A Technique for Classification of Printed & Handwritten text

Manpreet Kaur
M.Tech Research Scholar, Computer Engineering Department,
Yadavindra College of Engineering, Punjabi University,
Guru Kashi Campus, Talwandi Sabo (Punjab), India

ABSTRACT

Machine printed and handwritten words are sometimes mixed in a single document like in data entry forms. Since the algorithms for recognition of machine-printed and handwritten text are based on different techniques, so it is necessary to separate between these two types of texts before feeding it to respective optical character recognition systems. This separation will definitely increase the performance and overall system quality. Handwritten/machine-printed classification is the process to discriminate handwritten from machine-printed text and is a challenging task. It includes two issues first is detecting the letters and then classifier will classify the machine written text and hand written text. The proposed techniques is based on structural features of text i.e. aspect ratio of the text is calculated to discriminate between handwritten and machine printed text and results show the effectiveness of proposed approach.

Keywords: OCR, machine- printed text, handwritten text.

INTRODUCTION

The focus of current research is to discriminate printed and handwritten text written in any language. The Optical character recognition (OCR) technology employs different techniques to recognize either machine-printed or handwritten text, as classification approaches are quite different for machine printed and handwritten text. Therefore, separation of machine-printed and handwritten text has to be carried out before feeding text to respective module of OCR. The various terminologies used in current research are explained as under:

OCR is the 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.

MACHINE-PRINTED text includes the materials such as books, newspapers, magazines, documents and various writing units in the video or still image. Machine printed characters are uniform in height, width and pitch assuming the same font and size are used.

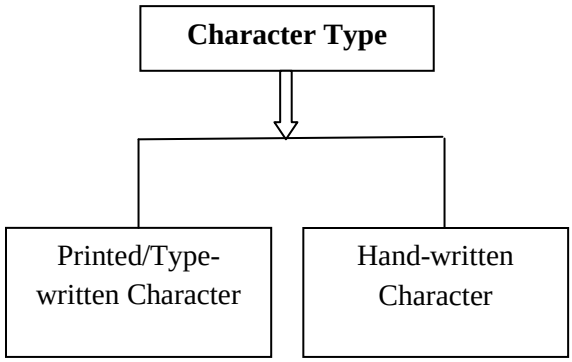

Fig.1. Character Type

HANDWRITTEN text can be divided into two categories: cursive and hand printed script. Characters are non-uniform and can vary greatly in size and style. Even characters written by the same person can vary considerably.

HMC Handwritten/ machine-printed classification is the process of labeling an image containing text segments, in order to discriminate handwritten from machine-printed text [1].

LITERATURE SURVEY

Tanzila Saba et al. (2015) [1]: This paper has presented a technique to classify the multilingual text block/text lines of data entry form into handwritten or printed text. This paper used a technique to explore new statistical and structural features of text lines to classify them into separate categories. Accordingly a set of classification rules is derived to explicitly differentiate machine printed and handwritten entries, written in any language.

Ranjeet Srivastava et al. (2015) [2]: This paper proposed an approach for separation of machine printed and handwritten text in Hindi documents. Statistical and structural features are used to distinguish between these two texts.

Abhishek Jindal et al. (2014) [3]: This paper proposed a technique to classify the handwritten and machine printed characters inside the intelligent character recognition cells. In future, to increase the recognition accuracy, other features such as stroke width and pixel distribution can be integrated with the proposed classification technique.

A. Saïdani et al. (2013) [4]: This paper elaborates the Different sets of features have been employed successfully for discriminating between Arabic and Latin words. They include few well-established features previously used and adapted in new structural features which are intrinsic features of Arabic and Latin scripts.

Konstantinos Zagoris et al. (2013) [5]: This paper used the Bag of Visual Words model (BoVW). Initially, blocks of interest are detected in the document image. For each block, a descriptor is calculated

based on the BoVW. The final characterization of the blocks as Handwritten, Machine Printed or Noise is made by a decisions scheme which relies upon the combination of binary SVM classifiers.

RESEARCH METHODOLOGY

The proposed technique has been implemented using image processing toolbox of MATLAB. Firstly the text image is scanned, then image is pre-processed for de-noising and thresholding is carried out. The text components in image are labelled and segmented. Aspect ratio of segmented characters is calculated to arrive at decision of machine-printed or handwritten text. The technique follows the steps as explained below:

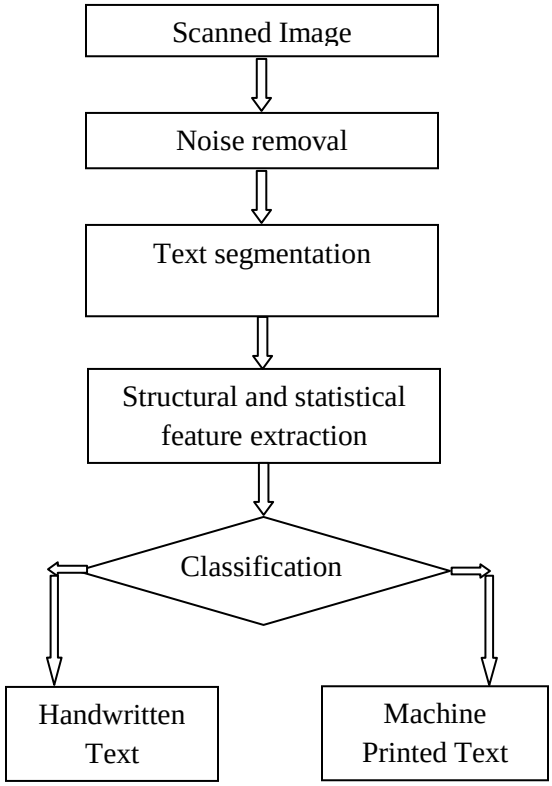

Fig.2. Flow diagram proposed scheme

Step 1: Input the source image and read as matrix of rows and columns.
Step 2: Convert the RGB image to Gray image.
Step 3: Threshold of the image using otsu technique.
Step 4: Identify the Labels of the thresholded image.
Step 5: Extract the text character based on labels of the image.
Step 6: Segmented the text characters from the image.
Step 7: Identify the aspect ratio of each section of character to identify the machine written and hand written text.
Step 8: Show the machine written text as label of 0 and handwritten text as 1.

RESULTS

The images with both types of text like hand written and machine written text are given as input. The input images to proposed system are shown in Figure 3 and Figure 5. The system is able to identify the machine printed and handwritten text correctly as shown in Figure 4 and Figure 6.

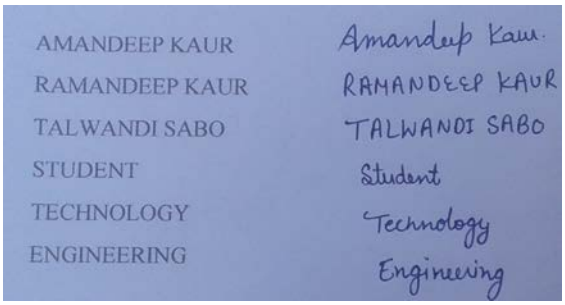

Fig.3. Input Image1

Fig.4. Output Image1

The Machine written text is shown as 1 and hand written text is shown as 0. In another input as shown below, results are also correct.

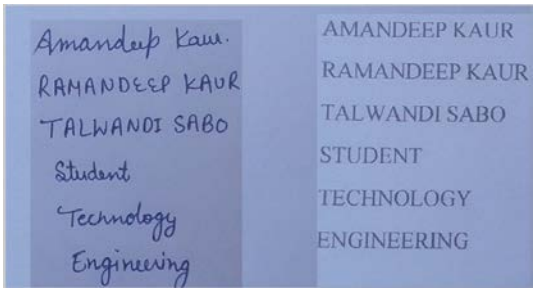

Fig.5. Input Image2

Fig.6. Output Image2

Conclusion and Future Work

An efficient technique for separation of machine printed text and handwritten text for english documents has been presented and its performance assessed. The approach is robust and fast enough as well as independent of language. Numbers of randomly selected text blocks (machine printed and handwritten in English) are tested. Experimental results of the approach are reliable for classifying machine printed and script in English language.

India is a multi-lingual country where a document page may contain more than one language scripts. In Future, this approach can be tested and extended to other Indian scripts like Gurmukhi.

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] Ranjeet Srivastava, and Ravi Kumar Tewari. "Separation of Machine Printed and Handwritten Text for Hindi Documents" Proceedings of the IEEE 2015

[3] 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.

[4] A. Saïdani and A. Kacem Echi. "Identification of Machine-printed and Handwritten Words in Arabic and Latin Scripts". Proceedings of the IEEE 2013

[5] Konstantinos Zagoris, and Ioannis Pratikakis. "Automatic Classification of Handwritten and Printed Text in ICR Boxes". Proceedings of the IEEE 2014.

[6] 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.

[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] Lincoln Faria da Silva, and Angel Sanchez. "Automatic discrimination between printed and handwritten text in documents". Proceedings of the IEEE 2009 .

[9] Ergina Kavallieratou, and Stathis Stamatatos. "Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics". Proceedings of the IEEE 2004.

[10] U.Pal. "Machine-printed and hand-written text lines identification". Proceedings of the IEEE 2001.