A System for Hand-Written and Machine-Printed Text Separation in Bangla Document Images

Purnendu Banerjee
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
203, B. T. Road, Kolkata-700 0108, India
e-mail: *purnendubannerjee@yahoo.com*

B. B. Chaudhuri
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute
203, B. T. Road, Kolkata-700 0108, India
e-mail: *bbc@isical.ac.in*

***Abstract*—In this paper, we describe an approach to distinguish between hand-written text and machine-printed text from annotated machine-printed Bangla Documents images. In applications involving OCR, distinction of machine-printed and hand-written characters is important, so that they can be sent to separate recognition engines. Identification of hand-written parts is useful in deleting those parts and cleaning the document image as well. In this paper a classification system is presented which takes a connected component in the document image and assigns them to two classes namely "machine-printed" and for "hand-written" classes, respectively. The proposed system contains a preprocessing step, which smoothes the object border and finds the Connected Component. Bangla script specific features are extracted from that Connected Component image, and a standard classifier based on SVM generates the final response. Experimental results on a data set show that the proposed approach achieves an overall accuracy of 96.49%.**

Keywords-Bangla Script Recognition; Printed and Handwritten Text Separation; SVM Classifier

I. INTRODUCTION

Automatic recognition of document text has been an active area of research for many years. Its main area of application is data processing oriented to the business world. The principal motivation for the development of these systems is the need to cope with the enormous flood of papers such as office documents, commercial forms, government records, postal mails, tabular forms, bank cheques etc. From the generic point of view, documents can be of three types: (a) Printed, (b) Handwritten and (c) Mixed (printed and handwritten) documents. Printed documents can be generated by printing technology such as laser, inject, offset, intaglio and screen printing etc. Handwritten documents are made by manual writing on paper. A Mixed document contains both printed and handwritten texts. A software, called Optical Character Recognition (OCR), is used to recognize the documents. The recognition of mixed documents is still a challenge because printed and handwritten scripts OCRs employ distinctly different algorithms. So, the first step in this case is to isolate handwritten parts from the printed parts and then send them to the respective OCR engine. This paper is an attempt to do the task for Bangla script documents.

The proposed approach is based on developing a two-class classifier (machine-printed versus handwritten) on individual connected components of the document image. Here we are concerned with Bangla, which is one of the most popular scripts in South Asia. It is used to write Bangla, Manipuri and Assamese language texts. About 300 million people of Eastern India and Bangladesh use these languages and script for their daily need.

In the proposed approach, features specific to Bangla printed script that can discriminate the handwritten text are chosen for the classifier. Three different classifiers were tested and the Support Vector Machine (SVM) classifier with a specific Kernel was found to be the best in performance.

This paper is organized as follows: A review of machine printed / handwritten text separation is presented in Section 2. In Section 3, the preprocessing and feature selection method are presented. The experimental results including the classifier selection are described in Section 4. Concluding remarks are provided in Section 5.

II. REVIEW ON MACHINE-PRINTED / HANDWRITTEN TEXT SEPARATION

Among earliest studies on the problem, Umeda and Kasuya [2] patented an approach where discrimination is performed by calculating the ratio between the number of slanted strokes and the sum of horizontal, vertical and slanted ones, where a predetermined static threshold is imposed. An overall recognition rate of 95% is reported under these conditions.

Later Kunuke et al. [3] proposed a method based on the extraction of scale and rotation invariant features: the straightness of vertical and horizontal lines and the symmetry relative to the centre of gravity of the character. Their results showed a recognition rate of 96.8% on a training set of 3632 and 78.5% on a test set of 1068 images. On the other hand, Fan et al. [4] used the character block layout variance and reported a 85% accuracy on English and Japanese textual images. In 2001 Pal and Chaudhuri [5]

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.171

presented a method for Bangla and Devnagari script based on structural regularities of the alphabet. It uses a hierarchy of three different features to perform the discrimination. Guo et al. [6] suggested a method with HMM classifier to separate type-written and handwritten words based on vertical projection profiles of the word. On a test-set of 187 words, they obtained a precision rate of 92.86% for the typewritten and 72.19% for the handwritten words. In [1], Jose et al. suggested content related as well as shape related features to characterize handwritten text on bank check images.

More recently Zheng et al. [7, 11] reported a robust printed and handwritten text segmentation approach from extremely noisy images. They used classifiers like k-nearest neighbors, support vector machine (SVM) and Fischer discriminator with features like pixel density, aspect ratio and Gabor filter output and achieved a segmentation accuracy of 78%. Meanwhile Jang et al. [8] described an approach, specific for Korean text, based on the extraction of some geometric features. They employed a multilayer perceptron classifier, achieving an accuracy rate of 98.9% on a test-set of 3,147 images. Kavallieratou [9] showed that a simple discriminant analysis on the vertical projection profiles performs as efficiently as many robust approaches. Farooq et al. [10] use an EM (Expectation Maximization) based probabilistic NN (Neural Network) model to identify Arabic handwritten text in mixed documents.

One interesting application of such classification is the detection and matching of signatures proposed by Zhu et al. [12]. A similar approach but using Conditional Random Field is described by Shetty et al. [13]. Peng et al. [14] suggested a novel approach based on three categories of word level feature and a k-means classifier associated with a relabeling post-procedure using Markov random field models; they achieved an overall recall of 96.33%. In a more general scenario of sparse data and arbitrary rotation Chanda et al. [15] recently described their approach based on the SVM classifier and obtained an accuracy of 96.9% on a set of 3958 images.

III. PROCESSING AND FEATURE SELECTION

Our system for machine-printed versus handwritten text recognition has three subsystems namely preprocessing, feature extraction and classification. The image preprocessing is used to smooth the object's border and to reduce the computational time of the subsequent components. The preprocessing steps are as follows:

- *Binarization*: A gray scale image of an input document is binarized by a thresholding operation. The threshold is determined by Otsu [16] or Sauvola algorithm [17].
- *Smoothing of object border*: A simple morphological processing is used to smooth the border. A black pixel with five to seven consecutive white border pixels is made white, or vice versa, provided the conversion does not disturb the topology of the component.
- *Bounding box extraction*: A component labelling algorithm is used to detect connected components. Small components containing five pixels or less are removed as noise. Bounding boxes are generated for remaining components.

Next, we consider feature selection both for training and classification task. Since most characters of Bangla alphabet contain straight lines (conversely, straight lines are rare in handwritten text) line straightness is an important feature to detect the printed part. We are concerned about the long run length black pixels which are greater than Stroke Thickness (ST). To calculate the stoke thickness, we first find the mode values namely $Mode(l_h)$ and $Mode(l_v)$ of horizontal black pixel run length frequency distribution and vertical black pixel run length frequency distribution, respectively. Then the stroke thickness can be represented as

$$ST = Min(Mode(l_h), Mode(l_v)) \quad (1)$$

Next, we calculate the mean value on the horizontal, vertical and diagonal black pixel run length frequency distribution for runs greater than ST, and used it for feature calculation. This is so because we are interested in run length of black pixels which are greater than the mean of the above frequency distribution.

The printed Bangle characters have a horizontal line called headline at the upper part of most characters which join to make longer line in a word. When hand-written, the writers normally do not draw this head line. If occasionally drawn, it is not as horizontal as in case of printed text.

So, a measure based on horizontal line at upper part of a connected component has been used by us. Since longer straight line provides more confidence to print script, the length is weighted by a factor proportional to its length. More specifically, if the width of the bounding box of the connected component is W and if l_{hi} is the length of the ith horizontal line, then we propose two weighted horizontal length features as

$$F_{H1} = \frac{\sum l_{hi} * f_{hi} * w_{1hi}}{\sum f_{hi}} \quad \text{Where } w_{1hi} = \left(\frac{l_{hi}}{W}\right)^{2} \quad (2)$$

$$F_{H2} = \frac{\sum l_{hi} * f_{hi} * w_{2hi}}{\sum f_{hi}} \quad \text{Where } w_{2hi} = e^{\frac{ST}{1+(W-l_{hi})}} \quad (3)$$

The horizontal feature value is related to the size of the connected component and the number of pixels belonging to that line. The first weighted function is maximum when the argument is equal to the length of bounding box ie, $l_{hi} = W$, making the maximum weight equal to 1. A shorter l_{hi} gets less weight which rapidly decreases to zero, since hand-written component has shorter horizontal line. In the second case, the exp function is maximum when the argument $(W - l_{hi})$ is equal to zero ie, $l_{hi} = W$, making the weight maximum, equal to e^{ST}, where the value of stroke thickness of machine printed documents is greater than that of

handwritten documents. In this feature we put some importance on the stroke thickness (ST).

Our third and fourth features are based on the vertical straight lines. Many printed Bangla characters have a vertical line stroke. Hand-written text also contain vertical strokes but in less number and they are not as straight as in case of printing. So, we can make a weighted vertical run length as our second feature. If the vertical run length and the height of the connected component are represented as l_v and H respectively, then the weighted vertical length features are proposed as

$$F_{V1} = \frac{\sum l_{vi} * f_{vi} * w_{1vi}}{\sum f_{vi}} \quad \text{Where } w_{1vi} = \left(\frac{l_{vi}}{H}\right)^2 \tag{4}$$

$$F_{V2} = \frac{\sum l_{vi} * f_{vi} * w_{2vi}}{\sum f_{vi}} \quad \text{Where } w_{2vi} = e^{\frac{ST}{1+(H-l_{vi})}} \tag{5}$$

The vertical feature values are related to the size of the connected component and the number of pixels belonging to that line. In F_{V1} the weighted function is maximum when the argument is equal to the height of bounding box ie, $l_{vi} = H$, making the weight equal to 1. A shorter l_{vi} gets less weight, which happens for hand-written component having shorter vertical line. In the second case, the exp function is maximum when the argument $(H - l_{vi})$ is equal to zero ie, $l_{vi} = H$, making the weight maximum. Here, the stroke thickness is the same as in equation (1).

The fifth and sixth features are based on the run length in the two diagonal directions. A large number of printed Bangla characters have delta shape and hence diagonal line stroke in 45^0 or 135^0 (↗ & ↖). In case of handwriting, the writers normally make it more roundish, not so straight. If occasionally drawn, they have less number of pixels compared to printed characters. So, the occurrence of diagonal line of a connected component is one feature used by us. If the height and width of the bounding box of the connected component is H and W, respectively and if l_d is the diagonal run length, then the weighted diagonal length feature may be

$$F_{D1} = \frac{\sum l_{d1i} * f_{d1i} * w_{d1i}}{\sum f_{d1i}} \quad \text{Where } w_{d1i} = \left(\frac{l_{d1i}}{H}\right) \tag{6}$$

$$F_{D2} = \frac{\sum l_{d2i} * f_{d2i} * w_{d2i}}{\sum f_{d2i}} \quad \text{Where } w_{d2i} = \left(\frac{l_{d2i}}{H}\right) \tag{7}$$

As the last feature we consider the foreground density (F_{FD}) of every component which may be expressed by equation (8). Usually the foreground density of printed text differs from that of the hand-written components.

$$F_{FD} = \frac{Total\ No.\ of\ Black\ Pixels\ in\ the\ component}{Total\ No.\ of\ Pixels\ Present\ in\ the\ Bounding\ Box} \tag{8}$$

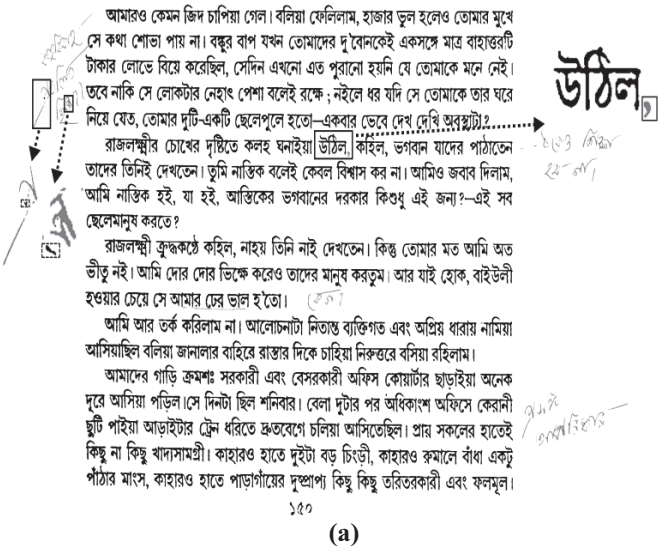
IV. Data Collection, Classification and Discussion

The dataset used for our experiment consists of 13830 components, of which 7258 are handwritten and 6572 are printed ones. We have collected these data from culturally heterogeneous group of Masters and PhD students. Two types of documents were considered. One type is the page of a book, on which some annotation has been made by hand (see Fig 1(a)). The other type is an application form which was filled by handwriting (see Fig 1(b)). These students produced the annotation and application form fill-up in their handwriting. In order to evaluate the effectiveness of the classification methodology, the training and testing database were created with machine-printed and hand-written components. The training set included 6937 components in which 4182 components are hand-written and 2755 components are machine-printed ones.

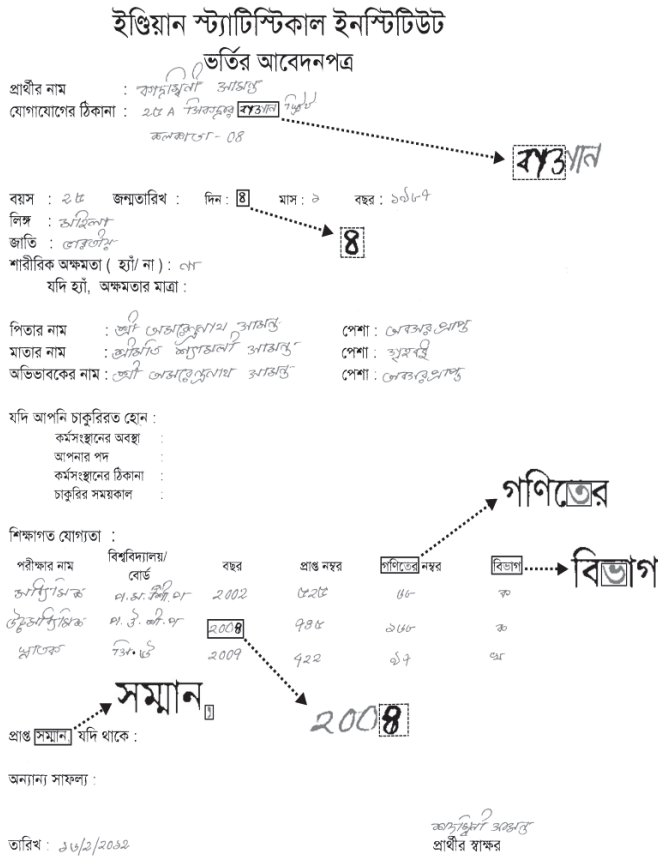
Three types of classifier have been implemented to test the proposed method of separating handwritten and machine printed text. They are (a) minimum distance classifier, (b) K-Nearest Neighbor (K-NN) classifier and (c) Support Vector Machine (SVM) classifier. To evaluate performance of the proposed method, we considered standard definition of recall (R), precision (P) and f-measures (F). Experiments in term of recall and precision for the pixels of handwritten component recognition results are presented in Table 1.

TABLE 1: Handwritten Region Recognition Performance

Classifier	Kernel Type	R	P	F
Minimum distance	-	0.85	0.89	0.87
K-NN	-	0.94	0.90	0.92
SVM Type nu-SRV	Linear	0.81	0.82	0.81
	Polynomial	0.83	0.88	0.85
	RGB	0.93	0.92	0.92
	Sigmoid	0.81	0.83	0.82
SVM Type epsilon-SRV	Linear	0.81	0.84	0.82
	Polynomial	0.85	0.87	0.86
	RGB	0.95	0.94	0.94
	Sigmoid	0.83	0.86	0.84
SVM Type C-SVC	Linear	0.83	0.85	0.84
	Polynomial (c=10, g=0.1)	0.88	0.87	0.87
	RGB (c=28, g=0.17)	0.96	0.97	**0.96**
	Sigmoid (c=9, g=0.14)	0.86	0.87	0.86



(a)



(b)

Figure 1. Example of typical recognition of text types.

For minimum distance classifier, the feature vectors are computed for all training samples of each class and the mean vector is used as the representative point for the class. For testing, the feature vector of the sample component is computed and its distance from the class representation point is found. The sample is assigned to the class for which the distance is minimum.

In case of K-NN, K nearest neighbors from the training pattern vectors are computed for any test sample component. The sample is assigned to the class from which majority of their neighbors come. For our experiment, K=11 resulted in the best classification.

SVM is a powerful classifier having several versions, each using one of several types of kernel. We have tested with three versions, each with four types of kernel. The results are shown in Table-1. It is noted that C-SVM with RGB kernel yields best f-measure of 0.96. Individually speaking, 0.97 precision is obtained for identifying the retrieved handwritten test part and 0.96 recall is obtained for identifying the total relevant handwritten text part. Since this is a two-class problem, the results of Table 1 can be converted into recognition accuracy for printed text identification as well. For example, in case of best result situation of Table 1 the printed text P, R and F values are 0.98, 0.97 and 0.97, respectively. Higher accuracy for printed text can be attributed to our choice of features. Note that, the chosen features emphasize on positive identification of printed parts, while no feature stress on positive identification of handwritten parts.

Two examples of typical recognition output are shown in Fig 1. Here correctly recognized printed and handwritten text parts are shown in black and intermediate gray, respectively. To show the error clearly, the region is magnified and the misrecognition is enclosed in box. Handwritten part misrecognized as printed part is surrounded by dashed line box while printed part misrecognized as handwritten part is surrounded by continuous line box.

V. CONCLUSION

A text identification system has been presented, which is able to discriminate between machine-printed and handwritten Bangla text. The proposed approach can handle document pages where printed and handwritten components are distinct components. A set of simple and easy-to-compute features have been used for the recognition engine. Though developed for Bangla script, the approach can be used for Devanagari, (the most popular script in India) since headline and vertical lines are dominant shapes in Devanagari script as well.

More difficult situation may arise when a handwritten text component touches the printed text at several places. Even more complex is when handwriting is overlaid on several text lines. This is a case of document inpainting [18, 19] where the task is to separate/ delete the handwritten text leaving all information of printed text intact. We plan to

work on such problem in future and report the useful results in another correspondence.

REFERENCES

[1] J. B. D. S. Eduardo, B. Dubuisson, and F. Bortolozzi. "Characterizing and distinguishing text in bank cheque images". Proc. *XV Brazilian Symposium on Computer Graphics and Image Processing*, pp. 203–209, 2002.

[2] T. Umeda and S. Kasuya. "Discriminator between handwritten and machine-printed characters", 1990. [Online Available: http://www.freepatentsonline.com/4910787.pdf]

[3] K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V. "A system for machine-written and hand-written character distinction". Proc. *ICDAR,* Vol 2, pp. 811-814, 1995.

[4] K.C. Fan, L.S. Wang, and Y.T. Tu. "Classification of machine-printed and handwritten texts using character block layout variance". *Pattern Recognition*, Vol. 31, No 9 pp. 1275– 1284, 1998.

[5] Pal. U. and Chaudhuri. B. B. "Machine-Printed and Hand-Written Text Lines Identification". *Pattern Recognition Letters,*Vol. 22, No. 3/4, pp. 431-441, 2001.

[6] J. K. Guo and M. Y. Ma. "Separating handwritten material from machine printed text using Hidden Markov Models". Proc. *ICDAR*, pp. 439-443, 2001.

[7] Y. Zheng, S. Member, H. Li, and D. Doermann. "Machine printed text and handwriting identification in noisy document images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 3,pp. 337-353, 2004.

[8] S. I. Jang, S. H. Jeong, and Y.-S. Nam. "Classification of machine-printed and handwritten addresses on Korean mail piece images using geometric features". Proc. *ICPR,* Vol 2, pp. 383–386, 2004.

[9] E. Kavallieratou, S. Stamatatos, and H. Antonopoulou. "Machine-printed from handwritten text discrimination". Proc. *IWFHR*, pp. 312–316, 2004.

[10] F. Farooq, K. Sridharan, and V. Govindaraju. "Identifying handwritten text in mixed documents". Proc. *ICPR*, pp. 1142–1145, 2006.

[11] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. "Multi-scale structural saliency for signature detection". Proc. *CVPR,* pp. 1–8, 2007.

[12] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. "Signature detection and matching for document image retrieval". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 2015–2031, 2009.

[13] S. Shetty, H. Srinivasan, M. Beal, and S. Srihari. "Segmentation and labeling of documents using conditional random fields". Proc. *Document Recognition and Retrieval IV, Proceedings of SPIE*, pp. 6500U–1–11, 2007.

[14] X. Peng, S. Setlur, V. Govindaraju, R. Sitaram, and K. Bhuvanagiri. "Markov random field based text identification from annotated machine printed documents". Proc. *ICDAR*, pp. 431–435, 2009.

[15] S. Chanda, K. Franke, and U. Pal. "Structural handwritten and machine print classification for sparse content and arbitrary oriented document fragments". Proc. *ACM Symposium on Applied Computing*, pp. 18–22. 2010.

[16] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 9, No. 1, pp. 62-66, 1979.

[17] J. Sauvola and M. Pietikainen, "Adaptive document image binarization", *Pattern Recognition,* Vol. 33, pp. 225-236, 2000.

[18] M. Bertalmio, G. Sapio, V. Caselles and C. Ballester, "Image Inpainting", Proc*. of ACM SIGGRAPH*, pp. 417-424, 2000.

[19] T. F. Chan and J. Shen, "Nontexture Inpainting by Curvature-Driven Diffusion", *Journal of Visual Communication and Image Representation*, Vol. 12, No. 4, pp. 436-449, 2001.