Sonam Jain et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 3, Sept 2017, pp. 153-159

Optical Character Recognition System for Multilanguage Script Recognition

Sonam Jain

Assistant Professor, Department of Computer Science & Engg., Delhi Institute of Technology and Management, Haryana

Abstract—**Optical Character Recognition system provides the renovation of scanned documents into editable form. The course of action of this scheme starts with scanning of input documents to digital image and translating colored or gray scale image into black and white 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. 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 Optical Character Recognition like pre processing, segmentation, feature extraction, correlation calculation and classification. After preprocessing by binarization and noise removal of the test image, image 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 character set of both the languages are analyzed and number of holes feature is decided to divide the whole character set into the groups on the basis of number of holes in any character. 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. Further 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. The experimental results of the proposed system shows the highest accuracy.**

Keywords— OCR, Multilanguage Script Recognition, Segmentation, Recognition

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. In these days many algorithms are available to recognize the characters. These existed algorithms can be grouped in various categories based on character recognition such as online character recognition, optical character recognition, Off-line character recognition magnetic character recognition, handwritten character recognition and printed character recognition. Optical Character Recognition fit in to pattern recognition which performs habitual detection using PC. It links a representative sense with image of printed or handwritten characters. The procedure of Optical Character Recognition initiate with scanning of input papers to digital image and renovating colored or gray scale 8 bit image into binary image. Then each character is move across segmentation practice and the resulted image of segmented character is fed into a pre-processor for removal of noise and normalization process. Some definite features are the mined from the character for recognition. The Optical Character Recognition system is able to recognize good quality documents with high accuracy. Some research papers on English and Punjabi have cited in literature survey. Authors described OCR based techniques only for recognition of one script. Hence these OCR based techniques are limited to only script recognition. So it is necessary to design an OCR based technique which is able to recognize multi-scripts such as English, numerals and Punjabi scripts. In present time lot of research has been done for printed and handwritten characters .So the present work is to develop multi-script OCR based system. This will help to handle the many documents of English and Punjabi scripts in government offices.

I. Related Work

The present work involves the Implementation of an Optical Character Recognition for Multilanguage Script Recognition. The available literature has been reviewed in this context.

Pal and Chaudhuri (2004) presented that rigorous study has been completed on OCR(Optical Character Recognition) and a huge amount of articles have been in print on this subject during the previous few decades. A lot of viable OCR systems now exist in the marketplace. However a large amount of these systems employ for Arabic, Roman Japanese and Chinese characters. Out of 12 major scripts in India there is not enough work done on Indian scripts. They presented the beginning and properties on Indian scripts. They also discussed unusual methods in OCR expansion as well as study effort completed on Indian scripts detection.

Lehal and Singh (2006) presented the complete multi-font and multi-size OCR (optical character recognition) system for Gurmukhi script. This paper shows the various stages for the development of complete Optical Character Recognition for

Gurmukhi script and described the potential solutions. OCR system has five main processing stages i.e

a) Digitization.
b) Pre-processing.
c) Segmentation.
d) Recognition.
e) Post-processing.

The accuracy of this system is 97%.

Shah et al. (2008) captured number from vehicle and then performed recognition of segmented characters of vehicle classification number. This paper presents OCR based method using artificial neural network for vehicle chassis number identification. This method used to test on several images of vehicle of diverse clarity. The tested images are captured form different distance and different views. Due to some facts the noise is added in image so passed it through pre-processing. In pre-processing involves the edge detection, segmentation, and normalization and thinning algorithms. Then noiseless segmented characters specified as input to ANN. The authors created the character of 28x18 size and organize in 504x1 input element vector form. The artificial neural network is trained for 0 to 9 numerals for recognition purpose. ANN gives the output on basis of 0.55, 0.65 and 0.75 threshold values. The conclusion of this paper is that at 0.75 threshold value achieved accuracy is 95.49% with zero wrong identification rates.

Rajashekararadhya et al. (2009) introduced a feature extraction method based on zoning and projection distance metric for extracting the features of Kannada numerals. The character image of 50x50 pixels is grouped in to 25 equivalent zones each of 10x10 pixels. Then one feature is extracted in vertical downward direction for each zone by evaluating column average. So10 features are extracted from zone with the repetition of this process for all columns of zone. Repeat the same process for all directions such as horizontal right direction, vertical upward direction and horizontal left direction. Hence 40 features are extracted by calculating column average pixel distance in four directions for every zone. At last total extracted features for 25 zones are 1000 used for identification of numerals. Zone may have empty centre pixels in row or column. The value of this type of empty pixels is zero in feature vector. K-nearest neighbor classified the input character by using the features which was extracted with above cited method. So 97.8% recognition accuracy was achieved for Kannada numerals.

Hamidreza Kasaei et al. (2010) depicted an actual period and a method for recognition and detection of car number plate. Tracing of number plate is done by using morphological operator and template matching is used for character recognition.

Rani et al. (2011) identification of bi-lingual or multi-lingual documents is most difficult task in Optical Character Recognition system. In this paper Rajneesh Rani, Renu Dhir made a system for classification of bi-lingual document containing English numerals and Punjabi characters by using Gabor features. Usually, such systems are developed by using two types of methods. One method is combined database method. That is this method contains complete database of all scripts which are printed on the document. So at recognition level large amount of database is available for recognition of single character. In the second method the script of every character is identified before recognition of characters. This made recognition of character easy task due to identified script. The various techniques are available for identification of printed and handwritten script which can be grouped into four categories as connected component analysis, text block level analysis, text line level analysis and word and character level analysis. The support vector machine classifier was used for classification of bi-lingual documents using 140 features extracted by Gabor Filter. The support vector machine classifier obtained 99% recognition accuracy with Polynomial Kernel functions.

Charles and Harish (2012) reviewed the various classification techniques based on the Optical Character Recognition to recognize printed and handwritten English character. The correlation method and neural network classifier was used for single character and continuous character recognition. In the correlation method for recognition of single character the input image passed through pre-processing, segmentation to locate text in image and recognition performed with correlation. For recognition the segmented image was correlated with all trained templates and find maximum correlation value and the character corresponding this value is acknowledged as classified character.

Mithe and Indalkar (2013) merged the functions of OCR and speech synthesizer. The goal of this method is to present user interface application which make text image to speech translation scheme with the use of android phones. The text image specified as input to OCR and OCR acquire text from this input image and then translate text into speech using speech synthesizer. The OCR performs

a) Scanning
b) Pre-processing.
c) Segmentation.
d) Feature extraction
e) Recognition
f) Store classified text.

The presented OCR system is valuable in various applications such as bank, companies, offices etc. This system is primarily presented for blind people. Recognition of characters is done using the Tersseract which is open source of Optical Character Recognition and gives more accurately classified text.

II. PRESENT WORK

A. Problem Formulation

OCR has steps forward in previous five decades from inimitable exceptional rationale reader to the versatile assembly and imposing on-line schemes. Due to this encroachment information capture cost reduced and has made the scheme more steady and precise. The literature on many different Optical Character Recognition methods for recognition of English, Numeral and Gurumukhi has been surveyed. Then problem is formulated from literature survey by observing the different script recognition methods. In literature we have observed that till now the work has been done for recognition of single script. But it is a challenge to develop the OCR for recognition of multi-scripts. Hence recognition of multi-scripts is a goal of recent research. This is a major motivation for the work. Every script has different structure and different characteristics. Due this reason methods of recognition of different scripts are also different. In present work system can recognize English text, Punjabi text, English Numerals and Punjabi Numerals. The segmentation of English text is different from the segmentation of Punjabi text due to different structure. But the common features one is statistical feature (Projection Histogram) and other is structural feature (Number of holes) are extorted for recognition of these scripts.

B. Objectives

The main research objective is to create a novel single platform for Multi-Script Recognition System. The research work aims to design a script recognition system which will capable to recognize two different languages such as English and Punjabi i.e. the designed platform will recognize more than one language to reduce the office work load where more than one language is official language.

It follows that the objectives of this work are to:

1) To design the English Language Script.
2) To design script recognizer system for Gurumukhi language.
3) To make multi language script recognition system.
4) To calculate the system efficiency to recognize the correct samples.

C. Present Work

The anticipated structure for detection of multi-scripts such as English text, English & Gurumukhi Numerals and Gurumukhi or Punjabi text is illustrated in this. Script Identification System Architecture is shown in Figure1 and this plan to be aware of the printed transcript of Arial font for English and GurbaniKalmi font of Gurumukhi 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 pre-processing 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. The segmentation of English words is difficult as compare to segmentation of Gurumukhi words.

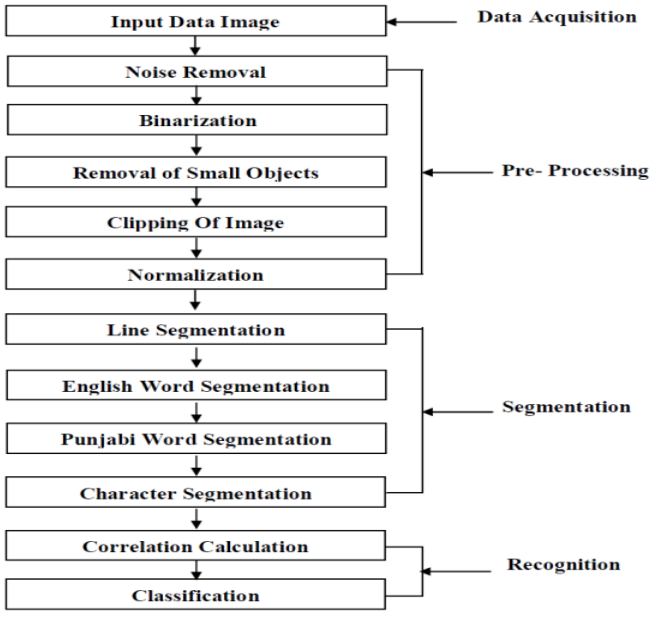


Figure 1: Script Identification System Architecture

The reason for this is that the English word enclose the uncouple characters which creates the confusion between character's gap and word gap. The segmentation steps include line segmentation, word segmentation and character segmentation

A. Line Segementation

Line segmentation is process which fragmented the text into lines. The lines are detached from one another by some zero pixels between them. The line segmentation is carry out by horizontal projection of input image. For this function the input image is scanned first in downward direction and after that in upward direction to dig out the lines until the line with zero pixels are not find. If the lines are well separated and not tilted, the horizontal projection will have well separated peaks and valleys. These marked valleys are very helpful to find out site of the line confines. As follows the lines are detached from one another.

a) Form the horizontal projection profile of image.
b) Find the line margins with the help of this profile.
c) First line is stated as first line and rest all lines are stated as remaining lines.
d) Then repeat this practice for remaining lines till all the lines are separated.

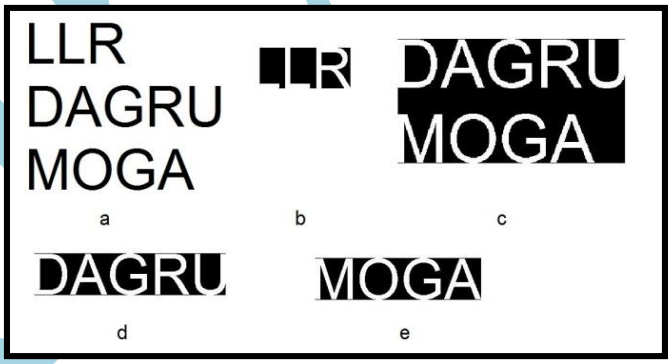


Figure 2: Segmentation of Lines of English Script (a) Input Image (b) First Line (c) Remaining Lines (d) First Line from Remaining Lines (e) Remaining Line

Figure 3: Segmentation of Lines of Punjabi Script (a) Input Image (b) First Line (c) Remaining Lines (d) First from Remaining Lines (e) Remaining Line

B. Word Segementation

Now the function of word segmentation is drag out the margins of the words from these sliced lines. Normally words are unconnected from each other with tolerable extent of gaps. Due to this inspection word segmentation does not stay on a complicated problem to solve. Word segmentation is achieved by vertical projection profile.

a) Vertical projection profile of line image is fashioned.
b) Then locate the borders of word.
c) Initial word fixed as first word and other words are fixed as remaining word.
d) Replicate the course for other words to disconnect the words.

English words segmentation is complicated than Gurumukhi words segmentation.

English Words Segmentation: The English word segmentation is also completed by vertical histogram projection, but it is compulsory to differentiate the intra character gap from word gap. For this purpose first the difference between the spaces are calculated. Then find maximum space whose difference is not equal to one. This maximum space noted as space length or word

space. If maximum space length is less than 25 it means there is only one word in the line.

Figure 4: Segmentation of English Words (a) Input Image (b) First Word (c) Remaining Words (d) First Word from Remaining Words (e) Remaining Word

Punjabi Words Segmentation: The characters in Punjabi word are associated with each other by means of headline. These characters share the common pixels in the headline. Hence during the Punjabi word segmentation the researcher does not have to face the problem like the English word segmentation. Punjabi word segmentation is ended in the same manner as segmentation of line, vertical projection profile.

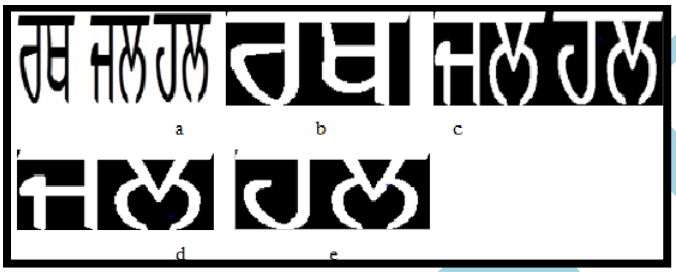
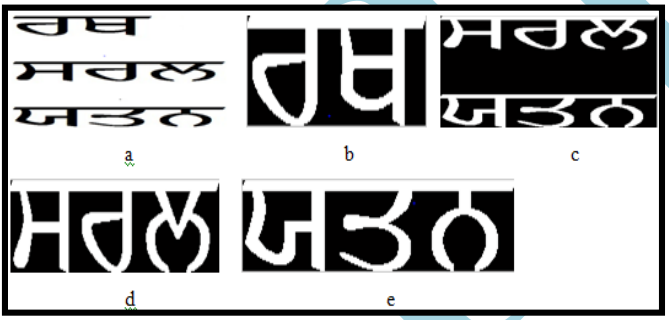


Figure 5: Segmentation of Punjabi Words (a) Input Image (b) First Word (c) Remaining Words (d) First Word from Remaining Words (e) Remaining Word

C. Character Segementation

Once the lines and words are sliced, the next task is to haul out the characters from these words. Character segmentation is compulsory for character recognition approaches which rely on the



isolated character. It is a key stage because most of the recognition errors are caused by erroneous segmentation of characters. Segmentation of characters plays an essential role in text recognition system.

Figure 6: Character Segmentation (a) Character size in scanned document (b) Character resize according to database

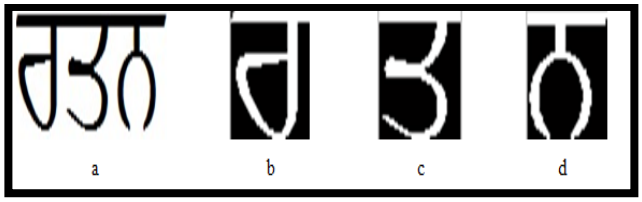


Figure 7: Segmentation of Punjabi Characters (a) Input Image (b) First Character (c) Second Character (d) Third Character

III. RESULTS

A GUI is built for an OCR system to recognize the script in multi languages the Multi-script recognition system as shown in figure 9. 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.

The four steps to use the designed system are shown in Figure 8,9,10, 11 and 12.

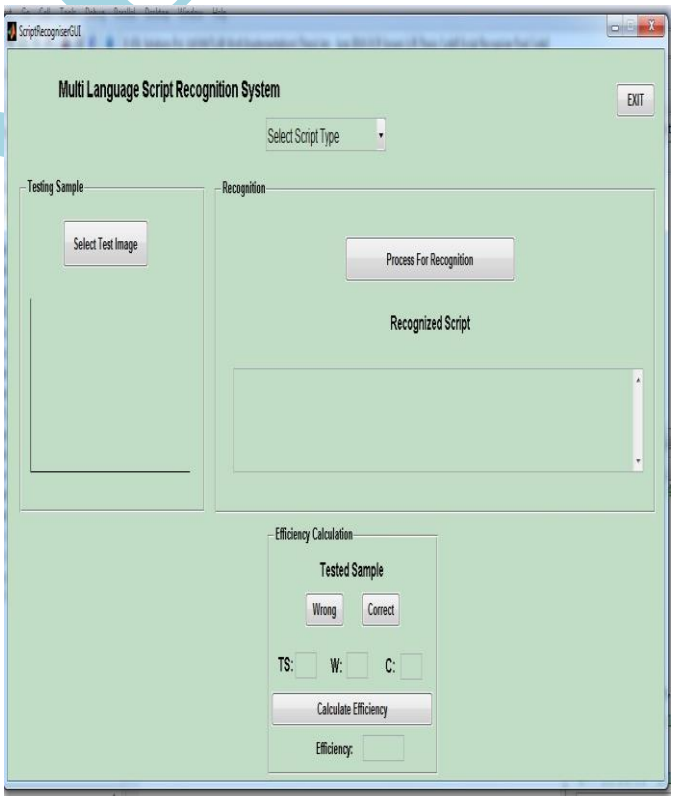


Figure 8: Multi-Script Recognition System – GUI

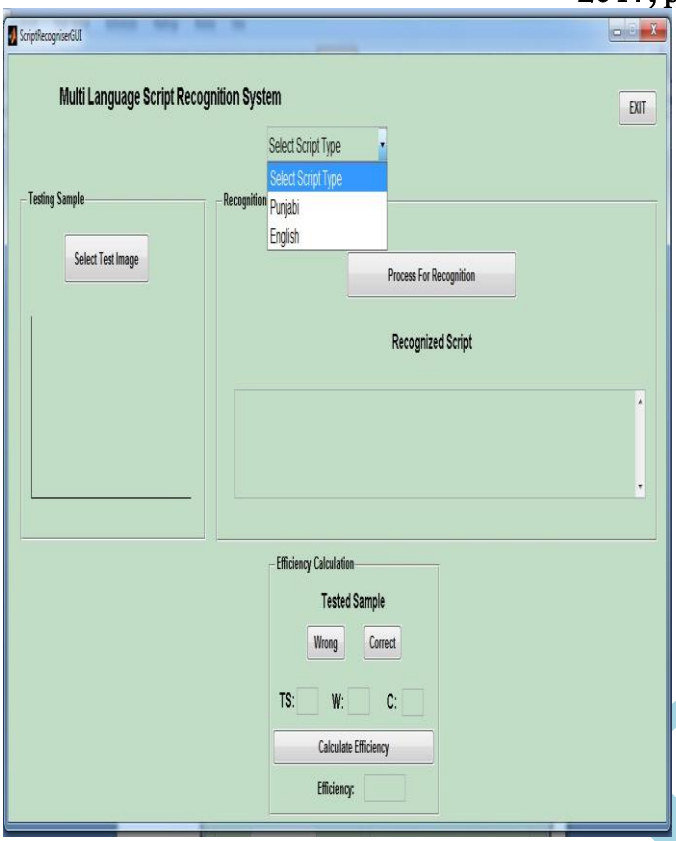


Figure 9: Step 1 of Multi language Script Recognition System

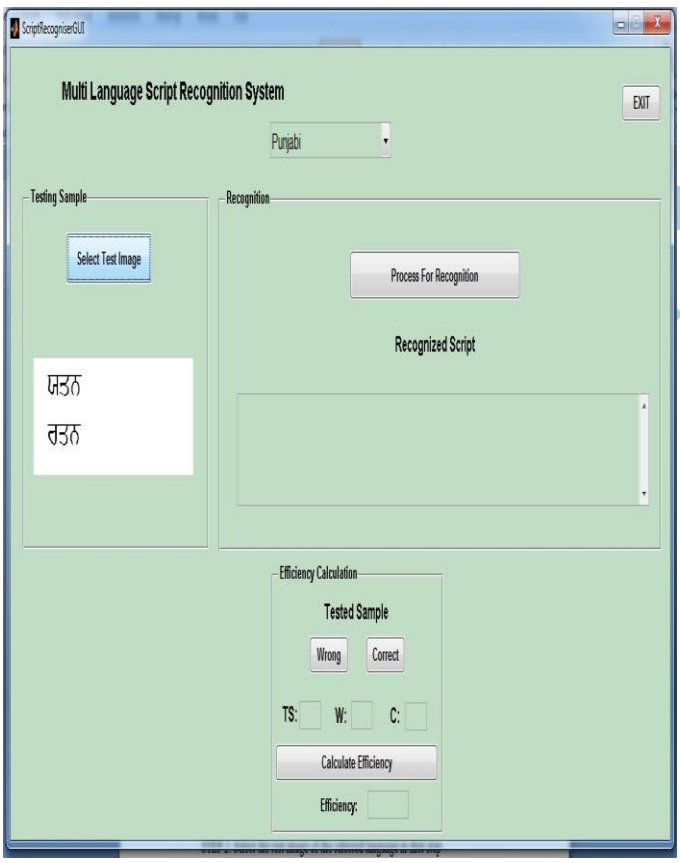


Figure 10: Step 2 of Multi language Script Recognition System

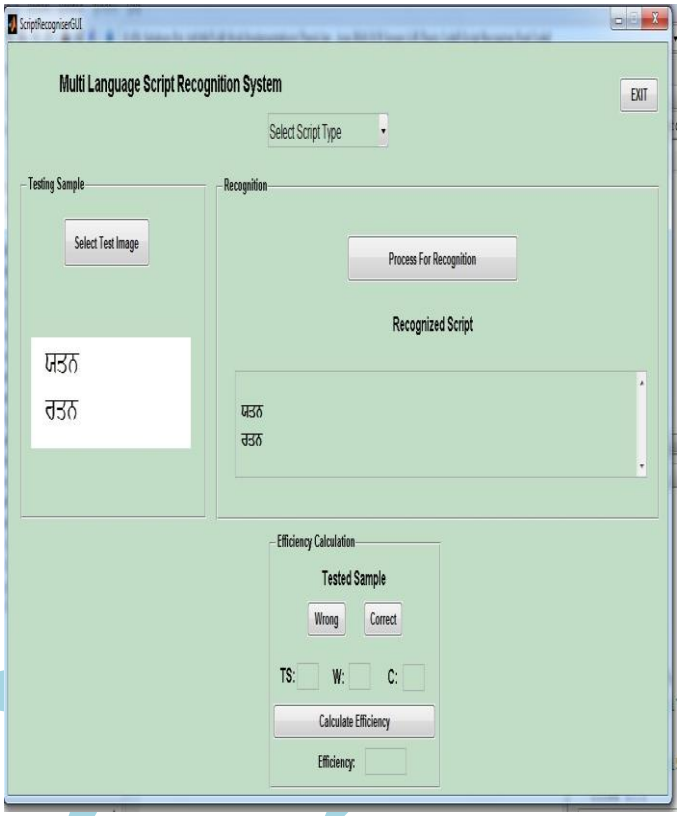


Figure 11: Step 3 of Multi language Script Recognition System

Figure 12: Step 4 of Multi language Script Recognition System

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.

1. Simulation Results for Punjabi Text

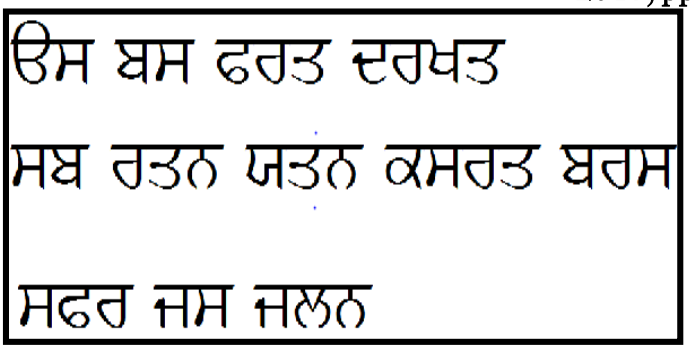


Figure 13: Input Image of Punjabi Text

Figure 14: Line Segmentation of Punjabi Text (a) First Line (b) Second Line (c) Third Line

Figure 15: Word Segmentation of first line of Punjabi Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word.

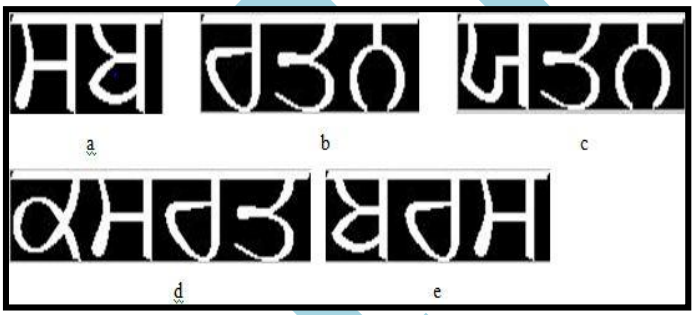


Figure 16: Word Segmentation of second line of Punjabi Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word (e) Fifth Word

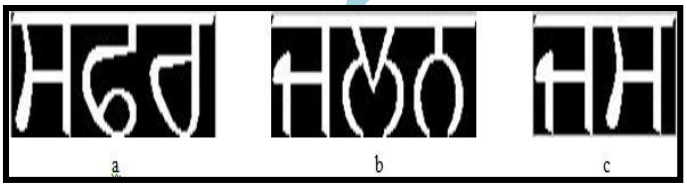


Figure 17: Word Segmentation of third line of Punjabi Text (a) First Word (b) Second Word (c) Third Word

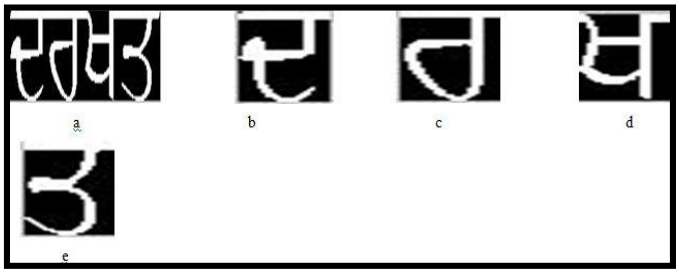


Figure 18: Character Segmentation of fourth word of first line of Punjabi Text (a) Fourth word of First Line (b) First Character (c) Second Character (d) Third Character (e) Fourth Character

2. Simulation Results for English Text and Numerals

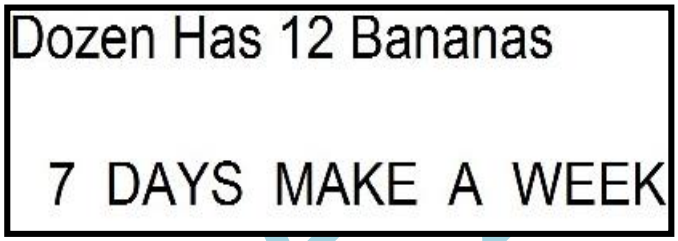


Figure 19: Input Image of English Text and Numerals

Figure 20: Line Segmentation of English Text (a) First Line (b) Second Line

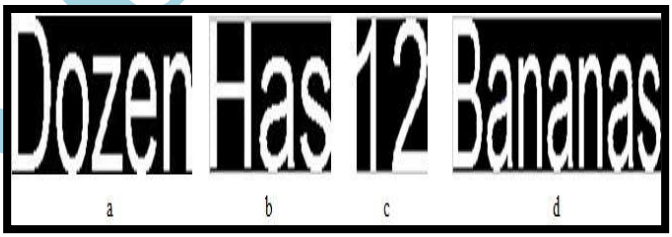


Figure 21: Word Segmentation of first line of English Text (a) First Word (b) Second Word (c) Third Word (d) Fourth Word

Figure 22: Word Segmentation of second line of English Text (a) first Word (b) Second Word (c) Third Word (d) Fourth Word

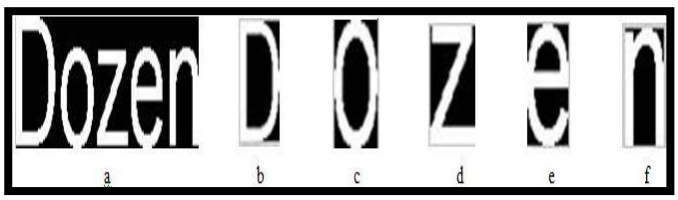


Figure 23: Character Segmentation of first word of first line of English Text (a) Fourth word of First Line (b) First Character (c) Second Character (d) Third Character (e) Fourth Character (f) Fifth Character

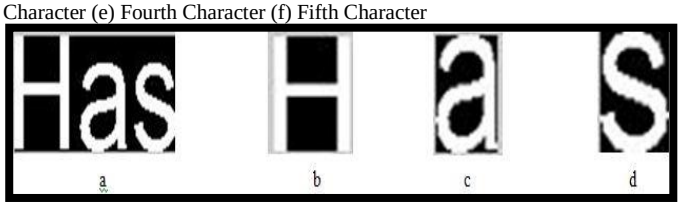


Figure 24: Character Segmentation of second word of first line of English Text (a) Fourth word of First Line (b) First Character (c) Second Character (d) Third Character

IV. CONCLUSIONS AND FUTURE WORK

A. Conclusions

In the proposed work a simple and efficient method for recognizing the multi-scripts such as English and Punjabi text is explained. The 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 shows that the proposed method is efficient to recognize English and Punjabi text.

B. Future Work

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.

REFRENCES

[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, pp. 216-233.

[2] Bingyu C. and Chen Y. (2012) "Reduction of Bleed-through Effect in Images of Chinese Bank Items", IEEE, pp. 174-178.

[3] Charles P.K., Harish V., Swathi M. and 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, pp. 366-373.

[5] Cheung A., Bennamoun M. and Bergmann N.W. (2001) "An Arabic Optical Character Recognition system using recognition-based segmentation", Published by Elsevier Science

[6] Kalas M.S. (2010) *"An Artificial Neural Network for Detection of Biological Early Brain Cancer",* International Journal of Computer Applications, vol. 1, no. 6, pp. 17-23.

[7] Kasaei S.H, Kasaei S.M. and Monadjemi S.A. (2010) "*New Morphology-Based Method for Robust Iranian Car Plate Detection and Recognition*", International Journal of Computer Theory and Engineering, vol. 2, no. 2, pp. 1793-8201.

[8] Kaur K. and Banga V. K. (2013) *"Number Plate Recognition Using OCR Technique",* IJRET, vol. 2, no. 9, pp. 286-290.

[9] Kumar R. and Singh A. (2010) "Detection and Segmentation of Lines and Words in Gurumukhi Handwritten Text", IEEE, pp. 353-356.

[10] Wong K.W., Chang S.J. and Lenug S.J. (2002) *"Handwritten Digit Recognition using Multi-Layer Feed forward Neural Networks with Periodic and Monotonic Activation Functions",* IEEE, vol. 3, pp. 106 - 109.

[11] Lehal G.S. and Singh C. (2001) *"A Technique for Segmentation of Gurumukhi Text",* Springer Berlin Heidelberg, pp. 191-200.

[12] Lehal G.S. and Singh C. (2002) "*A post-processor for Gurumukhi OCR"* Sadhana, vol. 27, pp. 99–111.

[13] Lehal G.S. and Singh C. (2006) *"A Complete Machine Printed Gurmukhi OCR System",* Vivek, vol. 16, pp. 10-17.

[14] Mithe R., Indalkar S. and Divekar N. (2013) *"Optical Character Recognition",* International Journal of Recent Technology and Engineering (IJRTE),vol. 2,no. 1,pp. 72-75.

[15] Pal U. and Chaudhuri B.B. (2004) *"Indian script character recognition: a survey",* Pattern Recognition Society. Published by Elsevier Ltd, pp. 1887-1899.

[16] Pan Y., Hou X. and Liu C. (2009) *"Text localization in natural scene images based on conditional random field"* , International Conference on Document Analysis and Recognition, IEEE, pp. 6-10.

[17] Patil V. and Shimpi S. (2011) *"Handwritten English character recognition using neural network"*, Elixir Comp. Sci. & Engg, vol. 41, pp. 5587-5591.

[18] Pradeep J., Srinivasan E. and Himavathi S. (2012) "Performance Analysis of Hybrid Feature Extraction Technique for Recognizing English Handwritten Characters", IEEE, pp. 373-377.

[19] Rajashekararadhya S.V. and Ranjan P.V. (2009) "Zone based Feature Extraction Algorithm for Handwritten Numeral Recognition of Kannada Script", IEEE, pp. 525-528.

[20] Rani R., Dhir R. and Lehal G.S. (2011) *"Identification of Printed Punjabi Words and English Numerals Using Gabor Features",* World Academy of Science, Engineering and Technology, vol. 5, pp. 317-320.

[21] Sharma D. and Jain U. (2010) *"Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron",* International Journal of Computer Applications, vol. 10, no. 8, pp. 10-20.

[22] Shing H.H. and Chan S.L. (1997) *"Hypertext-Assisted Video Indexing and Content- based Retrieval",* Proceedings of the eighth ACM conference on Hypertext, pp. 232-233.

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

[24] Taha S., Babiker Y. and Abbas M. (2012) *"Optical Character Recognition of Arabic Printed Text",* IEEE, pp. 235-240.

