A Study on Handwriting Analysis by OCR

Sukanya Roy[*], Tridib Dawn[*], Subrata Deb[*]

*Computer Science &amp; Engineering, University of Engineering and Management, Kolkata

***Abstract* –**OCR, commonly known as Optical Character Recognition also known as Optical Character Reader. It is used to collect information from handwritten documents or manuscripts, printed paper data records. It is a common method of digitizing handwritten manuscripts so that they can modified or used digitally in modern technology.

***Index Terms* –** Acknowledgement, Conclusion, OCR Procedure, Text Processing.

I. INTRODUCTION

THE recognition and conversion from images of text have always been a challenging task for automatic data processing and information retrieval and services. In particular, the task of scanning human handwriting and making them not only digital readable, but also searchable and digitally editable, is important to retrieve and collect information.

In this way, the information of old manuscripts or any handwritten document can be a valuable and interesting source to build a strong and complete information network. Different organizations are interested in the mass scale digitization of historic manuscripts or handwritten documents with a focus on offering improved full-text searching.

II. OCR PROCEDURES

Different digital collections and information systems digitalize handwritten documents, such as old manuscripts of any historic civilization or very old manuscript or any handwritten document. However, optical character recognition of old handwritten manuscripts often poses different challenges. In the following phrases we summarize the main issues.

III. OCR problems of Handwritten Documents

Working with handwritten documents, we face different problems.

Image problems: One of the major problems is the quality of the original manuscript and the quality of the scan. This includes issues such as curled pages, blurred fonts, or manually edited pages (e.g. stamps or hand-written notes).

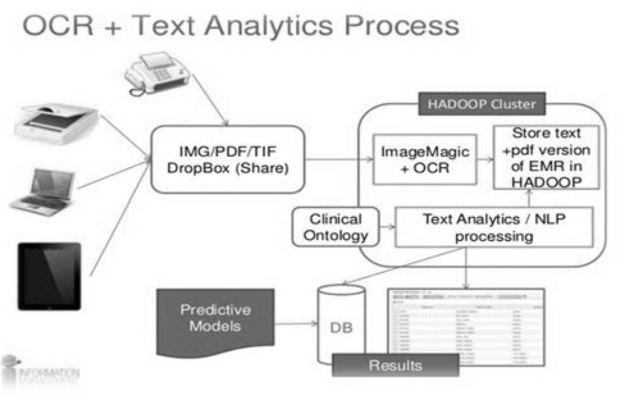
Font type and layout:Human handwritings are not supported by standard OCR software. The fonts or characters are different with the change of person.Old manuscripts are even harder to recognize, because they use font styles that are totally new to modern human society. Thus, also the spacing between words and characters is often not consistent. Additionally, historic papers often use different and inconsistent layout structures.

Missing knowledge base: Traditional OCR software uses knowledge bases based on contemporary dictionaries and grammatical structures to enhance the OCR procedure and does not provide manuscripts documents. Additionally, historic manuscripts often do not follow specific orthographic structures and rules, thus words can be written differently in the same text.

Every manuscript or handwriting is different from each other. Thus, very specific and unique problems can occur for every project. In the next section, we take a closer look at the overall OCR process, and possibilities to improve the different steps.

IV. OCR Process

Since technology develops by time the necessity of recording data in handwritten format decreases by time, now a days all records are being kept in format of digitized text document or media, so it is necessary to focus on these specific problems in the OCR process. In the following section, we describe the procedure of OCR with a focus on creating a learning / feature base for handwritten documents, which can be used for improving machine learning algorithm. To improve the accuracy of the OCR process, different actions can be taken in every single step of the process.


OCR + Text Analytics Process
HADOOP Cluster
IMG/PDF/TIF DropBox (Share)
ImageMagic + OCR
Store text +pdf version of EMR in HADOOP
Clinical Ontology
Text Analytics / NLP processing
Predictive Models
DB
Results
INFORMATION

- **Scanning:** The first phase of Optical Character Recognition is the scanning phase. This phase is one of the most important phases. If possible, scans should be made of well-preserved and clean originals. The scanning resolution should be at least 300 dpi and the output image a lossless image format (e.g. tiff).

- **Pre-processing:**This is the second phase of Optical Character Recognition. In this step, the scanned document can be manually optimized for the OCR process. This includes image editing processes such as increasing the contrast, reducing noise, or simplifying the colors.
- **OCR-process:** In this phase of Optical Character Recognition, the chosen OCR system reads the images and applies an algorithm to recognize the characters. It is crucial to choose OCR software that fits the current problem and supports a training/learning algorithm.
- **Create learning base:** To improve the OCR-process it is very important to create and improve the learning base for training the OCR system. This base consists of a dictionary fitting the document improved character pattern.
- **Post-processing:** In this phase, knowledge can be applied which had not yet been available to the OCR system. In a final step, the output can be corrected manually.

V. Conclusion

In this paper, we tried to throw light on the process of OCR of scanned old documents, historic books, manuscript of a very old document. To compare the accuracy of the OCR methods, a normalized version of the Levenshtein distance can be used. Since every historic book is different and poses its own and new challenges, the most important step of an OCR process is building a learning base. The main contribution of this work is a model for OCR processes of historic books with old fonts. With such a model, with respect to preened post-processing, the accuracy of OCR of manuscripts or human handwriting can be improved significantly, compared to related approaches.

VI. ACKNOWLEDGMENT

This work would not be possible without the help of Prof. Sukanya Roy, Lecturer at University of Engineering & Management, Kolkata. We would also like to thank the other group members for their hard work and dedication to complete the case study on Optical Character Recognition in Handwriting Analysis.

VII. REFERENCES

[1] Holley, R., "How good can it get? Analyzing and improving OCR accuracy in large scale historic newspaper digitization programs." D-Lib Magazine 15.3/4 (2009).

[2] “The challenges of historical materials and an overview on the technical solutions in IMPACT” [Online]. Available: https://impactocr.wordpress.com/2010/05/07/anoverview-of-technical-solutions-in-impact/

[3] Mori, S,, Ching Y. S., and Kazuhiko Y., "Historical review of OCR research and development."

Proceedings of the IEEE 80.7 (1992): 1029-1058.