HANDWRITTEN CHARACTER RECOGNITION TO OBTAIN EDITABLE TEXT

D.SARITHA, Assistant Professor , Dept.of Master of Computer Applications, Narayana Engineering College(Autonomous), Gudur.SPSR Nellore, AP, India
P.SWATHI, Research Scholar, Dept.of Master of Computer Applications, Narayana Engineering College(Autonomous), Gudur. SPSR Nellore, AP, India

Abstract

Developing an android application for character recognition to read the text from an image is a big area of research. Nowadays, there is a trend of storing information from the handwritten documents for future use. A simple way to store the information is image capturing of the handwritten document and save it in image format. The method to transform handwritten data into electronic format is 'Optical Character Recognition'. It involves several steps including pre-processing, segmentation, feature extraction and post-processing. Many researchers have been used OCR for recognizing character. This system uses the android phone to capture the image of the document and further steps are done by OCR. The main challenge is to recognize the characters from different styles of handwriting. Thus, a system is designed that recognizes the handwritten data to obtain an editable text.

Keywords:Scan Image,Scan Handwritten Data

Introduction:

The OCR engine process on the image data and convert it into a text. The technology behind the OCR is to first scan the document or image using a scanner. Once the image is scanned OCR software convert image into black and white version. Then the image is analyzed by dark and light areas. Where dark area identified as character and light area identified as background and the dark area is considered for further process.

Statement of the Problem:

Nowadays demand increases to create a paperless environment. Recognition of handwritten text is easy for a human but it is a complex task for computer systems. Many researchers have done work on this field but 100% accuracy is not achieved by the researchers. Our eyes can figure out the handwritten character of different people but the computer cannot do this easily

Objectives of the study

The method to transform handwritten data into electronic format is 'Optical Character Recognition'. It involves several steps including pre-processing, segmentation, feature extraction and post-processing. Many researchers have been used OCR for recognizing character. This system uses the android phone to capture the image of the document and further steps are done by OCR. The main challenge is to recognize the characters from different styles of handwriting.

Review of Literature

OCR accuracy can be increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. This technique can be problematic if the document contains words not in the lexicon, like proper nouns. Tesseract uses its dictionary to influence the character segmentation step, for improved accuracy.

The output stream may be a plain text stream or file of characters, but more sophisticated OCR systems can preserve the original layout of the page and produce, for example, an annotated PDF that includes both the original image of the page and a searchable textual representation.

"Near-neighbor analysis" can make use of co-occurrence frequencies to correct errors, by noting that certain words are often seen together. For example, "Washington, D.C." is generally far more common in English than "Washington DOC".

Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy.

The Levenshtein Distance algorithm has also been used in OCR post-processing to further optimize results from an OCR API.

Research Methodology:

OCR technology allows the conversion of a captured image into machine editable[5]. OCR technology do the work in three stages first is scanning of the document. Second is recognition of character and third is storing the text in the desired format. Optical character recognition is the electronic conversion of scanned image or printed text into machine editable text.

Open-source OCR-Tesseract software is used for our work. Tesseract is considered as one of the most accurate opensource engines. The steps followed by Optical character recognition are shown in the algorithm as follows,

- Image acquisition by the android camera
- Loading the image into created Graphical User Interface (GUI) in android studio.
- Preprocessing of the image.
- Extraction of features from the input image.
- Recognized data converted into text format using OCR.

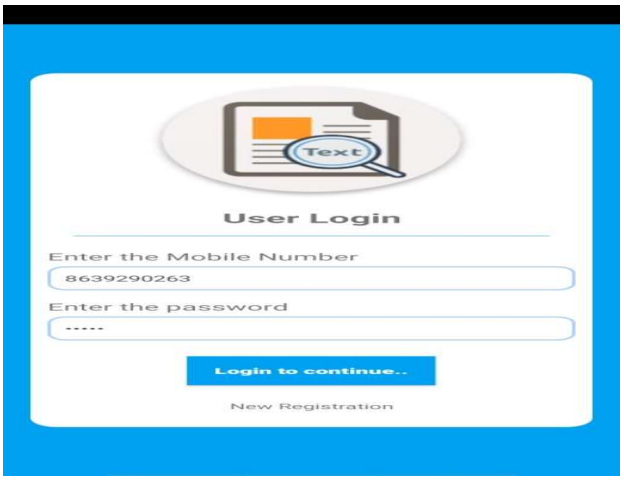
Results and Discussion

The result analysis describes that the entire project was executed successfully and also having quality and performance by analyzing the flow of data and output screens. In my project the modules like User, Scan Image Data, Scan Handwritten Data are independent modules. Because my project follows the top down approach and bottom up approach.

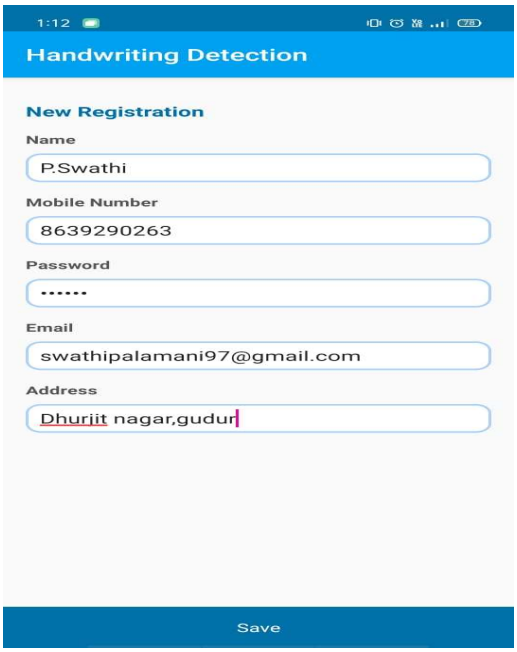
TC No	Test Case	Input	Expected Output	Observed Output	Result
TC1	Login	Enter mobile no and password	Login Successful	--do--	Pass
TC2	Login	Enter Wrong mobile no and password	Invalid login details	--do--	Pass
TC3	Registration	Enter all Fields Data	Registration Successful	--do--	Pass
TC4	Registration	Enter some fields data	All fields are mandatory	--do--	Pass

Test Cases Table

Login Page:



Registration Page:



Conclusion

This implementation gives conversion of Handwritten Character Recognition into editable text using the android app. The image is captured by the camera and loaded into the android app and choice is provided to the user to select a part of an image which is to be converted. Further processing is done by OCR engine and produces the converted text on the screen. The recognized text is saved in text format. To edit the recognized text a choice is given and save them in a proper location. More accuracy is achieved when text is in printed form rather than in handwritten.

References:

[1] https://www.tutorialspoint.com/android/android_resources.htm
[2] https://developer.android.com/guide/index.html
[3] Manwatkar, Pratik & Singh, K.. (2015). A Technical Review on Text Recognit ion from Images. 10.1109/ISCO.2015.7282362.
[4] Sharma, Ankit & Chaudhary, Dipt i & Ahemedabad, & India,. (2013). Character Recognit ion Using Neural Network.
[5] Robby, G. & Tandra, Antonia & Susanto, Imelda & Harefa, Jeklin & Chowanda, Andry. (2019). Implementat ion of Opt ical Character Recognit ion using Tesseract with the Javanese Script Target in Android Applicat ion. Procedia Computer Science. 157. 499-505. 10.1016/j.procs.2019.09.006.
[6] Vasudeva, Nisha & Parashar, Hem & Vijendra, Singh. (2012). Offline Character Recognition System Using Artificial Neural Network. International Journal of Machine Learning and Computing. 449-452. 10.7763/IJMLC.2012.V2.165.
[7] Sarkar, Deepayan. (2004). Optical Character Recognition using Neural Networks (ECE 539 Project Report).
[8] Chaudhuri, Arindam & Mandaviya, Krupa & Badelia, Pratixa & Ghosh, Soumya. (2017). Optical Character Recognition Systems. 10.1007/978- 3-319-50252-6_2.
[9] Liang, Jian & DeMenthon, Daniel & Doermann, David. (2008). Geometric Rectification of Camera-Captured Document Images. IEEE t ransactions on pattern analysis and machine intelligence. 30. 591-605. 10.1109/TPAMI.2007.70724.
[10] Ahmed, Muna & Abidi, Ali. (2019). REVIEW ON OPTICAL CHARACTER RECOGNITION.
[11] Sharma, Ankit & Chaudhary, Dipt i & Ahemedabad, & India,. (2013). Character Recognition Using Neural Network.

[12] N. Venkata Rao, Dr. A.S.C.S.Sast ry, A.S.N.Chakravarthy, Kalyan Chakravarthi " opt ical character recognition t echnique algorithms", Journal of Theoretical and Applied Information Technology , Vol.83. No.2,20th January 2016.
[13] Manoharan, Samuel. (2019). A SMART IMAGE PROCESSING ALGORITHM FOR TEXT RECOGNITION, INFORMATION EXTRACTION AND VOCALIZATION FOR THE VISUALLY CHALLENGED. Journal of Innovative Image Processing. 1. 31-38. 10.36548/jiip.2019.1.004.
[14] Chirag Patel, Atul Patel, Dhamendra Patel, "Optical Character Recognition by Open Source OCR Tool Tesseract A Case Study ", International Journal of Computer Applications Volume 55– No.10, October 2012.
[15] Metcalf, T.. (1986). Automat ic character recognition system. Patent Department of the Navy, Washington, DC..
[16] Tiwari, Usha & Jain, Monika & Mehfuz, Shabana. (2019). Handwritten Character Recognition—An Analysis. 10.1007/978-981-13-0665-5_18.
[17] Kaur, Dalvir & Sharma, Sukesha. (2019). Various Feature Extraction and Classification Techniques. 10.1007/978-981-10-8234-4_51.