ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 3, March 2013

Segmentation of Handwritten Text Document- A Review

Er.Naunita

***Abstract*— Optical character recognition (OCR) is very popular research field since 1950's. A great work has been done for various scripts. This paper provides a review of various segmentations. We present a method for line segmentation, word segmentation and character segmentation of handwritten text document. The goal of Segmentation is to divide a text document into line, word and character. The main challenge in the segmentation of handwritten language is the variation in handwriting styles.**

***Index Terms*— OCR, Line segmentation, word segmentation, character segmentation, Handwritten Documents.**

I. Introduction

OCR stands for **optical character recognition**. OCR has been one of the most challenging research areas in field of image processing in the recent years. Several research works have been done to evolve newer techniques and methods [1]. It is electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is widely used for data entry purposes because computerized printed texts can be electronically stored, searched, displayed online.

paper document
scanned document
editable document

OCR is a field of research in pattern recognition, Artificial intelligence and computer vision. OCR finds its applications in a wide area. Some of the important areas are as automatic number plate recognition, sound recording and reproduction etc.OCR consists of many phases as given below:-

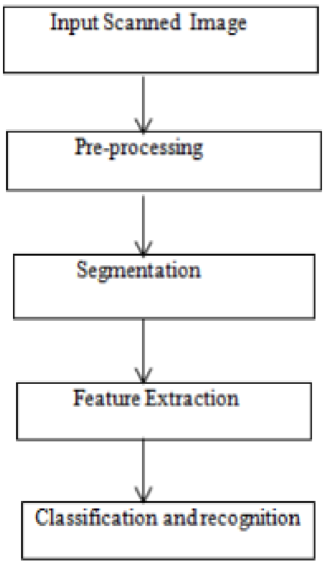

Figure 1: Phases in OCR

Segmentation is a technique, which partitions handwritten text into individual lines, words and characters[2]. Segmentation is the third step which comes after preprocessing step. Segmentation is of following types:

1) Line segmentation

2) Word segmentation

3) Character segmentation

In **line segmentation,** the lines of a text blocks are detected by scanning the input image horizontally. Frequency of black pixels in each row is counted in order to construct a row histogram. When frequency of black pixels in a row is zero, it denotes a boundary between two white pixels consecutive lines.

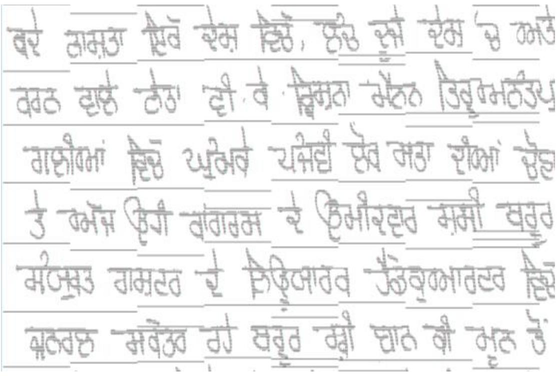

Figure 2: Segmented lines

Manuscript received Feb, 2013.

***Er.Naunita**,CSE Department, Punjabi University,Patiala,.Bhucho Mandi , India, 01642740007/ 9463841724*

In **Word segmentation** when a line has been detected, then each line has to scan vertically. Number of black pixels in each column is calculated to construct column histogram.

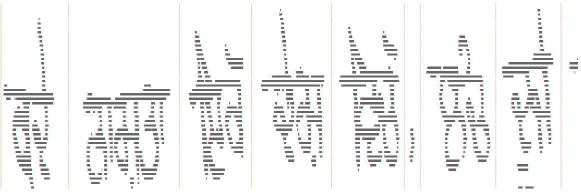

Figure 3: correctly segmented words

When there is no black pixel is found in vertical scan then there is a space between two words. Like this we can separate the words.

In **Character segmentation** a word is separated into characters, each individual character and composite character is separated for further identification.

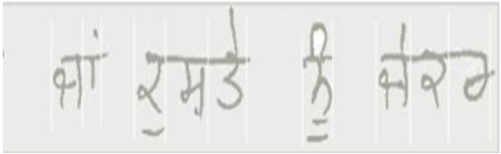

Figure 4: Word image after straighten header line

II. LITERATURE SURVEY

Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey by Pritpal Singh, Sumit Budhiraja. In this paper feature extraction of gurmukhi script has shown and proposed methods and techniques for this. There are many feature extraction techniques which are not implemented in case of handwritten gurmukhi script recognition e.g. wavelets, Fourier transform etc. So a lot of work can be done in field of Handwritten Gurmukhi Character Recognition **[1]**.

A method of Isolated Handwritten Words Segmentation Techniques in Gurumukhi Script has been proposed by Galaxy Bansal, Dharamveer Sharma**.** This paper presents the results of segmentation of words in various handwritings in Gurumukhi scripts. Handwritten words are also prone to the problem of overlapped, connected, merged and broken characters. Segmentation accuracy of 88.1% has been achieved for segmenting all types of handwritten words in Gurumukhi script **[2].**

A process of automation of script identification in Indian OCR is proposed by Tushar Patnaik, Bhupender Kumar, Deepak Kumar Arya. This paper represents a line based script identification that uses local approaches to identify devnagri, gurumukhi, bangla and telugu scripts in a multilingual document **[3].**

An automatic scheme is presented to identify text lines of different Indian scripts from a document has been proposed by U. Pal, S. Sinha, B.B. Chaudhuri, "Multi Script Line Identification from Indian Documents**[4].** At present, the system has been applied to 250 different multiscript document images containing about 4000 text lines. A least 250 text line images from each script were considered. Each line contains at least 10 characters. It was observed that the overall accuracy of the system is about 97.52%.

A process of Segmentation of Printed Text in Devanagari Script and Gurumukhi Script has proposed by Vijay Kumar, Pankaj K. Sengar. In this paper, line, word, character and top character segmentation for printed Hindi text in Devanagari script is done.Also a method has been described for line and word segmentation of printed text in Gurumukhi script. A performance of 100% at line level, approximately 100% at word level, 99% at character level, and 97% at top character level for Devanagari script and performance of 100% at line level and 99% at word level for Gurumukhi script is obtained**[5]**.

A method of line and word segmentation of handwritten document has proposed by G. Louloudis1, B. Gatos2, I. Pratikakis2, C. Halatsis1. In this paper, a methodology has been presented for segmentation of a handwritten document in its distinct entities namely text lines and words. Text line segmentation is achieved making use of the Hough Transform on a subset of the connected components of the document image **[6].**

A Complete Machine printed Gurumukhi OCR System by G.S Lehal1 and C. Singh. Gurumukhi script is used primarily for the Punjabi language, which is the world's 14th most widely spoken language. It is spoken by over 30 million people in India as well as people living in far flung countries such as UK, USA, Canada, UAE, Singapore, Kenya, Fiji and Malaysia. There is rich literature in this language in the form of scripture, books, poetry, etc. Gurumukhi is the first official script adopted by Punjab state. It is also the second language in many northern states of India; this is the first time that a complete multi-font and multi-size OCR system for Gurumukhi script has been developed. It has been tested on good quality images from books and laser print outs and has recognition accuracy of more than 97%. **[7].**

III. CONCLUSION

Segmentation of handwritten text is a complex task as various matras and the header lines are responsible for the complexity. There is proposed a new approach of Segmentation in handwritten document. The development architecture is vigorous in the recognition of handwritten documents. The algorithm worked well for all handwritten documents. In future there are some other technique may be tried for segmentation of overlapped characters. So character segmentation method can be changed to improve accuracy.

IV. REFERENCES

[1] Pritpal Singh, Sumit Budhiraja,"Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey".

[2] Galaxy Bansal, Dharamveer Sharma", "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script", ©*2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 24*

[3] Tushar Patnaik, Bhupender Kumar, Deepak Kumar Arya,"Automation of Script Identification in Indian OCR.",Proceedings of ASCNT-2011,CDAC- noida.

[4] U. Pal, S. Sinha, B.B. Chaudhuri, "Multi Script Line Identification from Indian Documents".

[5] Vijay Kumar, Pankaj K. Sengar", "Segmentation of Printed Text in Devanagari Script and Gurmukhi Scrip," International Journal of Computer Applications (0975 – 8887) Volume 3 – No.8, June 2010

[6] G. Louloudis1, B. Gatos2, I. Pratikakis2, C. Halatsis1. "Line and Word Segmentation of Handwritten Documents".

[7] "G.S Lehal1 and C. Singh", "A Complete Machine printed Gurumukhi OCR System"

Er. Naunita , I have received my B-Tech degree in Computer Science & Engineering from Shaheed Udham Singh College of Engg. & Technology, Tangori (Mohali) in 2009 and pursuing M-Tech degree in Computer Engineering from Yadwindra College of Engg, Talwandi Sabo (Bathinda).