ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLASSIFICATION OF PRINTED AND HANDWRITTEN TEXT: A REVIEW

Manpreet Kaur*[1], Balwinder Singh[2]
*[1]M.Tech. (Computer Engineering), Yadavindra College of Engineering, Talwandi Sabo, India
[2]Assistant Professor (Computer Science), Yadavindra College of Engineering, Talwandi Sabo, India

DOI: 10.5281/zenodo.848620

ABSTRACT
Separating handwritten and machine printed text from a document has many applications. Various types of documents like bank cheques and forms etc. are used in daily life which contains both handwritten as well as printed text. It is necessary to separate handwritten and machine printed text before processing it with optical character recognition system. Various strategies are used to discriminate between handwritten and printed text. Many methods are specific to a particular script and others can handle different type of scripts. The various steps to carry out this discrimination are performed in this sequence: scanning, pre-processing, feature extraction and classification. In this paper, an effort has been made to review the various techniques for discriminating handwritten and machine printed text.

KEYWORDS: OCR, discrimination, printed text, handwritten text

I. INTRODUCTION
Optical character recognition (OCR) system converts the scanned document image into machine editable form. The recognition techniques for handwritten and machine printed text are entirely different. It becomes necessary to isolate handwritten and printed text in mixed documents (Figure 1) before processing it with respective OCR systems.

Handwritten and printed characters havedifferent properties, which can be extracted from its features. Printed characters have fixed height and same width, equal spacing and regularity in writing [2]. But handwritten text height and width vary as per writer's style of writing.

Figure 1: Handwritten and Machine printed text (mixed character) form

Figure 2 shows the overview of automatic handwritten and printed text classification system [11]. The system takes scanned document image as input and performs binarization using a standard algorithm. During binarization, the grayscale image is converted into a two tone (black and white) image. Pre-processing also involves noise removal through morphological or filtering operations.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

The system performs text segmentation either at line, word or character level. After segmentation, structural or statistical features are extracted using various classification techniques. Structural features such as number of strokes or loops in characters are based on topology and structure of text script [1]. While statistical features are based on calculation of moments or standard deviation of stroke width etc. [2]. The classifiers such as Pyramid histogram of oriented gradients (PHOG) [4], Support vector machine (SVM) [8], k-nearest neighbors (KNN) algorithm and Radon transform [10] are used to separate the handwritten and machine printed character.

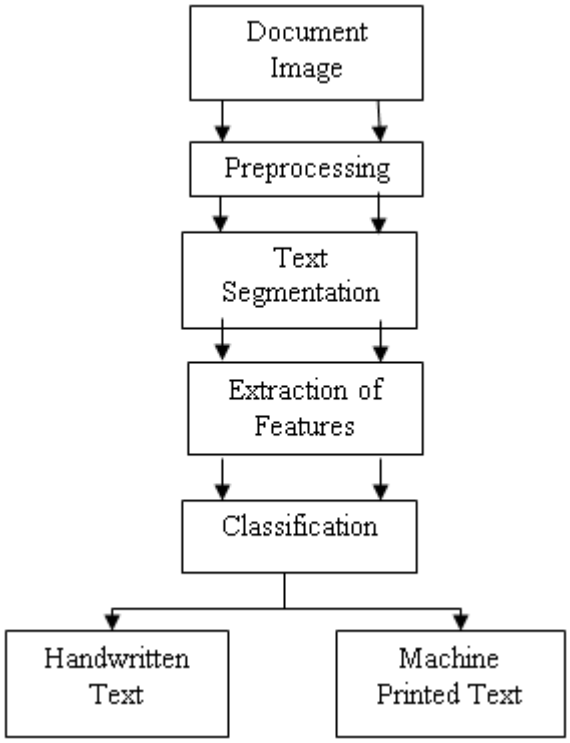

Figure 2: Handwritten and printed text classification system [11]

The paper is organized as follows. Section 2 discusses the relevant literature. Section 3 presents the conclusion and future scope of study.

II. LITERARURE SURVEY

Saba et. al [1] presented a language independent technique to discriminates the handwritten and printed text from data entryforms using new statistical and structural features of text lines. The method achieves 90% accuracy by employing simple classification rules and without using any training data. The technique has been tested on Arab and English scripts and is works for various font styles and sizes.

Srivastava et. al [2] proposed an approach for separation of printed and handwritten text for Hindi documents using structural and statistical features. The technique performs line, word and character segmentation and use stability in ratio of height of upper zone to middle zone and ratio of height of middle zone to lower zone for classification. The system achieved overall accuracy of 94.1%.

Jindal et. al [3] proposed a technique to classify the handwritten and printed characters inside form boxes. The classification is based on various features such as inter-character gap, height and identifying baseline of valid characters in one zone. The authors reported overall accuracy of 91%

Saïdaniet. al [4] elaborates a technique for identification of script (Arabic-Latin) and discrimination of handwritten printed text at word level. The authors used Pyramid histogram of oriented gradients (PHOG) features along with Bayes based classifier to achieve an accuracy of 98.3 % on standard database. In another work for identical scripts, same authors reported [5] an accuracy of 98.4% at word level with Bayes classifier by using existing features (connected component height, width, profile analysis, vertical projection variance and loop aspect ratio) and newly proposed features (bottom diacritics, loop position and elongate descender).

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Zagoriset. al [6] proposed an approach to separate handwritten from printed text using Bag of Visual Words (BoVW) model in three stages. Initially, area of interest has been identifiedusing page segmentation. Then, block descriptor has been calculated based on BoVW and final classification has been carried out by combination of SVM classifiers.

Narayan et. al [7] presented a novel approach to discriminate machine printed and handwritten text from uniform occurrence of characters using rough set theory. The technique has been tested on local samples as well as on IAM dataset.

Banerjee et. al [8] proposed a text discrimination system for Bangla printed and handwritten text. Firstly, connected components has been identified and then features has been extracted from it. Finally, SVM has been used to classify the text into printed or handwritten. The authors reported overall accuracy of 96.49%.

Mozaffari and Bahar [9] presented an approach to discriminate handwritten from printed text for Farsi and Arabic script using three different features. SVM and KNN classifiers have been employed to achieve overall accuracy of 97.1%.

Zemouri and Chibani [10] proposed a method to discriminate printed and handwritten text from separate regions in document images. After pre-processing operations, structural & statistical features for each word has been generated using the Radon transform and SVM classifier has been used to classify the words. The system has been tested on IAM database and reported a recognition rate of 98%.

Silva et. al [11] presented a technique for discrimination of handwritten and printed text. Pre-processing operations has been applied to enhance the image and bounding box (BB) around each word has been identified. Eleven features has been used for each BB and classification has been carried out using data mining rules. The system reported overall 80% accuracy on two public databases.

Zheng et. al [12] reported a technique for discrimination of noisy handwritten and printed text. Fisher classifier has been used to identify printed and handwritten text and Markov Random Field (MRF) based post-processing technique has been used to further improve the results. Experiment results show the robustness of method on extremely noisy documents.

Kavallieratouet. al [13] proposed an approach to to separate printed and handwritten text for Latin script. During pre-processing, document image has been segmented into text-lines. Structural features has been extracted and text-lines has been classified using discriminant analysis. The authors reported 98.2% average accuracy with minimum training set.

Guo and Ma [14] presented a method for separation of handwritten and printed text for English texts and Latin script. The technique even works well for overlaid annotated text over printed text. The classification has been performed by hidden Markov models (HMM) with overall accuracy of 92.86%.
Pal and Chaudhary [15] proposed a method to discriminate handwritten and printed text using structural and statistical features of Bangla and Devanagari script text-lines. As the method is based on headline property of scripts, the technique can be extended to other headline based Indian scripts like Gurmukhi and Marathi also. The method attained an overall accuracy of 98.6%.
Table 1 gives comparative analysis of various methods and techniques discussed in this paper.

Table 1. Comparison of existing methods

AUTHOR	FEATURES/CLASSIFIERS	SCRIPT	ACCURACY
[1]	Structural and Statistical	English and Arabic	90%
[4]	PHOG	Arabic and Latin	98.3%
[8]	SVM	Bangla documents	96.49%
[9]	SVM and k-NN	Farsi and Arabic	97.1%
[10]	Radon Transform	IAM database	98%
[15]	Structural and Statistical	Bangla and Devanagari	98.6%

III. CONCLUSION

In this paper, various methods for automatic separation of handwritten and printed text has been discussed. Several methods have been proposed for various scripts including Indian. Most of the existing techniques work for document images having separate region of interest or full text-lines for handwritten and printed text. Few techniques are available for discrimination of mixed handwritten and printed text document images. Also, no work has been reported for isolating Gurmukhi script.

In future, the existing discrimination techniques will be modified or new methods may be proposed for Gurmukhi script mixed character document images.

IV. ACKNOWLEDGEMENTS

The first author of paper would like to thank all faculty members of Computer Engineering section of YCoE for their help and encouragement.

V. REFERENCES

[1] Saba T., Almazyad A.S., Rehman A. "Language Independent Rule Based Classification of Printed & Handwritten Text", International conference on evolving and adaptive intelligent system (EAIS), pp.1-4 December 1, 2015.

[2] Srivastava R.,Tewari R.K., Kant S., "Separation of Machine Printed and Handwritten Text for Hindi Documents" International research journal of engineering and technology(IRJET), Vol.2, Issue 2, pp.704-708, 2015.

[3] Jindal A., Amir M., "Automatic Classification of handwritten & printed text in ICR Boxes", International advance Computing Conference(IACC), IEEE, pp.1028-1032, 21 Feb., 2014.

[4] Saïdani A, and EchiA.K..Belaid A., "pyramid histogram to oriented gradient for machine-printed/handwritten and Arabic/latin words discrimination", 6th international conference of soft computing and pattern recognition, IEEE, pp.267-272, 11 Aug., 2014.

[5] Saïdani A, and EchiA.K..Belaid A. "Identification of Machine-printed and Handwritten Words in Arabic and Latin Scripts", 12th International conference on document analysis and recognition, IEEE, pp.798-802, 25 Aug., 2013.

[6] Zagoris K.et. al, "Handwritten and Machine printed text separation in document images using the Bag of Visiual Words Paradigm", International Conference on Frontiers in Handwriting Recognition, IEEE, pp.103-108, 18 Sep., 2012.

[7] Narayan S., Gowda S.D., "Discrimination of handwritten and machine Printed text in Scanned document images based on Rough Set Theory" World Congress on Information and Communication Technologies, IEEE, pp.590-594, 30 Oct., 2012.

[8] Banerjee P., Chaudhari B.B., "A System for Hand-Written and Machine-Printed Text Separation in Bangla Document Images" International Conference on Frontiers in Handwriting Recognition, IEEE, pp. 758-762, 18 Sep., 2012.

[9] Mozaffari S., Bahar P., "Farsi/Arabic handwritten from machine printed words discrimination", International Conference on Frontiers in Handwriting Recognition, IEEE, pp. 698-703, 18 Sept., 2012.

[10] Zemouri ET-T., Chibani Y., "Machine printed handwritten text discrimination using radon transform and SVM classifier", 11th International Conference on Intelligent Systems Design and Applications, IEEE, pp.1306-1310, 22 Nov., 2011.

[11] Lincoln Faria da Silva, Aura Conci and Angel Sanchez, "Automatic discrimination between printed and handwritten text in documents", Brazilian symposium on computer graphics and image processing, IEEE, pp. 261-267, 11 Oct., 2009.

[12] Yefeng Zheng, Huiping Li and David Doermann, "Machine printed text and handwriting identification in noisy document images ", IEEE transaction on pattern analysis and machine intelligence, Vol 26, No 3, pp. 337-353, 26 Mar., 2004.

[13] Kavallieratou E., Stamatates S., "Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics", 17th international conference on pattern recognition (ICPR), IEEE , Vol.1, pp.437-440, 23 Aug., 2004.

[14] JinhongK.Guo and Matthew Y. Ma, "Separating handwritten material from machine printed text using hidden Markov Models", 6th international conference on analysis and recognition (ICAR), pp. 439-443,2001.

[15] Pal U., Chaudhuri B.B., "Machine-printed and hand-written text lines identification", Pattern recognition letter Vol.22, Issue 3, Elesevier, pp. 431-441, 31 Mar., 2001.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

CITE AN ARTICLE

Kaur, M., & Singh, B. (2017). CLASSIFICATION OF PRINTED AND HANDWRITTEN TEXT: A REVIEW. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 6*(8), 403-407. Retrieved August 25, 2017.