# Identification of Handwritten Text in Machine Printed Document Images

**Sandipan Banerjee**,

*Dept. of Computer Science and Engineering,*

*National Institute of Technology, Durgapur, India*

sandipan9008@gmail.com

***Abstract* — In our daily lives we come across many documents where both printed and handwritten text co-exist and sometimes intermingle. As the OCR techniques for processing the two are quite different it is necessary to classify and distinguish them first. In this paper, a scheme has been proposed by which handwritten, printed and "mixed" text regions in the same document image can be identified and demarcated from each other for Bangla, the second most popular Indian script. The proposed scheme has been established on the basis of the structural and statistical idiosyncrasies of printed and handwritten Bangla text.**

***Keywords- Document Processing; Optical Character Recognition ; Bangla Script; Machine-printed and Handwritten Text; Indian Language***

## I. Introduction

Document images are processed and analyzed successfully using the Optical Character Recognition (OCR) techniques for applications like natural language processing, text mining, human aid etc. Most of the OCR systems present in the market however are dedicated for machine printed texts only and some of the ones that do exist for handwritten ones are mainly for English and other Latin based scripts. For Indian languages however and especially Bangla, some work has been done in this field and they are mostly concerned with segmenting, extracting and recognizing individual characters from the given text.

But the whole equation changes when dealing with handwritten text as the recognition schemes for it are totally different from that of printed ones. In terms of pre-processing, segmentation, noise removal, feature selection and classification etc. handwritten text recognition is a different ball game totally. So, if a document page consists of both handwritten and printed text together in it, then it becomes very difficult to process it using OCR techniques and therefore the two types of text have to be separated and distinguished first before being fed to the OCR system.

The separated printed and handwritten text can then be recognized using the standard OCR systems available today.

Documents with printed and handwritten text together are found quite frequently in our daily lives. Some of the common ones are question papers or feedback forms where the printed questions or queries are put inside a table and the answers are to be provided by hand. Also, such documents can be found in printed text books where several lines are underlined, highlighted or annotated by hand for taking notes. Not only for proper processing but many historic documents, which are the last remaining prints of a text, if tainted with handwritings, can be cleaned up and the original text can be recovered and preserved by separating the handwritten text from it.

In this paper, the classification is done on the structural and statistical differences between machine-printed and handwritten text and a tainted text can be differentiated into any of the three categories: purely printed, purely handwritten or “mixed” i.e. where parts or whole of handwritten and printed text are very close or overlap on each other.

## II. Pre-processing

The experiment was performed by using tainted documents where both the printed and handwritten text is in Bangla script. The collected document pages had a variety of handwritings from different people with different writing styles. The pages were then scanned and digitized to get the document images for working on. The gray-scale images were then binarized into a two tone image with pixels having ASCII value either as 0 or 255 where the former represents a black pixel and the later signifies a white space. The two-tone image thus obtained is then ready for the next stage of pre-processing.

The text matrix now obtained was labeled on the basis of its connected components using the 4-component Connected Component Labeling (CCL) technique in [8]. The connected components, as marked according to the algorithm, can be a single word, multiple joined words or an overlapping piece of intermingled text. The next step was to calculate the Bounding Box (BB) for each of the connected components. The Bounding Box is the minimum rectangle (or box) that can contain a connected component within it, like the figure below.

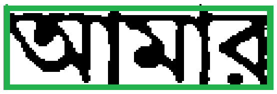

Fig 1. Connected Components inside the Bounding Box

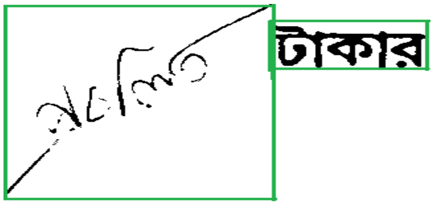

Fig 2. Mutually overlapping Bounding Boxes

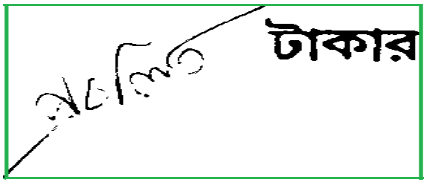

Fig 3. Bounding Box Reconstruction for this “mixed” region

There may occur cases where the BB’s of two adjacent connected components overlap each other and create a problem in distinguishing one from the other as shown in Figure 2. In that case, we treated both the connected components as a single entity and reconstructed the BB which can contain both of them together like Figure 3. So, the text components were now properly distinguished according to their connected components and their corresponding Bounding Boxes were also updated for future use.

## III. Classification Scheme

The classification strategy proposed here is a two stage classifier with the classification being done on the basis of the structural differences between the perfectly aligned and symmetric printed text and the coarse and skewed handwritings due to the human writing styles. The features are quite simple and easily detectable. The first stage classifier is based on three simple features that machine-printed text possesses and it separates the purely machine-printed text from the handwritten part of the document image. The second stage classifier discriminates the mixed regions of text, i.e. the BBs where machine printed text and handwritten ones are placed together, from the purely handwritten parts of text. The two stage classifier has been discussed below.

### A. First Stage Classifier

The first stage classifier is fed the updated list of BBs which contain any one of the three: purely printed connected components, purely handwritten components or mixed components. The task of the classifier is to mark the BBs which contain purely printed text and separate them from the handwritten and mixed text. For doing this, it uses very basic and intrinsic features that are possessed by Bangla text, especially the machine printed ones. The features sought after by the extractor are as follows.

### 1. Feature One: Headline

The common idiosyncrasy found among most of the Bangla alphabets is the presence of the “matra” or the headline on the upper region. And when two or more such alphabets are placed together adjacently to form a word, their headlines join together to form an extended headline - basically a horizontal run. Now the probability that a given word will possess one or two headlines in it is quite high as most of the Bangla alphabets, 32 out of the 50 present, consist of a headline. Moreover, 11 among the 12 most used Bangla characters also possess a headline as well. On top of that, there are 41 characters that can start a word in Bangla and of them 30 are found with a headline. So it is obvious that the percentage of Bangla words that consist of at least one prominent headline is very high and is found out to be 99.4%.

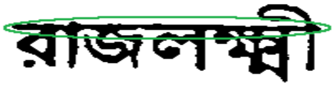

Fig 4: Headline or Matra in a Printed Bangla Word

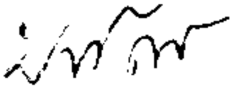

Fig 5: Absence of a Headline in a purely Handwritten text

This special feature of Bangla script can be utilized as a distinguishing feature between machine printed and handwritten text. In machine printed text the headline region is denoted by a straight horizontal run in the row of pixels. So, in a word that possesses a headline, there is at least one long horizontal run in its pixel row. But in handwritten text, proper care is not taken by the individual while writing and in most cases, approximately 91%, the headline is not straight at all and therefore doesn’t form a long horizontal run along its row of pixels. In some one off cases, some individuals do take proper care and write slowly and their headlines show a salient horizontal run in that case.

This variation between the headlines of handwritten and printed text has been taken as the first feature to be extracted in this scheme. A threshold T has been fixed and if the horizontal run of a word or text in a BB is found to be exceeding it, then it is likely to be a printed text and the BB is marked for being fed for the second feature extraction. And the BBs where such a horizontal run greater than T is not found, they are straight away kept to be fed to the second stage classifier. The value of T is taken as the average length of a Bangla alphabet of the printed text present in the document image that we are working on. So, the horizontal run in any text has to be at least longer than the average length of an alphabet for the text to be kept in contention for being a machine printed one. The BBs which satisfied this condition and were found to contain a long horizontal run were marked in the set $BB_1$ and the other set is named as $BB_2$.

### 2. Feature Two: Lowermost Point(s)

In Bangla, there are 39 consonants and 11 vowels. When a vowel is placed beside a consonant it usually takes a modified shape and is called a modifier thereafter. The modifier may be placed right, left, up or below the consonant which it modifies. So, the vowel ‘E’ when placed adjacent to the consonant ‘L’, it acts as a modifier and it changes its shape and the consonant now becomes ‘L’.

In machine printed Bangla text, the lower modifiers, if any, of the characters in a word lie on the same horizontal line, known as the lower line and the normal unperturbed characters of the word lie on another horizontal line, known as the base line. So basically the lowermost point of any character of a Bangla machine printed word lies in any one of the two horizontal lines mentioned above i.e. the base line or the lower line. But in handwritten text such care is not taken by the writer and the lowermost points of the characters of any word are distributed unevenly and therefore lie on more than two horizontal lines. And this feature has been used to further differentiate between printed and handwritten text.

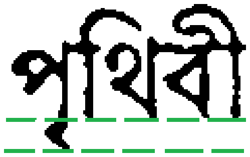

Fig 6: Printed text having its lowermost points on two horizontal lines only (the base line and the lower line)

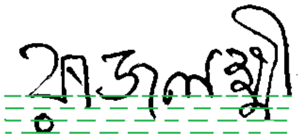

Fig 7: Handwritten text with some of its lowermost points on more than two horizontal lines

The extracted and marked BBs in $BB_1$ after the first feature extraction i.e. the headline, are now to be tested on the basis of this feature and the printed text can therefore be found out. For mathematical determination and programmatic execution the method mentioned in [4] has been used. The lowermost points of the different components of any word is divided here into two sets B and L based on their proximity to the base line row $B_r$ and $L_r$ respectively. A component which has its lowermost point at the row R belongs to B if $| B_r – R | <= | L_r – R |$ and vice versa. All such positions for the lowermost points of a text or connected component is calculated and is placed in either of the two sets B, comprising of $b_{1,}$ $b_{2,}$ $b_{3,\ldots,}$ $b_{m,}$ where m is the number of lowermost points belonging to B, and L comprising of $l_{1,}$ $l_{2,}$ $l_{3,\ldots,}$ $l_n$, where n is the number of lowermost points belonging to L. Now a feature called is the Character Lowermost Point Standard Deviation (CLPSD) is introduced and its value is calculated as:-

$$\text{CLPSD} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(b_i - b')^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(l_i - l}$$

$$\text{where} \quad b' = \frac{1}{m}\sum_{i=:}^{m} \quad \text{and} \quad l' = \frac{1}{n}\sum_{i=}^{n}$$

For machine printed text the lowermost points in B and L are almost all in the same row making the value of CLPSD very low i.e. the distribution is uniform. But for handwritten text, the distribution is usually uneven and therefore the individual row values in both B and L differ drastically from the mean value giving CLPSD a high value. The threshold used for the demarcation is kept as 0.1 of the mean height of the components in the text.

This feature helps to identify the set of BBs which contain printed text properly as in many cases where handwritten and printed text intermingle together, as in mixed regions, there can very well be a prominent headline due to the printed part of the text but the lowermost points of the components of such a BB are unevenly distributed giving it a very high value of CLPSD. For the BBs which have values of CLPSD less than the threshold they are identified as purely printed text and are separated out from the group and are put in the new set $BB_3$. The BBs which were found to have a high value of CLPSD are chalked out and mixed with $BB_2$ and we get a bigger set $BB_4$. This set is now fed to the second stage classifier.

### B. Second Stage Classifier

Modern pens have this quality that any text written using them are quite uniform in terms of thickness and rarely varies if the writing style of the user doesn't vary in between. In case of any sudden punctuation or an extra effort on some components of the text we can get words which have components with non-uniform thickness, but that is quite rare. In case of machine printed text and especially those in Bangla, due to the structural feature of certain alphabets and due to the curvy nature of the script, variable thickness is very familiar. Normally where any curve stops and changes its direction or a loop is encountered; a sudden change in the thickness of the text is noted. This feature can be utilized to distinguish the components of $BB_4$ into purely handwritten and mixed text.

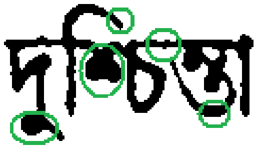

Fig 8: The different thicknesses in the different components of a machine printed Bangla word

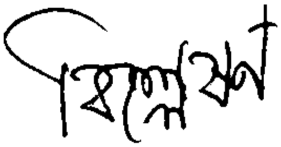

Fig 9: A handwritten Bangla word having almost the same thickness for all its components

The text in each of the bounding boxes in the set $BB_4$ is now analyzed and their respective contours are found out. Then the resulting set of bounding boxes is fed to the classifier. The BBs where no sudden variation of the thickness of the contour is noted, those are marked as purely handwritten text and in the BBs where such a huge change of thickness is seen, they are marked as the mixed regions. Therefore we get two distinct set of BBs which are either purely handwritten or mixed regions of handwritten and printed text intermingled together. For noting the varying thickness (t) in any component we take the thickness in every position across a full run of a component and calculate the mean thickness ($T_{mean}$) of the component. If the mean thickness is lesser than a threshold value we conclude that the component is handwritten otherwise printed.

$$T_{mean} = \sum_{i=1}^{p} t_i$$

where $t_1$, $t_2$,…, $t_p$ are the individual thicknesses for the p positions of the component.

The threshold for the variable thickness measurement is kept as 0.4 of the starting thickness for the connected component. That means for the thickness of the contour that we start with, an increase or decrease of only about 0.4 times of it is permissible. For free flowing handwritten text this condition is usually met as most of the times the writer writes the word in one complete go. But for machine printed text this varies heavily because of the presence of uneven zones of thickness like the headline, the modifiers, looping structures etc. So, there for printed text the threshold value is surpassed almost regularly and therefore we can separate it easily from the handwritten text. The initial thickness is recorded prior to every run across the contour of a component and is compared with the mean thickness found out in the end. If it is within the 0.4 bracket then it is marked as a handwritten text and if not then we conclude that it is a mixed region.

In the set $BB_4$ we were left with only purely handwritten text and mixed textual regions therefore the classifier now divides the set of BBs into two further sets: the BBs containing purely handwritten text, with a mean thickness lesser than or equal to the threshold, and the BBs which contain mixed regions of handwritten and machine printed text together and therefore has a mean thickness much greater than the threshold value. We mark the first of the two sets as $BB_5$ and the other one as $BB_6$.We already have a set of BBs which have purely machine printed text in $BB_3$, and now we get the two sets $BB_5$ and $BB_6$ as well marking the handwritten and mixed regions. Therefore the whole set of BBs has now been classified into three sets of purely printed, purely handwritten and mixed regions.

## IV. Results and Discussion

The scheme mentioned here was experimentally verified using a set of machine printed Bangla documents which were tainted by handwritings of different individuals. The set had over a total of 150 such samples of varying concentration of handwritten text. Most of the documents were pages from Bangla textbooks and question papers and some already scanned tainted documents were gleaned online. The printed text in them was of various styles and size. Each individual recorded three tainted pages of varying concentration and distribution of handwritten annotations: one sparingly tainted with handwritten text and machine printed text in the ratio of 1:20, one with a relatively moderate concentration of handwritten text with some of them overlapping the printed lines and the third one being heavily tainted with almost 30% concentration of handwritten text. A total of 41 individuals recorded their handwritings for this purpose.

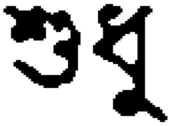

Fig 10: A machine printed Bangla word without any headline

The algorithm was coded and implemented on the set of document images using the C programming language. Roughly, the accuracy of the proposed approach was found out to be 88%. The errors were mostly encountered when some of the machine printed texts were devoid of any headline as exhibited in Fig 10. But as discussed earlier such a case is quite rare and therefore should be exempted. Also, as expected most of the handwritten texts were devoid of any prominent headline region and therefore were discarded from being printed at the first go. Moreover, the thickness variation was also much more pronounced in case of printed word components and were therefore instrumental in demarcating between the purely handwritten and mixed regions of text. Most importantly, as the features used in the approach is independent of the style and font of the printed text so the scheme is independent of such considerations. For different documents however the thickness was different and therefore the initial thickness was calculated while parsing each component.

| **No. of pages** | **No. of text boxes encountered** | **Correct Identification** | **Error** |
|---|---|---|---|
| 152 | 17560 | 88% | 12% |

The proposed approach can be also implemented on documents having Devnagari, Assamese or Punjabi as their script as all of them are similar to Bangla. Even, after making some modifications on the classifier, it can be used to identify handwritten text in Latin based script like English as well. For future course of work, the next step would naturally be to formulate and implement the algorithm for identifying as well as separating the overlapping handwritten texts from the machine printed text in the mixed regions.

## References

[1] Richard G. Casey and Eric Lecolinet, "A Survey of Method and Strategies in Character Segmentation", IEEE Transactions in Pattern Analysis and Machine Intelligence , Vol. 18 No. 7, July 1996

[2] A. Bishnu and B.B. Chaudhuri, "Segmentation of Bangla Handwritten text into characters by recursive contour following", Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 402-405, 1999

[3] U. Pal and B.B. Chaudhuri, "Automatic Separation of Machine Printed and Handwritten Text Lines", Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 645-648, 1999

[4] U. Pal and B.B. Chaudhuri, "Machine Printed and Handwritten Text Line Identification", Pattern Recognition Letters, V. 22 N. 3-4, pp. 431-441, 2001

[5] L.F. da Silva, A. Conci and A. Sanchez, "Automatic Discrimination between Printed and Handwritten Text in Documents", IEEE Xplore. Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), pp. 261-267, 2009

[6] A. Lemaitre, B.B. Chaudhuri and B. Couasnon, "Perceptive Vision for Headline Localization in Bangla Handwritten Text Recognition", Proceedings of the 9th International Conference on Document Analysis and Recognition, 2007

[7] U. Pal and S. Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proceedings of the 7th International Conference on Document Analysis and Recognition, 2003

[8] http://en.wikipedia.org/wiki/Connected-component_labeling