Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

NoteMate - A Note Making System Using OCR and Text Mining

Chetan Botre, Saad Patel, Shrinivas Kunjir, Swapnil Shinde
Comp. Department, SPPU, Maharashtra,
India

Abstract—in today's world, all aspects of our lives have a digital presence. There are high performance cell-phones, laptops and tablets all around us. 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. Note Mate is an application which uses this fact to attempt to digitize the activity of Note-Making. It is based on Android platform with cloud as the storage medium. In this project, we are implementing Optical character recognition in order to convert handwritten scripts on a touch screen device into text format, which can be stored for later retrieval or editing using any text-editor software. A simplified neural approach to recognition of optical or visual characters is portrayed. Tools and services for users to download educational contents interact with teachers as well as other pupils to discuss topics and also extract the accurate information as per requirement are also provided.

Keywords— OCR , NoteMate, Text Mining, Template Matching, Android.

I. INTRODUCTION

Our aim is to simplify and modernize note making activity, integrating services like OCR and Text mining to deliver a complete efficient and effective note making application, which can be used by users ranging from students, to professionals. We provide a GUI for handwriting based input on touch screen devices, working on android operating system. The user can create; share, save notes on a centralized and personalized cloud platform. The handwriting based input is converted into text format using Optical Character Recognition (OCR).

II. EXISTING SYSTEM

In presently available note making applications, the handwritten input data is saved in the device storage as it is in image form, without converting it to .txt format. Such stored data takes lot of space of device. To avoid this wastage of space we convert this data in text format and then store it to the device or on the cloud.

For example, an existing android application "Papyrus" [1] offers note making, but fails to convert the note into text format and also does not provide cloud based storage and advanced searching options.

III. DRAWBACKS OF THE EXISTING SYSTEM

There are numerous existing note making systems for handheld devices. However, these systems have the following drawbacks-

1. Existing systems for note making applications lack the functionality of converting handwritten digital notes to text format and provide a centralized storage option.
2. This makes it mandatory to have the corresponding software installed on the device in order to view or edit existing notes.
3. Thus, It becomes impossible to access these notes on any other unsupported devices and becomes difficult to further process, share or edit these notes.

IV. PROPOSED SYSTEM

In order to overcome the drawbacks of the existing note making systems currently available in the market, the following additions-

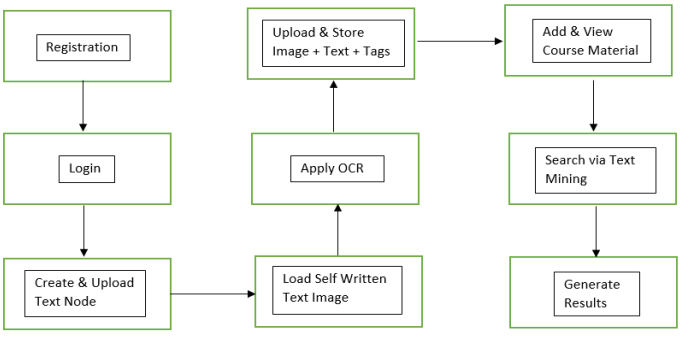

Figure 1: Proposed System

1. A framework to convert raw, handwritten data into systematic notes in text (.txt) format.
2. A novel optimization method to search from existing notes and on the cloud using text mining technique.
3. An online library for providing readymade content online, which can be viewed in the same application.

V. OPTICAL CHARACTER RECOGNITION

Optical Character Recognition is the process which can detect and identify and extract text from a text source, which may be digital or physical. There are a lot of techniques out there to implement Optical Character recognition (O.C.R). This system implements OCR using template matching. Template matching can be considered as the main step, to which all other prior steps lead to.

The following diagram is a flowchart showing the various steps involved in OCR.

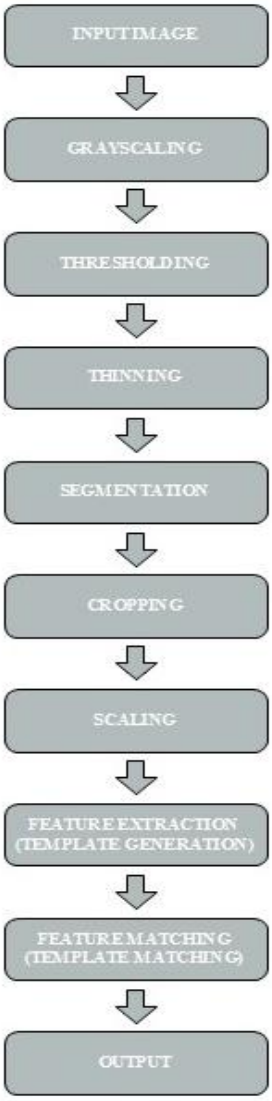

Figure 2: OCR Flow

A. Gray scaling

It is the process of averaging RGB (Red-Green-Blue) values. It converts input image to black and white image.

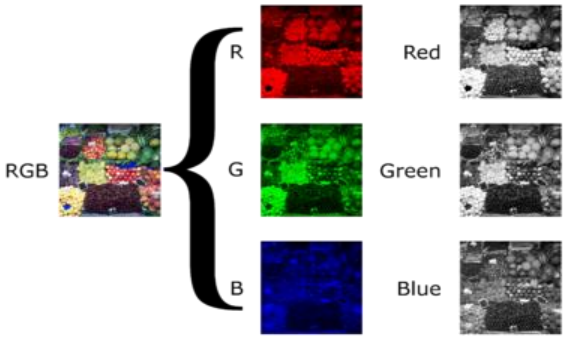

Figure 3: RGB Averaging

B. Thresholding:

It converts original image to binary form. Binary image pixel can contain only two possible color values.

Figure 4: Thresholding

C. Thinning:

Thinning is the process of reducing the dark parts of the handwritten text. It cleans the image so that only reduced amount of data needs to be processed in the next image processing stage.Thinning is the transformation of a digital image into a simplified, but topologically equivalent image.

Figure 5: Thinning

D. Segmentation:

The process of locating regions of printed or handwritten text is segmentation. Segmentation differs text from figures and graphics. When segmentation is applied to text, it isolates characters or words. Segmentation is the process of separating out individual alphabets.

Figure 6: Segmentation

E. Cropping:

Each individual token is cropped for further processing.

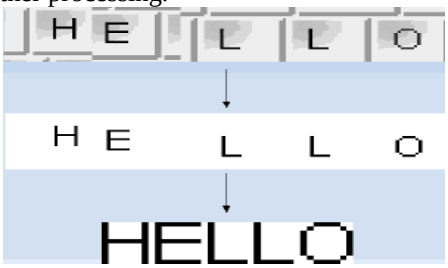

Figure 7: Cropping

F. Scaling:

The image is scaled to make sure that input token conforms to the size of data set tokens.

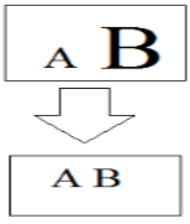

Figure 8- Scaling

G. Feature Extraction:

Feature extraction is the process of image digitization. It is an essential step required prior to template matching. Images are nothing but a large collection of pixels of varying characters like color, intensity and location. Feature extraction stores the image pixel information in a matrix containing 1's and 0's.

H. Template Matching:

This process involves the use of a database of characters or templates. There exists a template for all possible input characters. For recognition to occur, the current input character is compared to each template to find either an exact match, or the template with the closest representation of the input character. It is the technique to find small parts of an image which match a template image. [7]

VI. TEXT MINING

Text mining, also referred to as "data mining", refers to the process of deriving high quality information from text.
Text mining involves the following concepts-

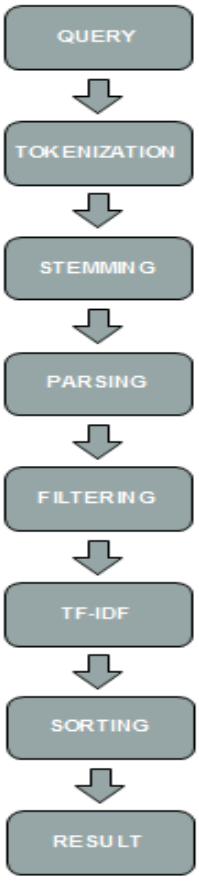

Figure 9: Text Mining Flow

A. Query:

A query is entered to retrieve the appropriate content according to the user's requirements. The query is the search query which is entered by the user and is the input for text mining.

B. Tokenization:

The entered sequence of string query is broken up into pieces like sentences and words by eliminating the whitespaces and punctuation marks.

C. Stemming:

Stemming is a technique used to reduce the inflected words to their root words. The stem is similar to the root word and it is generally mapped to it.

D. Parsing:

The body of the entered query is then extracted to which the regular expressions are applied upon which the further processing is done.

E. Filtering:

The stop words such as 'and' and 'the' are filtered out to enable a higher level of accuracy.

F. TF-IDF:

It stands for term frequency-inverse document frequency. It is a number that shows how important a word is to a document. In other words it's a weighting factor for information retrieval and text mining.

G. Sorting:

This is the technique wherein the documents are sorted in a prioritized manner and presented to the end user in that order. Here we are using sorting to sort the saved notes by the user for future reference. Also we are using the sorting technique for text mining purposes as in sorting according to relevance to the user query and presenting to the user in that order.

H. Result:

Based on the above steps, an output is generated by the system.

Abbreviations and Acronyms

- Tf - Term Frequency
- Df – Document Frequency
- Idf – Inverse Document Frequency
- w – tf-idf weight

q_i is the tf-idf weight of term i in the query
d_i is the tf-idf weight of term i in the document
$\cos(q,d)$ is the cosine similarity of q and d … or,
equivalently, the cosine of the angle between q and d.

II. Equations:

(1) Score $= \sum_{t \in q \cap d} (1 + \log \mathrm{tf}_{t,d})$

(2) $\mathrm{idf}_t = \log_{10} N/\mathrm{df}_t$

(3) $\mathrm{w}_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log N / \mathrm{df}_t$

VII. PROPOSED SYSTEM

1. The proposed system aims to apply the above given techniques and provide a means to create, store, edit handwritten digital notes and at the same time provide the ability to apply optical character recognition to the created notes in order to acquire the note in simple text.
2. The optical character recognition will employ Template Matching as its implementation technique.
3. Users of this application will have their notes saved on cloud, which provides the dual benefit that the notes will be easily available on any device hassle-free anywhere in the world, and also keep a backup of their notes.
4. Text mining will be used for advanced search to retrieve notes from the cloud, relevant to the keywords entered.

VIII. ADVANTAGES

1. The proposed system of note-making using neural networks learns the individual user's style of writing and adjusts itself to provide more precise results.
2. The use of text mining on the cloud provides advances search and retrieval features for retrieving notes from the cloud.
3. Another most obvious advantage our system provides is the conversion of handwritten text into a format which can be editable even on any basic text editor software.

IX. CONCLUSION

Thus, NoteMate is an application which uses Android platform with cloud as the storage medium. With NoteMate, we aim to provide a robust application for the purposes of note making which contains all the essential features that a note making application should provide, like OCR and more.

REFERENCES

[1] Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN ModelsSalvador Espana-Boquera, Maria Jose Castro-Bleda, Jorge Gorbe-Moya, and Francisco Zamora-Martinez.

[2] Mobile Camera Based Text Detection and Translation Derek Ma,Qiuhau Lin Tong Zhang.

[3] Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection, Lus Filipe da Cruz Nassif and Eduardo Raul Hruschka , IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 1, JANUARY 2013

[4] A. Vinciarelli, S. Bengio, and H. Bunke, O_ine Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models, IEEE Trans. Pattern Analysis and Machine Intel- ligence, vol. 26, no. 6, pp. 709-720, June 2004.

[5] A Review of Research on Devnagari Character Recognition Vikas J Dongre Vijay H Mankar, Department of Electronics Telecommunication, Government Polytechnic, Nagpur, India

[6] A. Gordo, A. Forns, and E. Valveny, Writer identification in handwritten musical scores with bags of notes, Pattern Recognition, vol. 46, no. 5, pp. 13371345, May 2013.

[7] Optical Character Recognition By Using Template Matching (Alphabet) Nadira Muda, Nik Kamariah Nik Ismail, Siti Azami Abu Bakar, Jasni Mohamad Zain Fakulti Sistem Komputer & Kejuruteraan Perisian.

