Synopsis

of

Development of OCR Techniques for Handwritten *Bangla* Text

Thesis submitted for the degree of

DOCTOR OF PHILOSOPHY (Engineering)

of Jadavpur University

By

SUBHADIP BASU

2005

Optical Character Recognition (OCR) involves computer recognition of characters from digitized images of optically scanned document pages. OCR systems facilitate large scale document transcriptions with huge saving of time and human effort and have potential applications in reading amounts from bank checks, extracting data from filled-in forms and interpreting handwritten addresses from mail pieces for automatic routing, and so on.

An OCR technique is either aimed at printed characters or handwritten characters of a particular *script*. The present work addresses certain key problems related to OCR of *handwritten* documents of *Bangla* script with successful development of techniques for skew angle detection and text line extraction [7], [9], word segmentation [8], and digit, character and word recognition [1], [2], [5], [6]. The work also includes development of applications for recognition of Pin codes from Indian Postal Documents [3], interpretation of handwritten form documents [4], and a Graphical User Interface (GUI) for editing machine recognized handwritten *Bangla* documents.

Despite the importance of *Bangla*, both as a script and language, evidences of research on OCR of handwritten *Bangla* characters, as observed in the literature, are a few in numbers. *Bangla* is the second most popular script and language in the Indian subcontinent. As a script, it is used for *Bangla*, Ahamia and Manipuri languages. *Bangla*, which is also the national language of *Bangla*desh, is the fifth most popular language in the world.

The existing techniques for extracting text lines from images of printed *Bangla* documents and segmenting printed *Bangla* words are not applicable to document images of handwritten *Bangla* text. Moreover, developments of sophisticated pattern recognition techniques are necessary for recognition of segmented characters of handwritten *Bangla* text. This is not only due to richness of the character set of *Bangla* script but also for dealing with wide variations of shapes and sizes of *Bangla* characters caused from person to person variations of writing styles. These are the major points of *motivation* behind initiation of the present work on the development of techniques related to OCR of handwritten *Bangla* documents.

Extraction of text lines from skewed document images is a critical problem related to OCR technology. Due to skewness of document images, text lines can not be directly extracted from them by identifying valleys of *horizontal pixel density histogram* of the images. Skewness is inherent in document images under all practical situations. It occurs primarily because a few degrees of misalignment of document pages with respect to the scanner or copier bed is unavoidable at the time of scanning. Text lines in the original documents may also be skewed for some special purpose. Moreover, skewness is inherent in handwritten text. So special techniques are required for extraction of text lines from document images.

Two different techniques are developed for extraction of text lines from multi-skewed document images under the present work. One of these techniques [7], may degrade the image quality. The other technique [9] is free from this difficulty. It is based on a novel idea of hypothetical water flows across the image frame from two opposite directions. Under this situation, stripes of areas containing text lines on the image are kept unweted as the hypothetical water flows are obstructed by the characters of these text lines. By labeling these stripes of areas, text lines from the document images are extracted under the said technique. The technique is experimentally shown to be applicable on document images of *Bangla* text and English text as well. Though the technique primarily targets handwritten documents, it can work well with printed documents also. On experimental observations, this technique is found to have successfully extracted 88.1% of text lines from samples of handwritten *Bangla* document images containing 405 text lines and 90.69% of text lines from samples of handwritten English document images containing 344 text lines. The technique is also tested on images of printed English and *Bangla* documents, which are all of uniform word and line spacings with skew angles assumed to vary within a specific range. The success rates of the technique as observed on these two types of sample documents of predefined features are cent percent. For successful operations of the line extraction technique, proper choices of values of flow angle θ and radius of structuring element k are necessary. It has been observed experimentally that certain text lines of some document images are divided into parts during extraction because of choices of a high

value of θ, compared to the average *skew angle* of each of these documents. This may also happen for large word spacings in documents. In such cases *user intervention* is necessary to tune the values of θ and k. Choices of the values of θ and k may be completely automated by computing the average skew angle and the average word spacing of the target document image prior to application of the said technique. The technique for computing the skew angle of a handwritten *Bangla* text line, which will be helpful for this purpose, is also developed under this work. For detection of the skew angle of a handwritten English text line, the usefulness of the *lower envelop* of the bottom most text line may be investigated. Considering all these, the task of completely automating the choices of values of θ and k may be taken up as an *extension work* related to the text line extraction technique described here.

A technique for separating touching text lines, developed under this work, is applicable on *Bangla* documents only. It may also be extended to include images of English documents, especially handwritten ones, by studying the utility of the lower envelope of an English text line as a *future course of work.*

Extracted words from this skew corrected text line need to be segmented into constituent characters prior to *recognition*. This can not be done simply by identifying valleys of the *vertical pixel density histograms* of the word images. For images of English words, there is a high chance that this technique may lead to *over segmentation.* It is also true for words of *Bangla* script. Appearance of consecutive characters in overlapping column positions on a text line makes the problem of *Bangla* word segmentation more complex compare to segmentation of English word. The problem becomes compounded with handwritten *Bangla* words because of variations in sizes and shapes of handwritten characters. Considering all this, a novel technique [8] for segmenting images of handwritten *Bangla* words is developed under this work. It is significant in the sense that the already existing technique of *Bangla* word segmentation based on isolating the *Matra* of a word for segmenting the word into its constituent characters cannot be applicable on handwritten *Bangla* words. Under the said technique, an approximate contour of the *Matra* of a word is determined first from the magnified horizontal pixel density histogram of the word image and certain tentatively chosen

segmentation points on the *Matra* is then located to isolate the possible word segments. The technique fails for the words, in which the *Matra* between two consecutive characters is not very prominent. To deal with such cases, two different approaches may be tried. In one approach, if the width of certain word segment exceeds the average value of the widths of individual segments of a word then the segment may be simply divided into two equal parts. The other approach may be to apply the segmentation technique afresh on one such segment. Which of these two approaches will be beneficial is a subject matter of *future investigation.*

Another major limitation of the said word segmentation technique is that it cannot handle *touching characters*. Such characters may quite often occur in handwritten words. The first one of the above two approaches, though a crude one, may also be helpful in separating touching characters. The possibility of application of the two existing techniques for segmenting touching characters, one using *water reservoir concept* [Pal et al. 2003] for segmenting handwritten numerals and the other for segmenting touching printed Devnagri and *Bangla* scripts using *fuzzy multifactorial analysis* [Garaian et al. 2002], may be explored in future for handling the problem of segmenting touching handwritten *Bangla* characters in a better way.

For recognition of handwritten digits, extracted from document images, two different approaches, one based on DS technique [1] and the other based on a Two pass technique [6] are developed under the present work. These are discussed in Chapter 4 of this thesis. On experimentation with a database of 6000 samples of handwritten *Bangla* digits, application of DS technique is found to improve recognition performances by a minimum of 1.2% and a maximum of 2.32%, compared to the average recognition rate of the two constituent Multi Layer Perceptron (MLP) based classifiers after 3-fold *cross validation* of results. The overall recognition rate as observed by this technique is 95.1% on an average. Improvement in recognition rate as observed by the DS technique is significant, especially in the context of high precision applications relating to handwritten digits. By including more non redundant features and other classifiers, a study may be conducted in future for further improvement of recognition rate under the DS technique.

The Two pass approach introduced under the present work is applied separately for recognition of handwritten digits and basic characters or their parts. Because of the exponential order of complexity in respect to the number of pattern classes, the experiments with DS technique is carried out only with the handwritten digit samples of *Bangla* script.

On being tested on the same database of 6000 samples of handwritten *Bangla* digits as mentioned before, the two pass approach has shown an average recognition rate of 95.25% after three fold cross validation of results. This recognition rate is achieved by improving the first pass recognition rate by 1.21% on average. The performances of the DS technique based approach and the two pass approach on handwritten digit samples are found to be more or less comparable.

A combination of the two approaches may be tried with samples of handwritten digits for further improvement of the recognition rate as a *future work*. In this work, the recognition results of the DS technique based approach may be used for producing the coarse classification decisions in the first pass of the two pass approach.

The Two pass approach, developed under the present work, is also applied for classification of individual word segments, already recognized as Basic character or part of a Basic character. The technique is tested on a database consisting of 300 samples of each of 36 classes of such handwritten word segments. That is, the database consists of 10,800 samples in all. All the samples of the database were written one after other, i.e. in isolation, at the time of data collection. The two pass approach is experimentally found to have shown an average recognition rate of 80.58% on this database after three fold cross validation of results. The recognition rate in the second and the final pass has been achieved after enhancing the first pass recognition rate by 2.28% on average. The recognition rate of the two pass classifier ensemble as observed here is not quite satisfactory for practical applications. It is so not only for intricacies of writing styles and richness of *Bangla* character set but also for improper segmentation of words and absence of algorithmically segmented samples from the training data.

To tune the MLP based classifiers with algorithmically segmented samples of word segments, the data set may be formed with samples of such segments for further

study of the technique. After recognition of words from the document images, they may be matched with the words of an already stored lexicon, if possible, to identify at least some of the misclassified words. Some arrangement may be made further for user intervention to rectify improper segmentation of words and improper classification of word segments.

Besides all these, in order to deal with intricacies of writing styles and richness of *Bangla* character set, a suitably chosen Adaptive Resonance Theory (ART) network, such as the ARTMAP or the Fuzzy ARTMAP [Carpenter et al. 1991, 1992], may be employed here as a pattern classifier because, unlike MLP, it continues to learn new data while retaining its stability by ensuring that the already learned knowledge is not erased or corrupted by this process. Centering this idea, a *future course of work* may be planned.

Finally, the present work has not addressed the problems of recognition of handwritten *compound characters* and some punctuation symbols of *Bangla* script. Compound characters are present in a large number in the *Bangla* script. The OCR technique developed under the present work can handle documents of *Bangla* text which can be only formed with the *Bangla* characters introduced in ‘Barna Parichay, Pratham Bhag (i.e. the first part)’, written by Ishwar Chandra Vidyasagar or ‘Sahaj Path’, written by Rabindranath Thakur. Both of the books are popular in *Bangla* speaking society as introductory texts of *Bangla* script and language. Compound characters of *Bangla* are introduced in the second parts of these two books. A potential area for extension of the present work is one which includes compound characters from the second parts of these books in document images supplied to *Bangla* OCR system. A suitably chosen ART network may again be a solution to this problem.

The present work can finally be considered as an important step towards the development of a complete OCR system for images of *handwritten Bangla* documents. It has not only addressed some fundamental problems of *Bangla* OCR, related to text line extraction, word segmentation and character recognition, but also opened up certain directions of future work with suitable hints for the solution.

List of Publications of the Author Related to the Thesis

[1] S. Basu, R. Sarkar, N. Das, M.Kundu, M.Nasipuri, D.K.Basu, “Handwritten Bangla digit recognition using classifier combination through DS technique,” accepted for publication in Proc. 1st International Conference on Pattern Recognition and Machine Intelligence, to be held at ISI Kolkata in Dec. 2005.

[2] S.Basu, N.Das, R.Sarkar M.Kundu, M.Nasipuri, D.K.Basu, "An MLP based Approach for Recognition of Handwritten ‘Bangla’ Numerals,” accepted for publication in Proc. 2nd Indian International Conference on Artificial Intelligence, to be held at Pune, in Dec 2005.

[3] S.Basu, S.S.Seth, P.Sarkar, B.Das, S.Dey, S.Ghosh, "Recognition of Pincodes from Indian Postal Documents," in Proc. of the National Conf. on Bioinformatics Computing, Patiala, March-2005.

[4] S.Basu, S.S.Seth, P.Sarkar, B.Das, S.Dey, S.Ghosh, "Development of a Multilingual Recognition Engine for Automatic Interpretation of Handwritten Form Documents," Proc. the 2nd National Conf. on Computer Processing of Bangla, Dhaka, Feb-2005, pp. 251-257.

[5] S.Basu, N.Das, R.Sarkar, M.Kundu, M.Nasipuri, D.K.Basu, "Handwritten Bangla Alphabet Recognition using an MLP Based Classifier," Proc. the 2nd National Conf. on Computer Processing of Bangla, Dhaka, Feb-2005, pp. 285-291.

[6] S.Basu, C.Chaudhury, M.Kundu, M.Nasipuri, D.K.Basu, "A Two Pass Approach to Pattern Classification,” proc. 11th Int’l Conference on Neural information Processing: ICONIP 2004 (LNCS 3316-Springer verlag), Science City, Kolkata, Nov. 20-25, 2004, pp. 781-786.

[7] S.Basu, C.Chaudhury, M.Kundu, M.Nasipuri, D.K.Basu, "Skew Angle Correction and Line Extraction from Unconstrained Handwritten Bengali Text," Proc. 5th

Intl. Conf. on Advances in Pattern Recognition, ISI, Kolkata, Dec-2003, pp. 271-274.

[8] S.Basu, C.Chaudhury, M.Kundu, M.Nasipuri, D.K.Basu, "Segmentation of Offline Handwritten Bengali Script," Proc. 28th IEEE ACE, Science City, Kolkata, Dec-2002, pp. 171-174.

[9] S.Basu, C.Chaudhury, M.Kundu, M.Nasipuri, D.K.Basu, "Text Line Extraction from Multi Skewed Handwritten Documents," Revised Paper Communicated to the Journal of Pattern Recognition.