Devanagari OCR: Issues and Analysis of Newspaper digitization

Deepak Kumar Arya
Centre for Development of Advanced Computing, Noida

Abstract- *OCR software attempts to replicate the combined functions of the human eye and brain, which is why it is referred to as artificial intelligence software. A human can quickly and easily recognize text of varying fonts and of various print qualities on a newspaper page, and will apply their language and cognitive abilities to correctly translate this text into meaningful words. This paper highlights some of issues that came up during the course of the newspaper digitization, how OCR software works on newspapers, factors that effect OCR accuracy, methods of improving accuracy, and testing methods and results for specific solutions that were considered viable for large scale text digitization projects.*

Keywords:- Digitization, Newspaper, Preprocessing, OCR

I. INTRODUCTION

Historical newspapers are one of the most significant sources of information for researchers due to the wealth of information they provide regarding every aspect of everyday political, social and intellectual life.

The issues associated with the newspaper digitization are how to prepare the scanned pages which can be used by OCR software[1] as input, how to handle bulk segmentation, OCR and post processing of newspaper pages[1] , how to perform quality check on the bulk data. The OCR software is used to analyze the structure of the newspaper page. It divides the page into elements such as blocks of texts (columns), images, etc. The lines are divided into words and then into characters. Once the characters have been singled out, the program compares them with a set of pattern images stored in its database. Finally, the software makes a best guess decision on the character. There are some common issues that can affect the OCR accuracy are deteriorated originals, unusual fonts, faded printing, shaded backgrounds, fragmented letters, touching/overlapping letters, skewed text, curved lines. There are lots of common problems associated with newspaper papers, so it is essential to measure OCR Accuracy. The OCR accuracy is determined by percentage of words on the page that are accurately "read" by the software. There are some of the techniques which we can use to improve OCR accuracy like improving the original source quality of input, scan at 300 dpi resolution or above, and use tiff files only for OCR, removing the skew and noise from pages, using the post processing dictionaries to minimize the manual editing and quality check effort.

II. WHY NEWSPAPER DIGITIZATION?

Information storage and management: Digitizing approximately 100 million pages or 200 billion characters of information. The storage requirements for the image files will be approximately 25 petabytes ? an order of magnitude larger than any publicly available information base. Creating and managing such a vast information base poses many technological challenges and provides a fertile test bed for innovative research in many areas[3]. Effort that will require the database to be globally distributed. For location independent access, this globally distributed database should appear to be a virtual central database from any place around the world. Mirroring the database in several countries will ensure security and availability. The network speeds at the various nodes would be different. Research in distributed caching and active networks would be needed to ensure that the look and feel of the database is the same from any location.

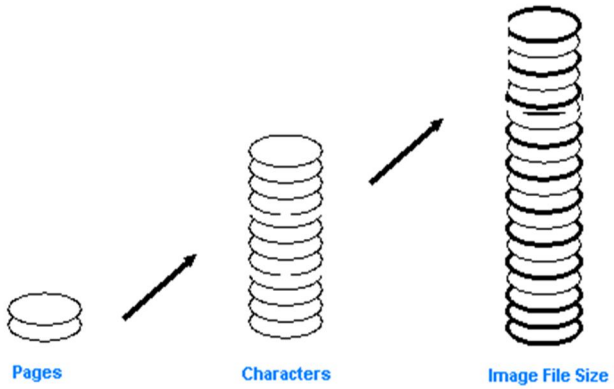

Figure 1.0: Information Storage and Management.

Search engines: The search engines of today work on the principle of keyword matching and perform searches in one language at a time. With a large corpus of multilingual data along with multilingual summarization and translation tools, a well-directed research effort would be needed to ensure concept- and content-based retrieval of knowledge from across multilingual data.

III. ABOUT THE SCRIPTS

Devanagari is the script for Hindi which is official language of India. It is also the script for Sanskrit Marathi, and Nepali languages. The script is used by more than 450 million[1] people on the globe.

Devanagari script is a logical composition of its constituent symbols in two dimensions. It is an alphabetic script. Devanagari has 11 vowels and 33 simple consonants. Besides the consonants and the vowels, other constituent symbols in Devanagari are set of vowel modifiers called *matra* (placed to the left, right, above, or at the bottom of a character or *conjunct*), pure-consonant (also called half-letters) which when combined with other consonants yield conjuncts. A horizontal line called *shirorekha* (a headerline) runs through the entire span of work. Some illustrations are given in Figs. 2.0 and 2.1. Devanagari script is a derivative of ancient Brahmi script[1] which is mother of almost all Indian scripts. Word formation in Indian scripts follows a definite script composition rule for which there is no coun- terpart in Roman[1].

A simplified Devanagari script composition grammar as proposed in [2] is presented here

< word > : = < composite char > + (shiro rekha)
< composite char > : = < vowel > * <conjunct > * < conjunct > + < matra >
< conjunct > : = <constant > < pure constant > * < conjunct >+

The script composition grammar imposes constraints on the symbols recognized by the OCR . A Devanagari word can be analyzed into three zones[1] : (i) a core zone (ii) upper zone (iii) bottom zone. The upper zone is the region above shiro rekha which contain all the top modifiers. The core zone contains all the characters and the vowel modifier. The lower vowel modifiers (matra) are in the lower zone. The character set of Devanagari can be divided into different groups based on the coverage pattern of the core region.

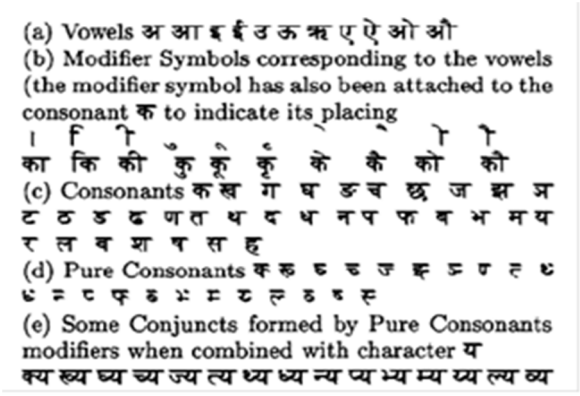
(a) Vowels अ आ इ ई उ ऊ ऋ ए ऐ ओ औ
(b) Modifier Symbols corresponding to the vowels (the modifier symbol has also been attached to the consonant क to indicate its placing
ा ि ी ु ू ृ े ै ो ौ
का कि की कु कू कृ के कै को कौ
(c) Consonants क ख ग घ ङ च छ ज झ ञ
ट ठ ड ढ ण त थ द ध न प फ ब भ म य
र ल व श ष स ह
(d) Pure Consonants क् ख् ग् घ् ङ् च् छ् ज् झ् ञ्
ट् ठ् ड् ढ् ण् त् थ् द् ध् न् प् फ् ब् भ् म् य्
र् ल् व् श् ष् स् ह्
(e) Some Conjuncts formed by Pure Consonants modifiers when combined with character य
क्य ख्य घ्य च्य ज्य त्य थ्य ध्य न्य प्य भ्य म्य य्य ल्य व्य

Figure 2.0: Characters and Symbols of Devanagari Script.

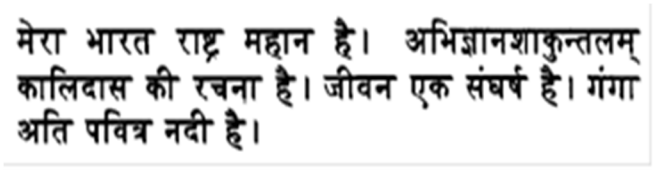

Figure 2.1: Sample Hindi text written in Devanagari script.

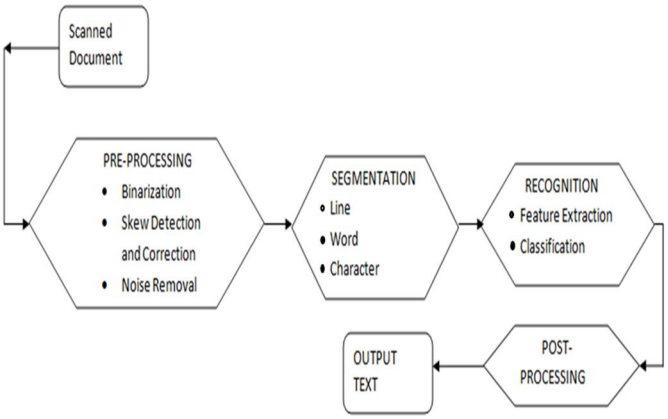

Figure 2.2: Overall framework of end-to-end OCR.

IV. CHALLENGES

Deteriorated originals, unusual fonts, faded printing, shaded backgrounds, fragmented letters, touching/overlapping letters, skewed text, curved lines

Figure 3.0 i) Deteriorated originals

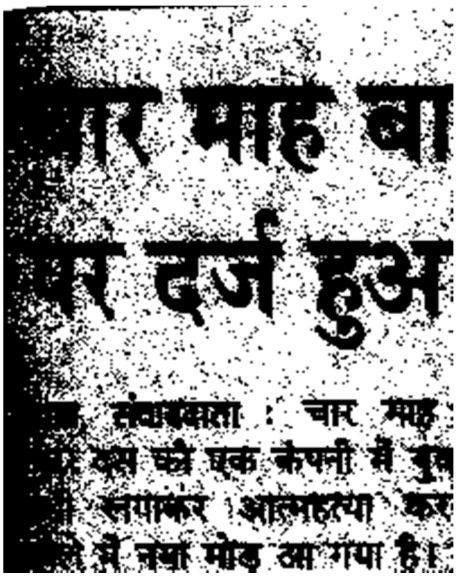

Figure 3.0 ii) Deteriorated originals

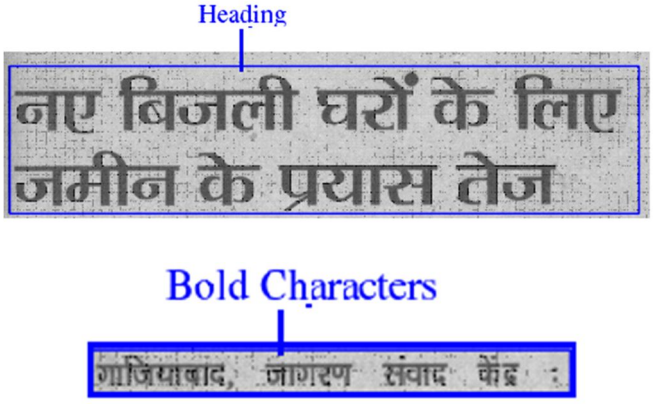

Figure 4.0 Unusual fonts and faded printing

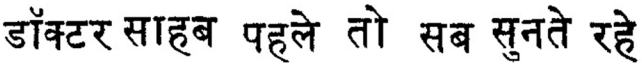

Figure 5.0 Local skew

Figure 6.0 Skewed text

V. PREPROCESSING

A. Scanning

Resolution	: -	300 DPI
Image Format	:-	Tiff Image
Image Type	:-	Gray Scale

B. Quality Check

In case the document does not meet the quality check, the same will be rejected. The Operator will be informed about this. The accepted documents will be queued for subsequent Preprocessing operation[5].

C. Preprocessing(Binarization / Noise Cleaning / Skew Correction)

The document images were Binarized[2], Noise Cleaned and Skew Corrected before extracting the required components.

Adaptive binarization method extends Otsu's method to a novel adaptive binarization scheme. The first step of our method is to divide images into NxN blocks[2], and then Otsu's method is applied straightaway in each of the blocks. Then each and every pixel is applied with a nonlinear quadratic filter to fine tune all the pixels according to the local information available.

In Otsu's method[2] we exhaustively search for the threshold that minimizes the intra-class variance, defined as a weighted sum of variances of the two classes:

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t)$$

Weights ωi are the probabilities of the two classes separated by a threshold t and σ_i^2 variances of these classes. Otsu shows that minimizing the intra-class variance is the same as maximizing inter-class variance

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = \omega_1(t)\omega_2(t)\left[\mu_1(t) - \mu_2(t)\right]^2$$

which is expressed in terms of class probabilities ωi and class means μi which in turn can be updated iteratively.

The morphological opening[2] and closing operators not only remove image noise but also connect discontinuities that are caused in the thresholding stage[2], in the character images that we have. The opening and closing operators are as follows:

$A \circ B = (A \otimes B) \oplus B$

and,

$A * B = (A \oplus B) \otimes B$

Which $\oplus$ and $\otimes$ are respectively the morphological erosion and dilation operators and B is the related structure element. The algorithm implemented is a two-pass algorithm to find the individual connected components in the document. While finding the component, the size of the component, ie. number of pixel in them are also computed. An elongation parameter for each component is also computed i.e. whether the component is elongated or not is decided. The users are the maximum size of components that will be considered as noise and hence deleted.

Next we used the mlskew algorithm[3] for correcting the skew in page which is generally caused due to incorrect scanning.

D. Manual Segmentation

Manual Segmentation of Image is done to separate out text and graphics region

नए बिजली घरों के लिए जमीन के प्रयास तेज

गाजियाबाद, जागरण संवाद केंद्र : जिलाधिकारी आर. रमेश कुमार के निर्देश के बाद बिजली घरों के लिए जमीन उपलब्ध कराने के प्रशासनिक स्तर पर प्रयास तेज कर दिए गए हैं।

सोमवार को डीएम के निर्देश पर अपर जिलाधिकारी (प्रशासन) सतीश कुमार ने पारेषण से जुड़े बिजली अधिकारियों की बैठक बुलाई। इसमें विचार विमर्श किया गया कि कहां-कहां पर बिजली घर बनाने में दिक्कत है और कहां से बिजली घरों के लिए भूमि उपलब्ध नहीं हो पा रही। इस पर अधीक्षण अभियंता आर.पी. सिंह और अधिशासी अभियंता ने बताया कि टीला मोड़ पर 132 केवीए बिजली घर के लिए भूमि प्रशासन उपलब्ध करवा रहा है। यहां पर बंथला ग्राम सभा की पांच एकड़ भूमि दी जा रही है। दोनों अधिकारियों ने बताया कि इंदिरापुरम में जीडीए वादे के बाद भी भूमि उपलब्ध नहीं करवा पा रहा है। यहां पर 400 केवीए का जीआईएस आधारित अत्याधुनिक बिजली घर तथा 220 केवीए का एक और बिजली घर बनना है। इस अत्याधुनिक बिजली घर में फाल्ट कम होंगे।

पुराने शहर में मालीवाड़ा में 132 केवीए बिजली घर बनना है। उन्होंने बताया कि गोविंदपुरम में बिजली घर के लिए जीडीए ने जो भूमि उपलब्ध करवाई है, वह कम है। अब नगर निगम से जीडीए द्वारा हस्तांतरित भूमि में कुछ जमीन उपलब्ध करवानी है। इसके साथ ही प्रताप विहार, मधुबन-बापू धाम और वसुंधरा में भूमि की आवश्यकता है।

Figure 7.0 Manual Segmenetation

E. Devanagari Ocr

Limitations of Devanagari OCR: -

OCR has never achieved a read rate that is 100% perfect. Because of this, a system which permits rapid and accurate correction of rejects is a major requirement. Exception item processing is always a problem because it delays the completion of the job entry, particularly the balancing function.

Of even greater concern is the problem of misreading a character (substitutions). In particular, if the system does not accurately balance data, customer dissatisfaction will occur. The success of any OCR device to read accurately without substitutions is not the sole responsibility of the hardware manufacturer. Much depends on the quality of the items to be processed.

F. Spell Checker

After the initial text is generated by the OCR, it is filtered through a Spell Checker [4] to ensure that only valid words are in the final text. we used a Devanagari dictionary of only 75583 words.

VI. CONCLUSION

The main Challenge of Digitization in Devanagari Script is the OCR engines are not developed which give high accuracy on Deteriorated originals, unusual fonts, faded printing, shaded backgrounds, fragmented letters, touching/overlapping letters, skewed text, curved lines newspaper images and devanagari newspaper printed quality is not at par as English Newspaper printed quality which make recognition a challenging task.

REFERENCES

[1] V. Bansal, R.M.K. Sinha, “Segmentation of Touching and Fused Devanagari Characters”, Pattern Recognition, vol. 35, pp. 875-893, April 2002.

[2] Tushar Patnaik, Shalu Gupta, Deepak Arya — Comparison of Binarization Algorithm in Indian Language OCR,ASCNT 2011.

[3] DIATHESIS: OCR based semantic annotation of newspapers. Martin Doerr, Georgios Markakis,. Maria Theodoridou, Minas Tsikritzis. June 2007

[4] Susan Haigh. Optical character recognition (ocr) as a digitization technology.Technical report, Information Technology Services National Library of Canada, November 15 1996.

[5] Microfilm, Paper, and OCR: Issues in Newspaper Digitization at the Utah Digital Newspapers Program (2004). Arlitsch, Kenning (1965)