Journal of Advances and Scholarly Researches in Allied Education
Vol. 15, Issue No. 10, (Special Issue) October-2018, ISSN 2230-7540

Issues in Devanagari Ancient Character Recognition: A Study

Sonika Rani Narang[1]* M. K. Jindal[2]

[1] Department of Computer Science, DAV College, Abohar, Punjab, India

[2] Department of Computer, Applications, P. U. Regional Centre, Muktsar, Punjab, India

Abstratct – A number of OCRs have been developed for the identification of fine printed Devanagari text. Some work has been done on recognition of handwritten Devanagari text. However very little work has been done in recognition of text for ancient handwritten Devanagari manuscripts. These manuscripts are very old and in delicate condition. These are not easily accessible. Ordinarily used techniques for handwritten recognition can't be applied directly to the ancient manuscripts as ancient manuscripts have unique characteristics. In this paper, these characteristics and different problems encountered while designing OCR for ancient handwritten Devanagari manuscripts have been identified.

Keywords – Problems during OCR, Devanagari Ancient Manuscripts, preprocessing

----------X----------

1. INTRODUCTION

OCR is the process of converting scanned images of machine printed or handwritten text into a computer processable format. In nation like India, there is abundance of data as manuscripts, ancient texts, books in that capacity. These manuscripts are wealth of knowledge. Due to their delicate condition, these ancient documents are not available to everyone. Additionally, such material is inadequate when it comes to searching data among a huge number of pages. It must be digitized and converted to a textual frame with the end goal to be recognized by machines. Optical character recognition assumes an imperative role in achieving this. There are different stages in any OCR process like image obtaining, preprocessing, segmentation, feature extraction, arrangement and post processing etc. Existing techniques of OCR can't be applied in straight forward manner to ancient manuscripts because there are numerous problems in these documents at each phase of OCR process. This paper describes such problems in ancient Devanagari manuscripts collected from different libraries and museums at different stages of OCR process.

This paper is organized as pursues Prior work is given in section 2. Section 3 describesvarious stages of a normal OCR process. Problems amid recognition of Devanagari ancient manuscripts are listed in Section 4. Finishing up notes and future work are given in Section 5.

General Terms

Pattern Recognition, Optical Character Recognition, OCR, Ancient Manuscripts

2. RELATED WORK

G.S. Lehal and R. Dhir[1] presented a range free skew detection technique for machine printed Gurmukhi documents. They discussed the problem of skewness in documents. M.K. Jindal, G. S. Lehal, and R. K. Sharma[2] studied segmentation problems in printed degraded Gurmukhi content and proposed answer for segmenting contacting characters in upper, middle and lower zones of machine printed Gurmukhi content. L. Sulem, A. Zahour, B. Taconet[3] presented a survey on text line segmentation of authentic documents. They likewise mentioned different types of problems encountered amid recognition of ancient manuscripts. M. K. Jindal, R. K Sharma., and G. S. Lehal[4] presented an investigation of contacting characters in degraded Gurmukhi text. They divided contacting characters in different categories and proposed new calculation to segment contacting characters in the middle zone of the machine printed content. F. Kleber and R. Sabletnig[5] presented a method to extrapolate missing parts of a degraded ancient document based on from the earlier knowledge. This paper presented a calculation for decision estimation of glagolitic texts based on text line extraction and is suitable for degraded manuscripts. M. Diem and R. Sablatnig[6] showed approach to recognize

degraded characters in glagolitic original copy utilizing neighborhood features. This work is done on ancient composition where characters were washed out(partially visible) due to maturing. G. Angelica , M. Diem , R. Sablatnig[7] proposed a vigorous method to separate text area from decorative area in ancient documents utilizing scale Invariant feature transform(SIFT) and nearby descriptors. K. R. Shah and D. Dattatray[8] gave a review on Devanagari Handwritten Character Recognition (DHCR) for Ancient Documents. They listed different challenges in OCR for ancient Devanagari documents like recoloring of text, maturing of text, overlapping of text line and detection of characters contacting to page lines.

3. VARIOUS STAGES OF OCR PROCESS

There are a number of stages in any OCR process. A typical OCR process may involve following stages:

3.1 Image Acquisition:

Computerized images of printed or handwritten text are acquired utilizing some advanced scanner or advanced Camera.

3.2 Pre-processing:

The digitized images obtained in the main stage might be of low quality. These may contain noise or inclination etc. The pre-processing stage changes an image. It takes in a crude image, enhances it, reduces noise, and removes skewness and inclination. This phase simplifies processing of the rest of the stages.

3.3 Segmentation:

Segmentation is the process of extracting imperative objects by apportioning an image into foreground and foundation pixels. Much emphasis is laid on this phase. This stage is further divided into three sub-phases:

3.3.1 Line segmentation:

In this phase, Boundaries of various lines are identified in the image.

3.3.2 Word Segmentation:

In this phase, words are separated in a line.

3.3.3 Character Segmentation:

In this phase, characters of a word are separated from the already identified words.

3.4 Feature extraction:

The feature extraction stage analyzes a character and selects a set of features which can be used to uniquely identify the character. This phase is likewise very essential in the OCR process. It is very essential to select a vigorous and representative set of features for any OCR system.

3.4 Classification:

The grouping stage uses features extracted in the previous stage. It then classifies the characters as per some preset rules.

3.5 Post-processing:

The post-processing stage deals with error detection and error correction. It improves recognition by refining the decisions taken in the previous stage and recognizes words by context and punctuation.

4. PROBLEMS DURING RECOGNITION OF DEVANAGARI ANCIENT MANUSCRIPTS

Ordinarily used techniques for handwritten recognition can't be applied in a straight forward manner to the ancient manuscripts as ancient manuscripts have several unique characteristics. There might be problems in every stage of OCR for ancient manuscripts as discussed below.

4.1 Problems in digitization phase:

4.1.1 As Ancient documents might be fragile due to maturing, it is difficult to handle these. Numerous a times scanner can't be used to acquire computerized images of the documents. Sometimes an advanced camera hasbeen used for procuring computerized images of these documents**.**

4.1. Noise added due to the scanner source brightness: Sometimes there might be brightness due to scanner source especially in case of image acquired utilizing advanced camera. It results in noise. Some libraries have preserved ancient pages by overlaying these. This cover results in noise.Fig 1 indicates one such document where noise due to camera streak happens.

Fig1: Noise due to the scanner source brightness

4.2 Problems in pre-processing phase:

Pre-processing phase of Ancient manuscripts may face numerous problems as discussed below:

4.2.1 Decorative border:

Numerous ancient documents use decorative borders around the text. It is hard to extract text from such pages.Fig 2 demonstrates a document having decorative border. Before beginning OCR process, text must be separated from border in these documents.

Fig2: Decorative Border

4.2.2 Paper color change:

Paper shading may change due to maturing. It might result in noise. It becomes hard to recognize foundation shading and foreground colour.Fig 3 demonstrates a document where foundation and text shading have changed due to maturing.

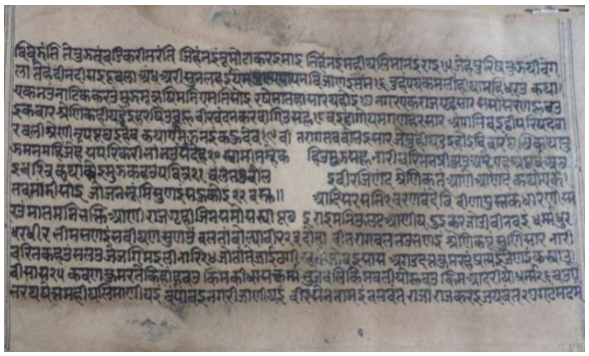

Fig3: Changed paper colour due to aging

4.2.3 Torn Paper:

Paper might be torn. Examining may not be clear because of this and some characters might be missing.Fig 4 demonstrates a document with torn corner. It is very hard to retrieve these lost characters.

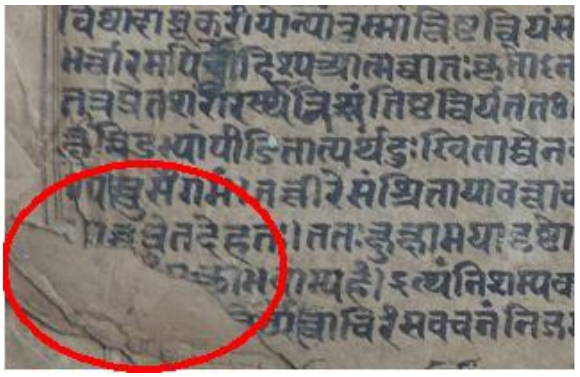

Fig4: Torn paper

4.2.4 Ink Seepage:

There can be ink seepage. It is called "Back Paper Noise". This noise is due to imprinting on the two sides of paper.Fig 5 demonstrates a document which contains noise due to ink seepage.

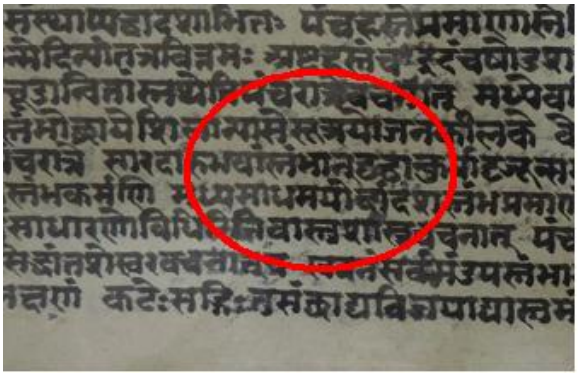

Fig5: Noise due to ink seepage

4.2.5 Skewed Text:

Ancient documents might be skewed or these may have slant.Skewed or slanted text makes segmentation troublesome. Sometimes this inclination or skewness must be corrected in preprocessing phase. Fig 6 demonstrates a document with skewed text.

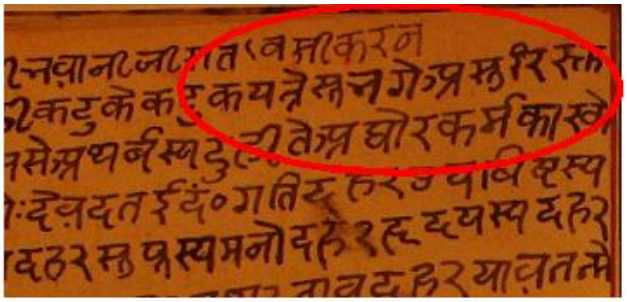

Fig 6: Skewed text

4.2.6 Holes or spots in documents:

Sometimes ancient documents are of low quality due to swoon composing. These include different aggravating elements, for example, holes or spot etc. These holes or spots may result in loss of data. Fig 7 demonstrates a document with a hole and some spots.

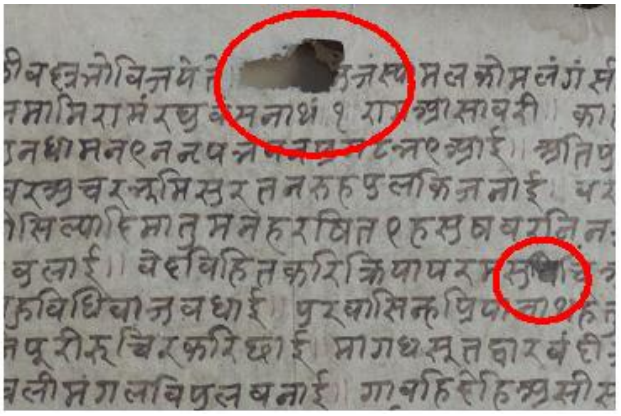

Fig 7: Holes in the document

4.2.7 Low visibility of characters:

Due to maturing, sometimes quality of printing is very low. Some characters may not be visible properly. It makes recognition of characters very troublesome. Fig 8 demonstrates a document with faded characters.

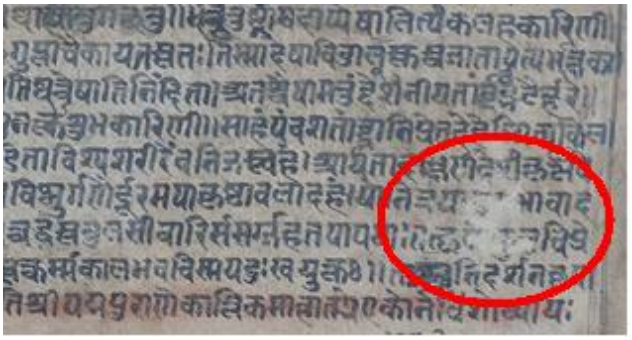

Fig8: Low visibility of characters

4.2.8 Ink stains:

In ancient documents fluid ink was used. Sometimes, this ink made stains on the paper. It makes preprocessing difficult. Fig 9 demonstrates a document containing different ink stains. These ink stains are in foreground shading. It is very hard to recognize stains and real characters.

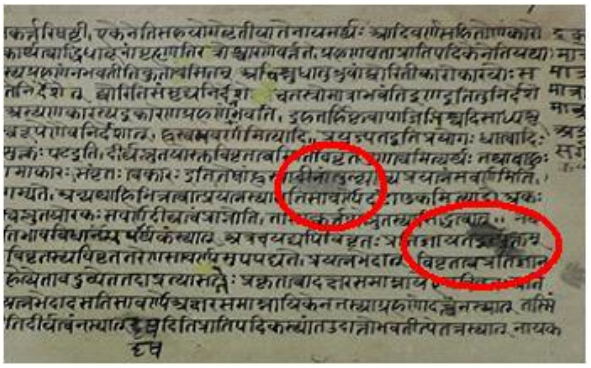

Fig9: Noise due to ink stains

4.2.9 Unwanted marks:

Sometimes, there might be unwanted stamps on the pages which make preprocessing troublesome. These unwanted imprints may change the shape of real characters as appeared in fig 10.

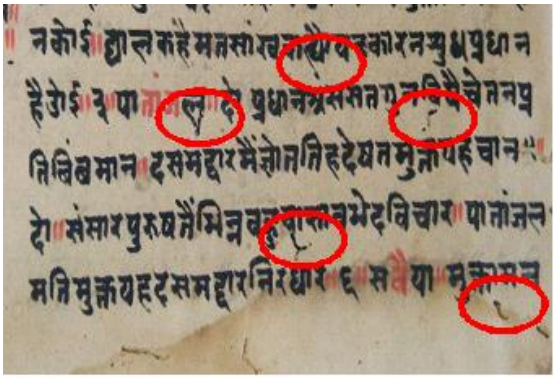

Fig10: Noise due to unwanted marks

4.3 Problems during segmentation:

4.3.1 Loose layout formatting:

In Ancient manuscripts, design arranging requirements were loose. So their physical structure is harder to extract. Additionally there might be more than one type of orientation on a single page.Fig 11 demonstrates one such document.

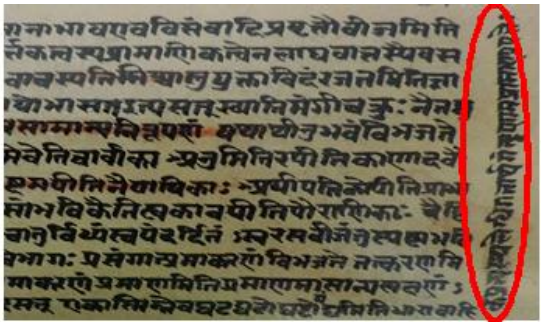

Fig11: Loose layout formatting

4.3.2 Narrow space between lines:

Sometimes there is narrow space between lines.

4.3.3 Overlapping characters:

These documents have overlapping components. So Line segmentation requires new algorithms.

4.3.4 Heavily printed and touching characters:

Characters are for the most part heavily printed. Additionally there are contacting characters. It makes recognition troublesome.

4.3.5 Uneven or no space between words:

There is uneven space between different words. Sometimes there is no space between words. So segmentation into characters and words isn't easy. It requires new calculations to segment words and characters. Fig 12 demonstrates a document which contains slender space between lines, some contacting characters and in addition overlapping characters.

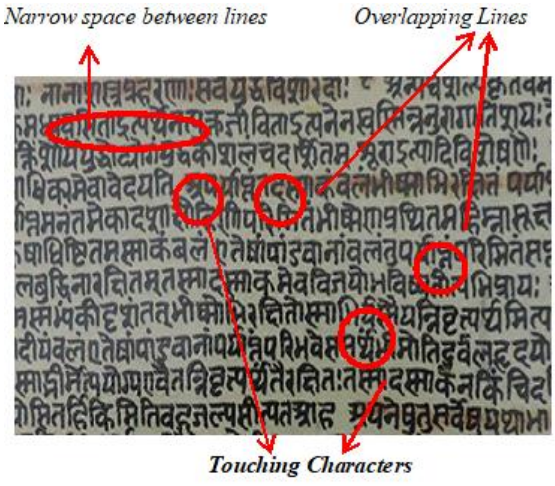

Fig12: Noise due to narrow space between lines, touching or overlapping characters

4.3.6 Words out of line boundaries:

Some words might be out of the standard line boundaries making segmentation troublesome. As appeared in fig 13, some words are written out of line boundaries.

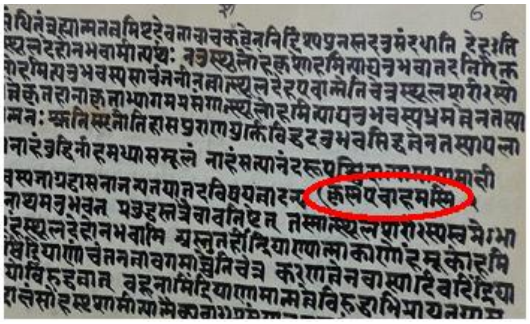

Fig13: Words out of line boundaries

4.4 Problems during recognition Phase:

4.4.1 Unusual characters:

Characters and words have bizarre and shifting shapes, depending on the writer, the period and the place concerned.

Manuscripts use some characters which are not used in modern Devanagari content. So some special features are required to recognize such characters. Fig 14 demonstrates a document that contains characters which are rarely used in modern Devanagari content.

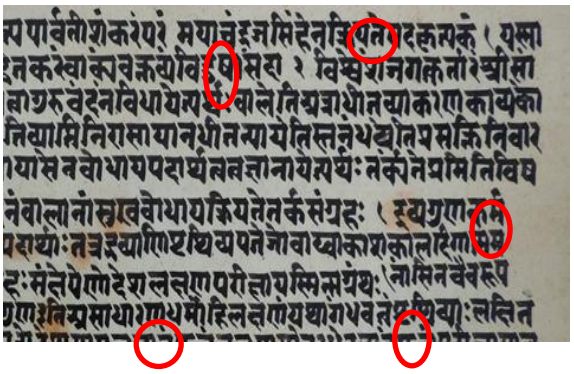

Fig14: characters with unusual and varying shapes

4.4.2 Deformed characters due to uneven spreading of ink:

In ancient documents, fluid ink was used. Sometimes there was uneven spreading of ink. More ink deformed some characters and less ink resulted in faded characters. Recognition of such characters is difficult. Fig 15 demonstrates a document containing deformed characters due to uneven spreading of ink.

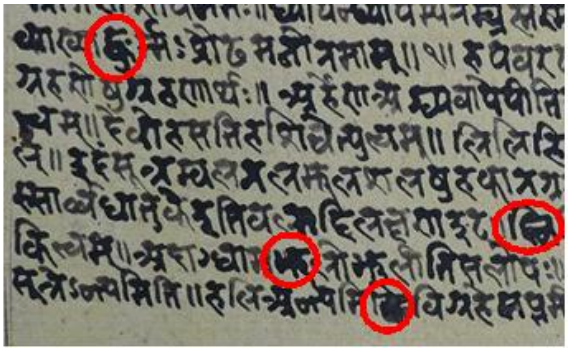

Fig15: Deformed characters due to uneven spreading of ink

4.5 Problems during post-processing phase:

The vocabulary in Ancient manuscripts is likewise large and may include strange names and words. So post-processing requires extra efforts.

5. CONCLUSION AND FUTURE WORK

The above discussed problems are found in ancient handwritten manuscripts collected from different libraries and museums. These documents were digitized utilizing a computerized camera. A few of the above discussed problems can be solved basically by utilizing some other method of digitizing documents yet some other problems are very run of the mill and require special attention. Different calculations need to be developed to remove these problems and to correctly identify text in ancient manuscripts. This investigation provides an understanding into the different research areas related to these documents. Every type of problem discussed in this paper needs special attention. Future work may involve discovering answers for these problems and more documents might be collected to uncover more problems in developing OCR for such documents. In future, researches may handle the problems discussed in the paper one by one and develop calculations for the same.

6. REFERENCES

[1] Lehal G.S., Dhir R. (1999). “A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents", ICDAR, pp. 147-152.

[2] Jindal M. K., Lehal G. S., and Sharma R. K. (2006). " Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script", International Journal of Signal Processing, pp. 258-267.

[3] Sulem L., Zahour A., Taconet B. (2006). "Text line segmentation of historical documents: a survey", International journal on document analysis and recognition, Springer, vol.9, pp. 123-138.

[4] Jindal M. K., Sharma R. K., and Lehal G. S. (2007). "A study of touching characters in degraded Gurmukhi text”, World Academy of Science, engineering and Technology, pp. 1069-1072

[5] Kleber F. and Sabletnig R. (2008). "Ancient document analysis based on text line extraction", 19th international conference on pattern recognition, pp. 1-4.

[6] Diem M. and Sablatnig R. (2009). "Recognition of Degraded Handwritten Characters Using Local Features", *10th International Conference on Document Analysis and Recognition*, pp. 221-225 Barcelona, Spain, 2009.

[7] Angelica G., Diem M., Sablatnig R. (2010). "Detecting Text Areas and Decorative Elements in Ancient Manuscripts", 12th International Conference on Frontiers in Handwriting Recognition, IEEE, pp. 176-181.

[8] Shah K.R. and Dattatray D. (2013). "Devnagari handwritten Character Recognition(DHCR) for Ancient Documents: A Review", Proceedings of 2013 IEEE Conference on Information and Communication Technology, pp. 656-660.

Corresponding Author

Sonika Rani Narang*

Department of Computer Science, DAV College, Abohar, Punjab, India

E-Mail – sonikanarang@davcollegeabohar.com