A Novel Technique of Optical Character Recognition of Printed Text for Multi Languages

Ramandeep Singh[1] Jasdeep Singh Mann[2]
[1,2]Department of Computer Science & Engineering
[1,2]BMSCE Sri Muktsar Sahib, & Sri Muktsar Sahib, India

Abstract— Multi-scripts recognition systems are requisite in the countries like India where multi-languages are spoken in numerous states of the country. Multi-Scripts Recognition is a demanding problem and research work for expansion of optical character recognition scheme for bi-scripts and multi-scripts is in infancy. Here in presented work a multi-script recognition system is proposed for the English and Punjabi scripts. For recognition the image is processed through basic steps of OCR like pre-processing, segmentation, feature extraction, correlation calculation and classification. After binarization and noise removal of the test image, it is segmented using line segmentation, words segmentation, and character segmentation technique of proposed algorithm. The lines of both the languages are segmented using the horizontal projection profile of the image. The words segmentation of the Punjabi language is very easy as compared to the English language due to the fact that all characters of a Punjabi word is connected through a connecting line and two simultaneous words are separated by blank space. After the segmentation, the number of holes in image is calculated and then the correlation of this character is calculated with the particular group of trained database. And then the decision is made on the basis of the highly correlated image in the database. Grouping of the database is done to reduce the correlation calculation time for whole database. In last system efficiency is calculated by using the test images of various sizes. Experimental results show the high accuracy of the proposed system.

Key words: OCR, Segmentation, Binarization, Correlation

I. INTRODUCTION

Optical Character Recognition translates printed or handwritten scanned text in editable form. The course of action of this scheme starts with scanning of input documents to digital image and translating colored or gray scale 8 bit image into binary image. Then every character is goes through segmentation process and the consequence image of subdivided character is promoted to a pre-processor for diminution of noise and normalization. Some explicit features are extracted from the character for recognition. The feature extraction is decisive and several different schemes exist with merits and demerits. After recognition the recognized characters are gathered to refurbish the original text and then detect and correct misrecognized text through post-processing. An OCR system enables to get a printed or handwritten manuscript or a published item supply it straight into an electronic computer file and then change the file using a word processor. OCR's are used to convert books and documents into electronic files for computerized record keeping in an office or to publish on web.OCR is being used by libraries to digitize and preserve their holdings. A large number of magazines and letters are sorted every day by OCR machines. In OCR technology performance of the OCR is highly depends on the quality of the input image. The quality of input image is depends up on the quality of the scanner. Scanner with good color quality and with high speed is preferred.[4][5]

OCR deals with these three steps:

a) Document scanning process
b) Recognition process
c) Verifying process

A. Processing Steps of OCR

In OCR text of scanned image is converted into text file .This process of converting is passes through various steps.

The basic processing steps

a) Scanning
b) Pre-processing,
c) Segmentation
d) Feature extraction/correlation calculation,
e) Classification and post-processing of OCR are discussed as following

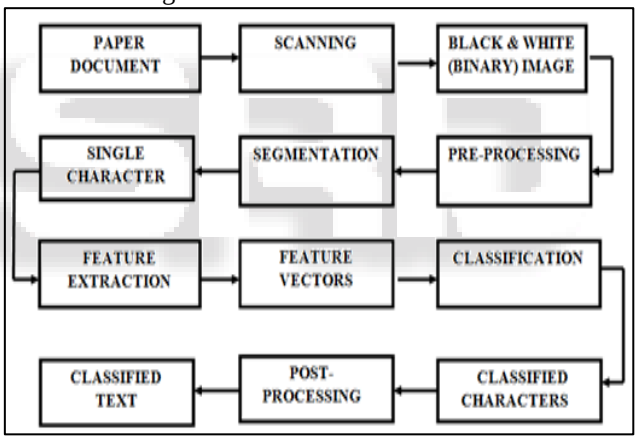

Fig. 1: Processing Steps of OCR System

II. METHODOLOGY

The anticipated structure for detection of multi-scripts such as English text, English & Gurumukhi Numerals and Gurumukhi or Punjabi text is illustrated in this. The categorization of multi-scripts is major task of present work and its explanation is given as below. Script Identification System Architecture is shown in Figure 3.1 and this plan to be aware of the printed transcript of Arial font for English and GurbaniKalmi font of Gurmukhi language. The text of the scanned image is converted into text file and it includes some processing steps. The processing steps tracked by the arrangement are data acquisition, pre-processing, segmentation and recognition. Pre-processing is used for scanning, clipping, removal of noise, removal of unnecessary small objects and normalization. After preprocessing text lines are segmented and then these lines are segmented into words. The segmentation of English text is different from Gurumukhi text because the words of Gurumukhi language are connected with a headline. So, in word segmentation it

includes different techniques for English word segmentation and Punjabi word segmentation. When words are segmented then individuals characters are segmented from the words. Thus segmentation is the heart of the Multilanguage optical character recognition system. In recognition correlation value is calculated and the one which have highest correlation value is recognized as character and write into the text file.[12][11]

a) The main steps of system are:
b) Data Acquisition
c) Pre-processing
d) Segmentation
e) Recognition

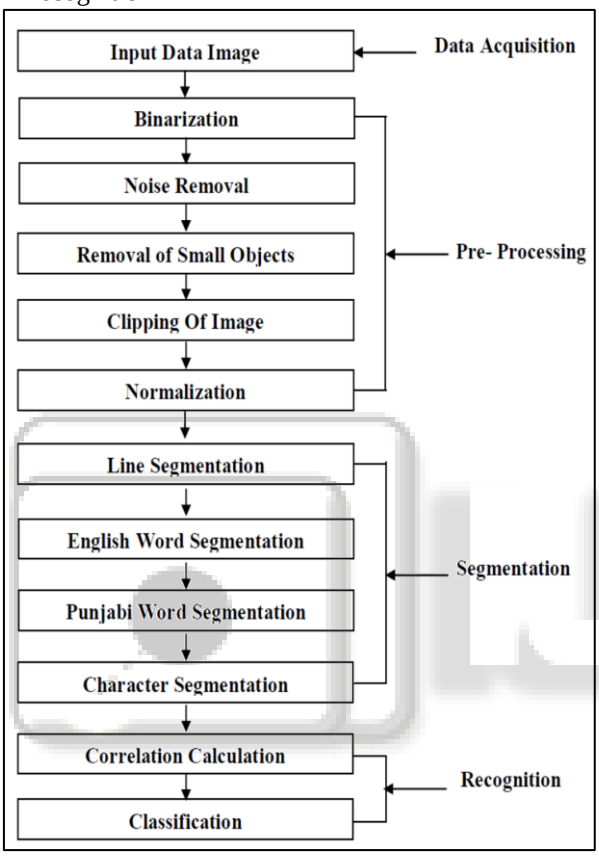

Fig. 2: Script Identification System Architecture

III. RESULTS & DISCUSSION

A GUI is built for an OCR system to recognize the script in multi languages the Multi-script recognition system as shown in figure 3. Arial Font is used for database of English script characters and GurbaniKalmi font is used for the database of Punjabi script characters to train the system. Testing samples of various sizes were prepared to test the efficiency of the system for both the scripts. Before passing the test sample user selects the type of the language to be recognized and then testing sample is passed to the system and process for the recognition. After recognizing the sample it is displayed in the edit text box of the GUI and stored in the text file. After that for the purpose of the efficiency calculation, user has to update the status of the recognized sample as right or wrong. The system will not take any further input sample until status is not updated. When system has updated the status of recognized sample then system is ready for further use. When user wants to calculate the efficiency of the system then "Calculate Efficiency" button can be used to calculate it.

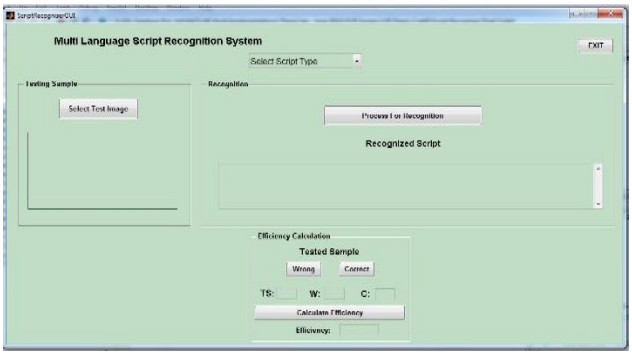

Fig. 3: Multi-Script Recognition System - GUI

A. *Simulation Results*

The proposed system is tested on 500 samples containing single characters, single words, single lines and multi lines of both the languages. Out of these 496 samples were segmented & recognized perfectly and in 4 samples of single line a single letter is not recognized correctly. The results of processes such as segmentation, features extraction and recognition of Punjabi and English text and numerals are shown as below.

B. *Simulation Results for Punjabi Text*

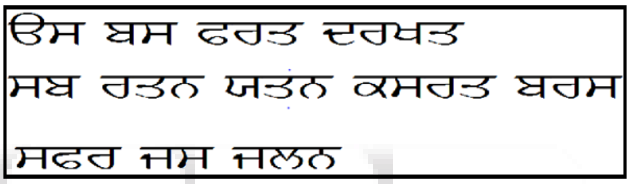

Fig. 3.1: Input Image of Punjabi Text

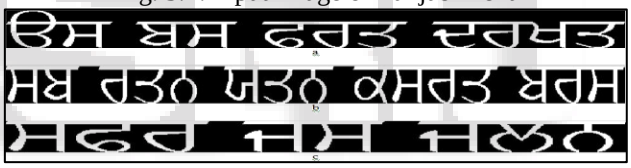

Fig. 3.2: Line Segmentation of Punjabi Text (a) First Line (b) Second Line (c) Third Line

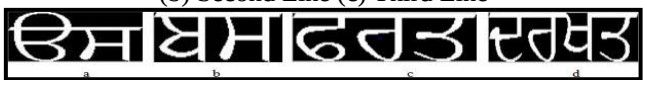

Fig. 3.3: Word Segmentation of first line of Punjabi Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word

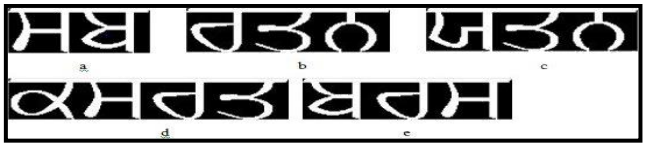

Fig. 3.4: Word Segmentation of second line of Punjabi Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word (e) Fifth Word

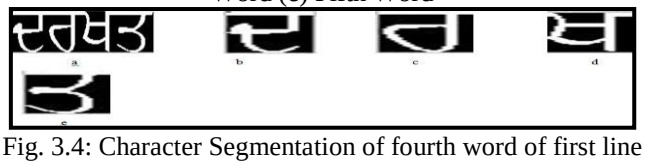

Fig. 3.4: Character Segmentation of fourth word of first line of Punjabi Text (a) Fourth word of First Line (b) First Character (c) Second Character (d) Third Character (e) Fourth Character

C. Simulation Results for English Text and Numerals

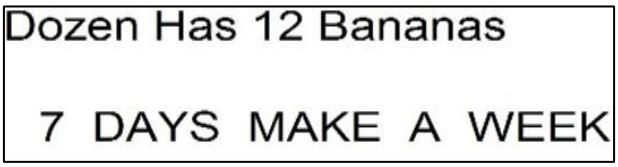

Fig. 3.5: Input Image of English Text and Numerals

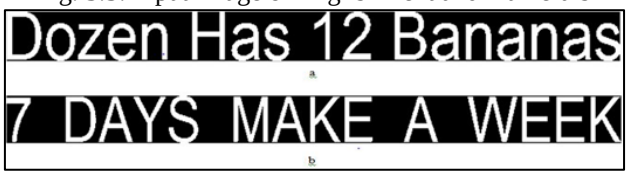

Fig. 3.6: Line Segmentation of English Text (a) First Line (b) Second Line

Fig. 3.7: Word Segmentation of first line of English Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word

Fig. 3.8: Word Segmentation of second line of English Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word

Fig. 3.9: Character Segmentation of first word of first line of English Text (a) Fourth word of First Line (b) First Character (c) Second Character (d) Third Character (e) Fourth Character (f) Fifth Character

Fig. 3.10: Character Segmentation of second word of first line of English Text (a) Fourth word of First Line (b) First Character (c) Second Character (d) Third Character

The proposed system gives the good results but during recognition process, sometimes it get confused with uppercase and lowercase English alphabets of similar structure such as S with s, W with w, O with o.

IV. Conclusion

In the proposed work a simple and efficient method for recognizing the multi-scripts such as English and Punjabi text is explained. The main motive of this proposed method is to provide the multi-script recognizer which is capable to recognize more than one scripts as English, Punjabi text and Numerals. The system is trained for English text and numerals using Arial font and for Punjabi text and numerals using GurbaniKalmi font. The number of holes feature is extorted from the segmented characters of any above cited scripts which were segmented using proposed segmentation algorithm. In this system it has been observed that the segmentation of English word is difficult than Punjabi word segmentation due to confusion of space between the character and words. This problem is solved by calculating the maximum space length between characters. If the space length is less than 25 then it is consider as line has one word. On the basis of the number of holes characters are grouped and then correlation is found to take decision of any segmented character. The concept of grouping increases efficiency of the system and reduces time consumption of correlation matching. The experimental results show that the proposed method is efficient to recognize English and Punjabi text.The proposed system worked for complete English script but for Punjabi scripts without vowels. In future, the system can be trained for Punjabi script for upper and lower zone. This system can further improved to work on different fonts for both the English and Punjabi script.

References

[1] Arica N. and Yarman-Vural F. T. (2001) "An Overview of Character Recognition Focused on Off-Line Handwriting" IEEE Transactions On Systems, Man And Cybernetics—Part C: Applications And Reviews, vol. 31, no. 2.

[2] Bingyu C. and Chen Y. (2012) "Reduction of Bleed-through Effect in Images of Chinese Bank Items" Frontiers in Handwriting Recognition (ICFHR), IEEE.

[3] Charles P. K., Harish V., Swathi M., Deepthi CH. (2012) "A review on the various techniques used for Optical Character Recognition" International Journal of Engineering Research and Applications (IJERA), vol. 2, pp. 659-662.

[4] Chen X. And Yuille A. (2004)"Detecting and reading text in natural scenes", Computer Vision and Pattern Recognition, vol. 2.

[5] Cheung A., Bennamoun M., Bergmann N.W. (2001) "An Arabic optical character recognition system using recognition-based segmentation", Published by Elsevier Science Ltd. pp.215-233.

[6] Devijver P. A. and Kittler J. (1982) "Pattern recognition: A Statistical Approach", London: Prentice-Hall.

[7] Faaborg A.J. (2002) "Using Neural Networks to Create an Adaptive Character Recognition System", Cornell University, Ithaca NY.

[8] Fan X. and Fan G. (2009) "Graphical Models for Joint Segmentation and Recognition of License Plate Characters" IEEE Signal Processing Letters, vol. 16, no. 1.

[9] Fu K. S. and Mui J.K. (1981) "A Survey on Image segmentation", Pattern Recognition,vol. 13, pp.3–16.

[10] Hinton G. (1989) "A fast learning algorithm for deep belief nets" Neural Computation, vol. 18, no. 7, pp.1527–1554.

[11] Sankaran N., Jawahar C.V. (2012) "Recognition of printed Devanagari text using BLSTM Neural Network" Pattern Recognition (ICPR), 21st International Conference, IEEE, pp. 322 – 325.

[12] Shah P., Karamchandani S., Nadkar T., Gulechha N., Koli K., Lad K. (2009) "OCR- based Chassis-Number Recognition using Artificial Neural Networks", IEEE, pp. 31-34.

[13] Sharma D., Jain U. (2010), "Recognition of Isolated Handwritten Characters of Gurumukhi Script using

Neocognitron", International Journal of Computer Applications (0975 – 8887) vol.10, no. 8.

[14] Singla G. and Kumar P. (2013) "Extract the Punjabi Word from Machine Printed Document Images" International Journal of Engineering Research and Application vol. 3, Issue 5, pp.343-348.

