International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-3, February 2013

# A Survey on Techniques of Separation of Machine Printed Text and Handwritten Text

**Ranjeet Srivastva, Aditya Raj, Tushar Patnaik, Bhupendra Kumar**

***Abstract*—In many documents such as admission form, bank cheques, memorandums, letters and application forms machine printed and handwritten characters are mixed. Since the algorithms for recognition of machine-printed texts and handwritten texts are different, it is necessary to distinguish between these two types of texts before giving it to respective OCR systems to process it. This separation will definitely increase the performance and overall system quality. The paper discusses some observations about characteristics of these two types of texts and various techniques of separation of machine printed and handwritten text into three categories (Structural and statistical features, Gradient features and Geometric features) based on feature extraction method.***

***Keywords*-- Feature Extraction, Handwritten Text, Machine Printed Text, OCR.***

## I. INTRODUCTION

The presence of mixed type texts in a document image is an important obstacle towards the automation of the optical character recognition procedure. Machine printed character recognition and handwritten character recognition techniques are quite different in every aspect like preprocessing, character segmentation, feature extraction etc. Hence it is better to separate these two types of texts before feeding them to respective OCR system. The classification of machine printed and handwritten text is typically performed at the block or zone level, text-line level, word level or character level. Until recently most of research work has been done on word level and some work on line level, because at word level it is possible to analyze more complex pages which contain both type of words even within the same line and a single word is typically uniform with respect to writing style. Line level segmentation of a document image is quite simple in comparison of word level segmentation but it is less accurate on separation of the mixed type of texts.

Separation of machine printed and handwritten text is a challenging task and it becomes even more tough in the documents in which printed script is cursive in nature or the handwritten texts are written very close to the printed texts or handwritten texts are overlapping with the printed texts. Another major hindrance in separation is presence of noise in the image and skewness in the image. Detection and removal of noise is a difficult task due to the irregular variations in its size, shape and nature. Various algorithms have been used for noise removal and skew angle correction in the preprocessing step. Lincoln Faria da Silva, Aura Conci, Angel Sanchez [1] uses 3*3 median filters for noise removal. E. Kavallieratou, and S. Stamatatos [2] have performed skew angle correction by employing horizontal histogram and Winger-ville distribution. The paper is strutured as follows- section-II describes observations about the characteristics of printed and handwritten text, various techniques of separation of text are discussed in section-III and section-IV compares various approaches. Section-V contains the conclusion.

**Manuscript published on 28 February 2013.**
***** Correspondence Author (s)
**Ranjeet Srivastva**, M.Tech(IT), Department of IT, C-DAC, Noida, India.
**Aditya Raj,** M.Tech(CSE), Department of CSE, C-DAC, Noida, India.
**Tushar Patnaik**, Sr. Lecturer/ Sr. Project Engineer, Department of CSE, C-DAC, Noida, India.
**Bhupendra Kumar**, Sr. Technical Officer, Department of CSE, C-DAC, Noida, India.



## II. OBSERVATIONS OF CHARACTERISTICS OF MACHINE PRINTED & HANDWRITTEN TEXT

Some features of machine printed and handwritten characters which can be used for separation are given as follows-

(1) Machine printed characters are written straight whereas handwritten characters may or may not written straight. A large proportion of printed text be linear and aligned properly either horizontally or vertically while edges in the handwritten characters may not be linear [10].

(2) Machine printed characters have proper spacing and are less likely to overlap, whereas handwritten characters may have overlapping and touching character which results in major challenge to preprocessing and segmentation step[4].

(3) Since the machine printed characters are written in proper alignment they have larger regularities in projection profile but the handwritten annotations do not have regularities in projection profile because of varying style of different author and the environment [12].

(4) A machine printed text line has relatively stable height compared to handwritten text line, and the mean and variance of width of each character is consistent [2].

(5) Horizontal run and gradients are uniform in machine printed text. If the text is repeated then it have stroke in same direction in all occurrences [11].

## III. TECHNIQUES FOR SEPARATION

The techniques for separation of machine printed and handwritten text can be categorized based on (i) Structural and Statistical features (ii) Gradient features (iii) Geometric features. All the techniques have mainly three steps i.e. image preprocessing, feature extraction and classification. This paper discusses various techniques based on the above features and within these three steps. The steps involved in separation are shown in "Fig.1" and categorization of techniques of separation is shown in "Fig.2".

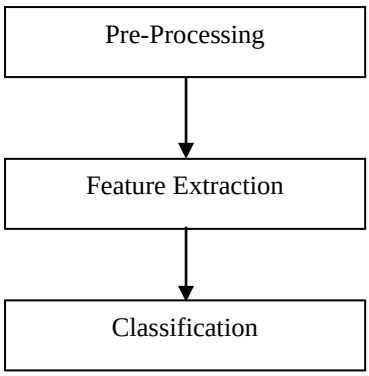


Fig. 1 Steps of Separation

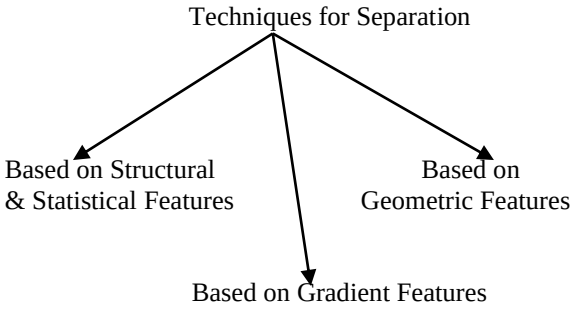


Fig. 2 Techniques for Separation

***A. Techniques based on Structural and Statistical Features:*** K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V [3] have proposed an approach of separation of Roman alphabet. In preprocessing they have segmented the image at character level and extract the bounding box and contour of the characters and used two features: straightness of lines (horizontal & vertical) in the character and symmetry related to different points. Most of the roman alphabets have straight line. Symmetry relative to center of gravity is another feature for the characters which do not have straight line. They have used Hough Transform method to search for the straight line in a character. In the last a feed forward neural network is used for classification.

E. Kavallieratou, and S. Stamatatos [2] used a trainable approach to localize text areas and split them into text lines. In the preprocessing step two stage (first on page level and second on text area level) skew angle corrections is performed using horizontal histogram and Winger ville distribution, then document is segmented into smaller areas called first order connected components(CC) using constrained run length algorithm(CRLA) also known as "smearing". In order to minimize the number of first order CCs they have eliminated the CCs based on following criteria:

(1) If the area of their corresponding bounding boxes (BB) < 100 pixels. Those CCs are assumed to be noise.

(2) If the aspect ratio i.e. width and height of corresponding BBs < 1/20. E.g. Vertical line

(3) If the aspect ratio> 20. E.g. Horizontal line

After the preprocessing, following structural features are used:

(1) The height of printed character is more or less stable within a text-line whereas it may not be stable in case of handwritten text-line.

(2) A Hindi text-line is partitioned into three zones, first is upper zone or ascender which is portion above the sirorekha or headline, second is middle zone or main body which contain basic characters below headline and the third is lower zone or descender where some lower modifier resides. The ratio of ascender's height to main body height and ratio of descender's body to main body height would be stable in printed text and variable in handwriting.

U.Pal and B.B Chaudhuri [4] present a scheme of separation in both Bangla and Devnagri script. They have used histogram based thresholding approach to convert gray scale image to binary image, detect the individual column using run length smoothing approach after that mode of the text is detected i.e. whether the text is in portrait or in landscape mode based on the criteria that white space between characters is always much smaller than white space between lines, and in the final step of preprocessing the text lines are segmented using horizontal projection profile. Three levels of features are used in feature extraction:-

(1) Machine printed Hindi text lines generate long horizontal run as all the characters are attached to a headline, handwritten text lines may or may not generate long horizontal run i.e. less than threshold.

(2) If we delete the headline region in Bangla and Devnagri script then in the case of machine printed words all the characters will be isolated but this may not be true for handwritten words because of irregular alignment of characters in the word.

(3) Distribution of character's lowermost point is regular in machine printed text and irregular in handwritten text. These three features are given to the nodes of a tree classifier for classification of printed and handwritten text.

*B. Techniques based on Gradient Features:* S.Imade, S.Tatsuta and T.Wada [5] present a technique for segmentation and classification of mixed text/image document using neural network. They have separated a document image into printed and handwritten kanji & kana characters, photographs and painted image regions. The input image is converted into monochromatic binary image by an appropriate thresholding method which is transformed into 8*8 image blocks, spatial filtering is used for noise removal and finally the regions are segmented based on the criteria that if 16 white elements continue in vertical or horizontal direction the elements are converted to white line segment any other element than the white line segments is converted to black elements. Each region connected with black elements is enclosed by a rectangle. The proposed approach uses two features that are

(1) Histogram of gradient vector directions which are computed for all pixels in a block image & the local sum for the histogram is computed at every quantized direction of 15 degree. Printed Characters have significant peaks in both vertical and horizontal direction on the histogram of gradient vector directions, handwritten characters have same peaks but smaller than that of printed characters, painted images do not have significant peak in both directions and the gradient vector direction of photographs cannot be characterized by any direction.

(2) The histogram of the luminance levels of printed & handwritten characters have two peaks because of their binary gradation.


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

These two histograms are input to layered neural networks (NN) which consist of three layers: an input layer, a hidden layer and an output layer. This NN has already been trained by sample images.

*C. Techniques based on Geometric Features:* Seung Ick Jang, Seon Hwa Jeong and Yun-Seok Nam [6] proposed a method for classifying machine printed and handwritten addresses on Korean mail piece images. They have divided their approach in three blocks: (1) Valid Connected Component (CC) grouping which is consists of five steps, first the CCs are generated from input image then small CCs are removed considering them as noise, then extract only those CCs that forming skyline which detect CCs in the upper most text lines, again they eliminate the CCs not belonging to upper most text line and finally grouped the small CCs. (2) Feature Extraction where they have used the feature that machine printed characters have similar width and are arranged regularly but it is not guaranteed in handwritten characters. Considering this property they have used three geometric features (i) Width magnitude histogram (ii) Width variance histogram (iii) Position histogram. (3) Classification where a multi layer perceptron network which has 3 layers: an input layer with the same number of nodes as the dimension of the feature set, a single hidden layer with 20 nodes and an output layer with 2 nodes is used for classification.

J. Franke and M. Oberlander [7] proposed an approach for form reading applications. They have used different statistical classifiers whose output was combined and the combination was carried out by another statistical classifier. Features used by them are: (1) Machine fonts with constant spacing have more or less a constant center distance between centers of two successive rectangles (BBs). (2) Machine fonts with proportional spacing have a constant gap between two successive rectangles. Feature vector was divided into four parts and a linear classifier was adopted for each of them. These are (a) Width Histogram Classifier (dW1, dW2), (b) Height Histogram Classifier (dH1, dH2), (c) Gap Histogram Classifier (dG1, dG2) and (d) Center Distance Histogram Classifier (dC1, dC2) where components of each classifier corresponds to decision vector of each data field i.e. Machine printed or handwritten. A new feature vector (dW1, dH1, dG1, dC1) is calculated from the first component of these classifiers because second component is linearly dependent on first component. With this new feature vector a combination classifier D was adopted on the learning set.

## IV. COMPARISON OF VARIOUS TECHNIQUES

**C**omparisons of various techniques for separation of machine printed text and handwritten text is shown in Table I. The comparison is done by considering the following parameters: features used size of data set, classifier used and accuracy of the system.

Table I
*Comparisons of various methods of Separation*

| *Method Proposed by* | *Feature* | *Classifier* | *Data Set* | *Accuracy %* |
|---|---|---|---|---|
| U. Pal, and B. B. Chaudhuri [4] | Statistical | Tree Classifier | 100 | 98.3 |
| K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V [3] | Statistical | Feed Forward Neural Network | 1068 | 78.5 |
| E. Kavallieratou, and S. Stamatatos [2] | Structural | Statistical Classifier | 50 | 98.2 |
| S.Imade, S.Tatsuta and T.Wada [5] | Gradient | Feed Forward Neural Network | 1000 | 56 |
| Seung Ick Jang, Seon Hwa Jeong and Yun-Seok Nam [6] | Geometric | Multi Layer Perceptron | 3147 | 98.9 |
| J. Franke and M. Oberlander [7] | Geometric | Statistical Classifier | 4659 | 98.27 |
| Yefeng Zheng, Huiping Li, David Doermann [8] | Geometric | Fisher Classifier | 94 | 78 |
| Ranju Mandal, Partha Pratim Roy, Umapada Pal [9] | Gradient | Support Vector Machine | 16743 | 98.56 |

## V. CONCLUSION

Methods for treating the problem of separation of machine printed and handwritten text have developed remarkably in the last two decades. In this paper we have discussed an organization of these methods under three basic categories. It is hoped that this comprehensive discussion will provide insight into the concept involved. We have noticed that most important problem of separation methods is segmentation because of cursive text lines, overlapping text lines, closeness of handwritten annotations with printed text, noisy document or skewed document image. We have included a list of references sufficient to provide more detailed understanding of the approaches described. We apologize to researchers whose important contributions may have been overlooked.

## ACKNOWLEDGMENT

The authors are thankful to the referees for their critical comments. We are also thankful to Mr.Alok Kumar (M.Tech-CSE) and Ms.Madhuri Yadav(M.Tech-IT) at C -DAC Noida, for their helpful suggestions.

## REFERENCES

[1] Lincoln Faria da Silva, Aura Conci, Angel Sanchez, *Automatic discrimination between printed and handwritten text in documents*, IEEE, 1530-1834/2009.
[2] E. Kavallieratou, and S. Stamatatos, *Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics*, Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, v. 1, 23 - 26 Aug., pp.437 - 440, 2004.
[3] K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V, *A System for Machine-Written and Hand-Written Character Distinction*, Proceedings of the Third International Conference on Document Analysis and Recognition,v.2,14 - 16 Aug.,pp 811 - 814, 1995.

[4] U. Pal, and B. B. Chaudhuri, *Automatic separation of machine-printed and hand written text lines*, ICDAR '99. Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 645-648, 1999.
[5] S.Imade, S.Tatsuta and T.Wada, *Segmentation and Classification for Mixed Text/Image Document Using Neural Network, In Proc. 2nd ICDAR*, pp. 930-934, 1993.
[6] Seung Ick Jang, Seon Hwa Jeong and Yun-Seok Nam, *Classification of Machine Printed and Handwritten Addresses on Korean Mail Piece Images Using Geometric Features*, Proc. 17th International Conference on Pattern Recognition (ICPR'04) 1051-4651/04.
[7] J. Franke and M. Oberlander,*Writing Style Detection by Statistical Combination of Classifiers in Form Reader Applications, Proceedings of 2nd ICDAR*, 1993, pp. 581-584.
[8] Yefeng Zheng, Huiping Li, David Doermann, *Machine Printed Text and Handwriting Identification in Noisy Document Images*, IEEE Transactions on Pattern Analysis & machine intelligence, Vol. 26, No. 3, 0162-8828 march 2004.
[9] Ranju Mandal, Partha Pratim Roy, Umapada Pal, *Signature Segmentation from Machine Printed Documents using Conditional Random Field*, ICDAR.2011.236, IEEE, 1520-5363.
[10] Sean Violante, Robert Smith and Mike Reiss, A computationally efficient technique for discriminating between handwritten and printed text, IEE, 1995.
[11] Faisal Farooq, Karthik Sridharan, Venu Govindraju, Identifying handwritten text in mixed document, IEEE, 0-7695-2521-0/2006.
[12] Jinhong K.Guo,Matthew y.ma, Separating handwritten material from machine printed text using hidden markov models, IEEE, 0-7695-1263-1/2001.

**Mr. Ranjeet Srivastva** received his B.Tech in IT from U.P.T.U Lucknow, Uttar Pradesh, India in 2008.Currently, he is doing M.Tech in IT from C-DAC Noida(Affiliated to G.G.S.I.P.U New Delhi), India. He is working on the project "Separation of Machine Printed and handwritten text in Hindi". His interest areas are Digital Image Processing, Theory of Computation, Data Structure, Computer Networks, Network Security and Operating Systems.

**Mr. Aditya Raj** received his B.Tech in IT from U.P.T.U Lucknow, Uttar Pradesh, India in 2010.Currently, he is doing M.Tech in CSE from C-DAC Noida(Affiliated to G.G.S.I.P.U New Delhi), India. He is working on the project "OCR of machine printed Oriya script". His interest areas are Digital Image Processing, Theory of Computation, Data Structure, Operating Systems.

**Mr. Tushar Patnaik** (Sr. Lecturer/Sr. Project Engineer) joined C-DAC in 1998. He has thirteen years of teaching experience. His interest areas are Computer Graphics, Multimedia and Database Management System and Pattern Recognition. At present he is leading the consortium based project "Development of Robust Document Image Understanding System for Documents in Indian Scripts".

**Mr. Bhupendra Kumar** (Senior Technical Officer) joined C-DAC in 2005, he received his M.Tech.degree from IIIT Allahabad with the specialization in wireless communication and computing. His interest areas are Advanced Image processing, pattern recognition, computer network, wireless network, MANETs. Currently he is involved in project "Development of Robust Document Image Understanding System for Documents in Indian Scripts".