Review paper on Implementation of Language Independent Rule Based Classification of Gurmukhi Printed & Handwritten Text

Jaswinder Kaur[1] Mrs.Rachna Rajput[2]
[1]M.Phil Student [2]Assistant Professor
[1,2]Gurukashi University, Talwandi Sabo

Abstract— Printed & handwritten text recognition is the part of pattern recoginition.In the era of computer science, it has become important to collect database for future use. This data may handwritten & machine written In current work the OCR(optical character reader) for Gurmukhi Punjabi text is performed. Basic aim is to differentiate between machine written text and hand written text. It includes two issues first is detecting the letters and then classifier will classify the machine written text and hand written text with the different parameters like as PSNR,SSIM,MSE.
Key words: OCR, HMC

I. INTRODUCTION

The focus of current research is to discriminate printed and handwritten entries in data entry forms written in any language. Data entry forms are too common in all kinds of enterprises to collect customer information. These information are filled by customers manually in English or Arabic handwriting. Such data entry forms are collected in substantial number on daily basis in these enterprises that needs an automatic processing system to extract this information from data entry forms for further processing /classification into Arabic/English. Form field separation is the basis of optical character recognition(OCR). Therefore, in automatic form processing, printed and script entries discrimination is mandatory aspect.

Handwritten/machine-printed classification (HMC) is the process of labeling an image containing text segments, in order to discriminate handwritten from machine-printed text. It has numerous applications, particularly in (improving) Optical/Intelligent Character Recognition (OCR/ICR), automatic document analysis and anonymization [1].

A. OCR (Optical Character Reader)

is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

There are number of difference between hand written and machine written text. These are given below:

- In machine, written language has less no. of strokes as compared to hand written text.
- Machine written text is more readable than the hand-written text.
- Machine written language has fixed sized text compare to the hand-written text.
- Machine written text has more options for formats compare to the hand-written text.
- Machine written text is equally spaced compare to the hand-written text.

II. LITERATURE SURVEY

1) Tanzila Saba(2015) et al: Handwriting in data entry forms/documents usually indicates user's filled information that should be treated differently from the printed text. Therefore, prior to segmentation & classification, text distinction into Printed & script entries is mandatory.
2) B. B. Chaudhuri (2015) et al: There are many types of documents where machine-printed and hand-written texts intermixed appear. Since the optical character recognition (OCR) methodologies for machine-printed and hand-written texts are different, it is necessary to separate these two types of text before feeding them to the respective OCR systems.
3) Surabhi Narayan(2012) et al: Discrimination of handwritten and machine printed text in a scanner document image is an important phase as processing and recognizing machine printed and handwritten text cannot be done using a single OCR. Uniformity has been depicted from the transitions occurred due to the overlay of component structures on the null background.
4) Mrs.Saniya Ansari (2015) et al: Online handwritten character recognition is having wide areas of application in real life environment. Therefore the accuracy of such systems should be more, efficient and faster to process applications. The feature extractor determines which properties of the preprocessed data are most significant and should be used in further stages. In this paper different feature extraction methods are discussed and presented related with Devnagari script and proposed efficient and optimized extraction method with their comparative analysis.
5) U. Pal(2001) et al: there are many types of documents where machine printed and hand written texts intermixed appears. Since the optical character recognition methodology for machine –printed and hand-written texts are different , to achieve optimal performance it is necessary to separate these two types of texts before feeding them to their respective OCR. System. In this paper they have done this for bangle and devnagari.
6) Ergina Kavallieratou(2004) et al: in this research paper they have given the approach that able approach to discriminate between machine-printed and handwritten text. An integrated system able to localize text areas and split them in text-lines is used. A set of simple and easy to- compute structural characteristics that capture the differences between machine-printed and handwritten text-lines is introduced.

III. PROBLEM DEFINITION

Handwriting in data entry forms/documents usually indicates user's filled information that should be treated differently from the printed text. In Arab world, these filled information are normally in English or Arabic.but now Gurmukhi is also available for filling important data so Gurmukhi become need of databases in a number of case related with specific areas .Secondly, classification approaches are quite different for machine printed and script. Therefore, prior to segmentation & classification, text distinction into Printed & script entries is mandatory.

IV. METHODOLOGY

1) Step 1 input the image from any of the outer source.
2) Step 2 Convert the image in to the matrix of rows and columns.
3) Step 3 convert the RGB image to Gray image.
4) Step 4 Identify the threshold of the image based on OTSU technique.
5) Step 5 Identify the Labels of the thresholded image.
6) Step 6 extract the text character based on labels of the image.
7) Step 7 segmented the text characters from the image.
8) Step 8 Identify the aspect ratio of each section of character to identify the machine written and hand written text.
9) Step 9 Show the machine written text as label of 0 and handwritten text as 1

REFERENCES

[1] Tanzila Saba, and A Nikolaidis. "Language Independent Rule Based Classification of Printed & Handwritten Text". Proceedings of the IEEE 2015 tenth workshop on Multimedia Signal Processing, pp.393-398, 2015.
[2] B. B. Chaudhuri. "Automatic Separation of Machine-Printed and Hand-Written Text Lines ", Pattern Recognition Letters ,2015.
[3] Surabhi Narayan, and A Nikolaidis. "Discrimination of handwritten and machine Printed text is Scanner document Images based on Rough Set Theory". Proceedings of the IEEE 2012 tenth workshop on Multimedia Signal Processing, pp.393-398, 2012.
[4] Mrs.Saniya Ansari. "Optimized and Efficient Feature Extraction Method for Devanagari Handwritten Character Recognition", Pattern Recognition Letters ,2015.
[5] U.Pal. "Machine-printed and hand-written text lines identification". Proceedings of the IEEE 2001.
[6] Ergina Kavallieratou, and Stathis Stamatatos. "Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics". Proceedings of the IEEE 2004.
[7] Purnendu Banerjee, and A Nikolaidis. "A System for Hand-Written and Machine-Printed Text Separation in Bangla Document Images". Proceedings of the IEEE 2012
[8] Konstantinos Zagoris, and Ioannis Pratikakis. "Automatic Classification of Handwritten and Printed Text in ICR Boxes". Proceedings of the IEEE 2014.
[9] Abhishek Jindal, and Mohd Amir. "Language Independent Rule Based Classification of Printed & Handwritten Text". Proceedings of the IEEE 2015 tenth workshop on Multimedia Signal Processing, pp.393-398, 2015.
[10] Ranjeet Srivastava, and Ravi Kumar Tewari. "Separation of Machine Printed and Handwritten Text for Hindi Documents ". Proceedings of the IEEE 2015
[11] Lincoln Faria da Silva, and Angel Sanchez. "Automatic discrimination between printed and handwritten text in documents". Proceedings of the IEEE 2009 .
[12] A. Saïdani and A. Kacem Echi. "Identification of Machine-printed and Handwritten Words in Arabic and Latin Scripts". Proceedings of the IEEE 2013.

