE Jan 2007.*