Bengali Character Recognition using Feature Extraction

Thesis Paper for Department of Computer Science & Engineering

Of

BRAC University

By

Samiur Rahman Arif

Student ID: 04201007

December 2007

DECLARATION

I hereby declare that this thesis is based on the results found by myself. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree.

Signature of
Supervisor

Signature of
Author

ACKNOWLEDGMENTS

I have to thank specially Dr. Mumit Khan for over viewing my thesis work to completion. I also thank Mr. Abul Hasnat (Research Programmer) of Center for Research on Bangla Language Processing for carefully reading my pre- thesis paper and assisting me in every step.

ABSTRACT

The Character Recognition Problem can be assumed as a classification task in which a (portion of an) image is to be given a label among a set of possible labels that represent the characters under consideration. This is the fundamental aspect of feature extraction technique .This generic formulation may lead to quite different settings. Also, if the images of the characters can be obtained optically, we speak of *"Optical Character Recognition"* (OCR), as opposed to other settings in which input data is obtained by other means. OCR itself can be considered as a subtask of the more general problem of *"Document Analysis or Understanding",* where the goal is to obtain a symbolic representation of a digital image of the document under consideration that include not only the recognized text (characters), but also other document components and their relationship. In this thesis I will discuss various feature extraction techniques and later I will see how zoning can be used to build an efficient Bengali character recognition system.

Different feature extraction techniques are used to recognize different representations of characters for example binary characters, character contours, skeletons (thinned characters) or gray level sub images of each individual character. The feature extraction methods are distinguished in terms of

invariance properties, re-constructability and expected distortions and variability of characters. When a feature extraction method is chosen we need to consider it in terms of efficient application of the system and time consideration for building such system.

TABLE OF CONTENTS

Chapter 1

Introduction

Optical character recognition (OCR) is one of the most successful applications of automatic pattern recognition. In 1929, Gustav Tauschek obtained a patent on OCR in Germany, followed by Handel who obtained a US patent on OCR in USA in 1933 (U.S. Patent 1,915,993). In 1935 Tauschek was also granted a US patent on his method (U.S. Patent 2,026,329). The first commercial system was installed at the readers 1955, which, many years later, was donated by Readers Digest to the Smithsonian where it was put on display. Since 1950 OCR has been a very active field of research and development.

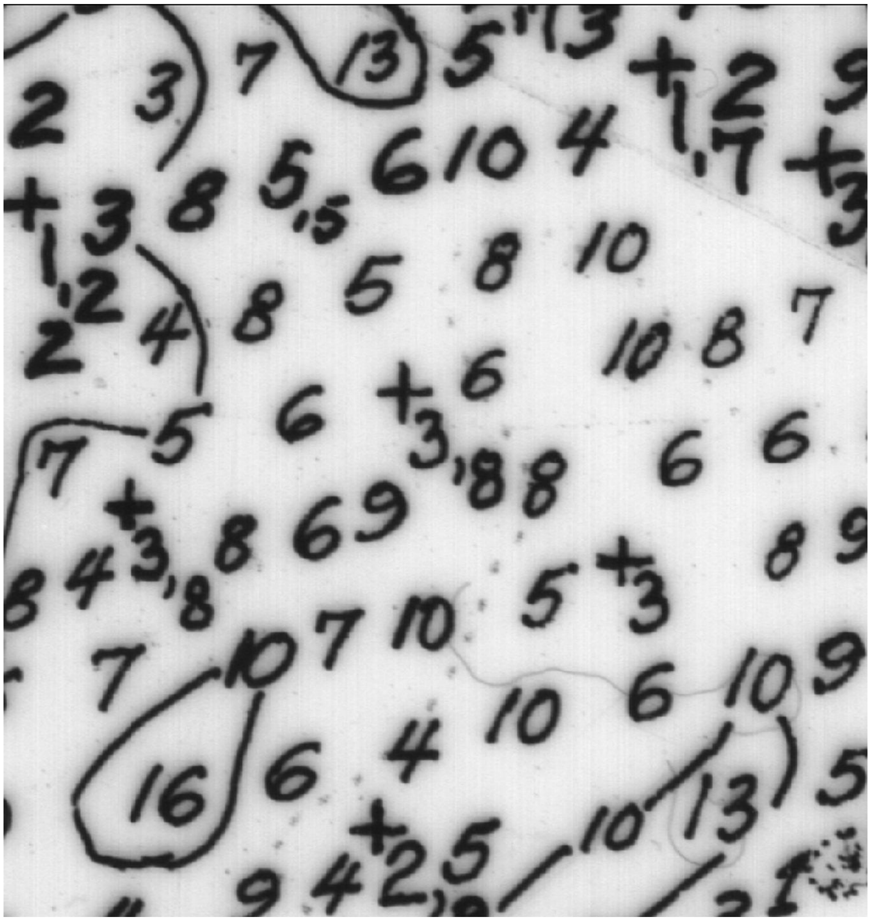


Figure 1: A gray scale image of a part of a hand printed hydrographic map

The current research on OCR is now concerning with the documents that are not well handled by the system, including severely degraded, omnifont machine printed text, and handwritten text. Also efforts are being made to achieve lower error rates without losing the speed hence efficiency.

Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance. This leads us to the question which available feature extraction method is the best for Bengali alphabets. An experimental evaluation of the methods for each application must still be performed strenuously to select the best method for a specific problem. In this process, one might find that a specific feature extraction method needs to be further developed. In this thesis paper a concise evaluation and performance of the system would be discussed in later chapters.

An OCR system has the following particular processing steps:-

(1) Gray level scanning at an appropriate resolution level typically 300-1000 dots per inch.

(2)Preprocessing:

a) Binarization (two-level thresholding) using a global or locally adaptive method.

b) Conversion to another character representation.

(3) Feature extraction.

(4) Recognition using one or more classifier.

(5) Contextual verification or post processing.

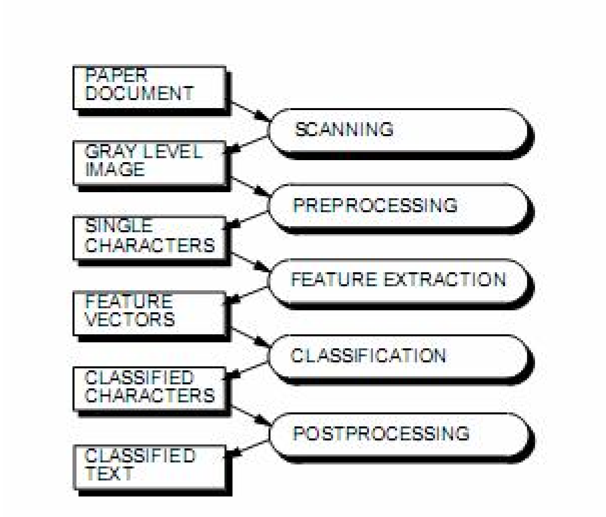


Figure 2: Steps in a character recognition system

Among the steps in character recognition system feature extraction is by far the most important step. There is an argument that only a limited number of independent

Features can be extracted from a character image so that which set of features in used is not important. The extracted features must be invariant to the expected distortions and variations. The other steps along the OCR processing also need to be optimized to obtain the best possible performance and these steps are not independent. The choice of the feature extraction method dictates the nature and output of the preprocessing steps.

Gray Scale subimage	Binary (solid outer contour)	Vector (Skeleton)
Template matching	Template matching	Template matching
Deformable templates	Contour profiles	Deformable Templates
Unitary Transforms	Projection histograms	Graph Descriptions
	Zoning	Discrete features
Zoning	Geometric moments	Zoning
Geometric moments	Spline Curve	
Zernike moments	Unitary Transformation	Fourier descriptors

Table 1: Overview of feature extraction methods for different image representation

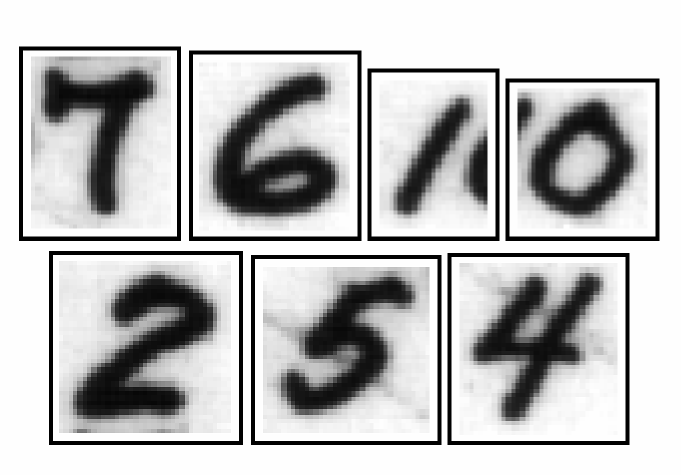
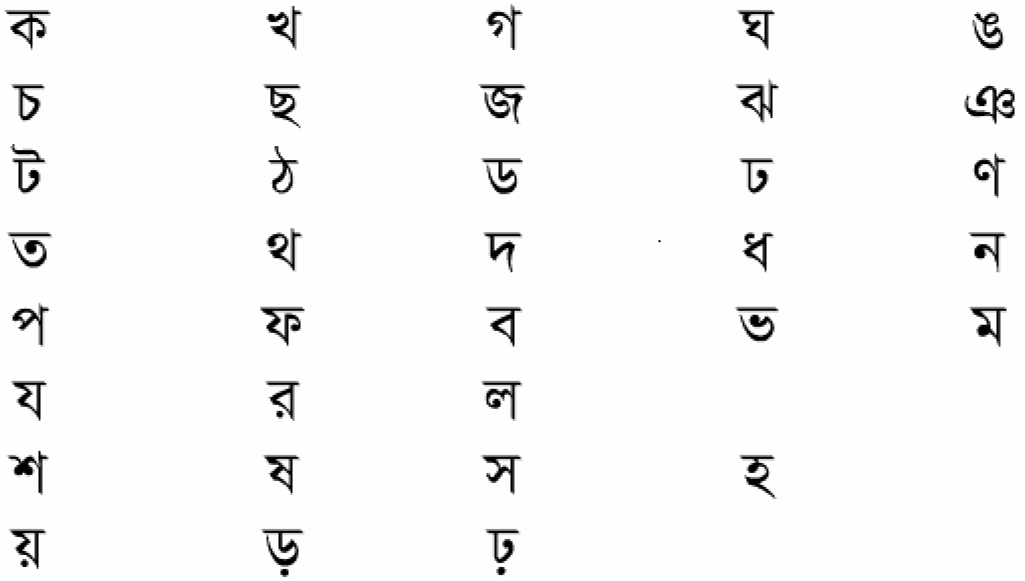


Figure 3: Gray Scale Images

For the Bengali Character Recognition system that was developed for this thesis binary characters were used.



ক	খ	গ	ঘ	ঙ
চ	ছ	জ	ঝ	ঞ
ট	ঠ	ড	ঢ	ণ
ত	থ	দ	ধ	ন
প	ফ	ব	ভ	ম
য	র	ল		
শ	ষ	স	হ	
য়	ড়	ঢ়		

Figure 4: Bengali alphabets, binary representation

In order to recognize many variations of the same character, features that are invariant to certain transformations on the character need to be used. For feature extraction methods, the characters can be reconstructed from the extracted features. By reconstructing the character from the extracted images one may visually check whether sufficient number of features for a single character is captured to reconstruct it. So we have to consider invariants and reconstructability of the feature extracted. However, not all variations among characters from the same character class can be modeled by using invariants. Size and translation invariance is easily achieved. The segmentation of individual characters can itself provide estimates of size and location but the feature extraction method may often provide more accurate estimates. Rotation invariance is important if the characters to be recognized can occur in any orientation. In our test case for this thesis we have not consider characters with multiple orientations. Skew invariance is useful in the test case as we have considered hand printed text, where the characters may be more or less slanted and multifont machine printed text where some fonts are slanted and some are not. If invariant features can not be found, an alternative is to normalize the input images to have standard size, rotation, contrast and so on. For some feature extraction methods the characters can be reconstructed from the extracted features. This property ensures that complete information about the character shape is present in extracted features. By reconstructing the character images from the extracted features, one may visually check that a sufficient number of features are used to capture the essential structure of the characters. Reconstruction may also be used to informally control that the implementation is correct.

Chapter 2
Different Feature Extraction Methods for Binary Images

Binary raster image is obtained by a global or locally adaptive binarization of the gray scale input image. The segmentation of the characters in done simply by isolating connected components. The problem arises when there are overlapping characters. Another problem occurs when the characters are fragmented in to two or more connected components.

The binary raster representation of a character is a simplification of the gray scale representation .The image function Z(x, y) now takes on two values either 0 or 1.In a way gray scale image methods can also be applicable to binary raster images. Generally, illumination invariants will be missing in binary raster images while recognition phase.

The goal of the OCR system is to preserve the relevant information about the character shape and discard some of the unnecessary information. Here, we will discuss some of the methods for feature extraction in binary raster images.

2.1 Template Matching

In binary template matching, several similarity measures other than mean square distance and correlation have been used. To detect matches let n_{ij} be the number of pixel positions where the template pixel x is I and the image pixel y is j with i,j ~~e~~{0,1} :

$$n_{ij} = \Sigma\ \delta m(i,j)$$

$$\text{where, } \delta m(i,j) = \begin{array}{l} 1 \text{ if } (x_m = i) \wedge (y_m = j) \\ 0 \text{ otherwise} \end{array} \qquad \ldots\ldots\ldots(2)$$

i,j ~~e~~{0,1} and y_m and x_m are the m-th pixel of the binary images Y and X which are being compared.

We can also use the mean square distance like gray scale images in binary images. Suppose a character with has Tj template is matched with different with vector and one with the most similarity another vector Tk is identified and if that is over a threshold then that character is given the label j else the character is unclassified. In the case of dissimilarity measure lowest dissimilarity is given the level k. A common dissimilarity is the mean square distance measure:

$$D_j = \sum (Z(x_i, y_i) - T(x_i, y_i))^2 \quad .....(3)$$

Where it is assumed that the character image and the template are of the same size and summation is taken over a m pixel image. For convenience the images are taken with same height and width size. If the value of D_j is below the threshold then we can label a testing character with a template character one.

2.2 Unitary Image Transform

The NIST form-based hand print recognition system uses the Karhunen Loeve transform to extract features from the binary raster representation. Its performance is claimed to be good and this OCR system is available in the public domain for English language. The complexity of the methods makes it difficult choice for Bengali characters.

2.3 Projection Histograms

Projection histograms were introduced in 1956 in a hardware OCR system by Glauberman. Today this technique is mostly used for segmenting characters, words and text lines , or to detect if an input image of a scanned text pages is rotated .For a horizontal projection $y(x_i)$ is the number of pixels with $x=x_i$ (figure 5). The features can be made scale independent by using a fixed number of bins on each axis and dividing by the total number of print pixels in the character image. However, the projection histograms are very sensitive to rotation and to some degree, variability in writing style. Also, important information about the character shape seems to be lost.

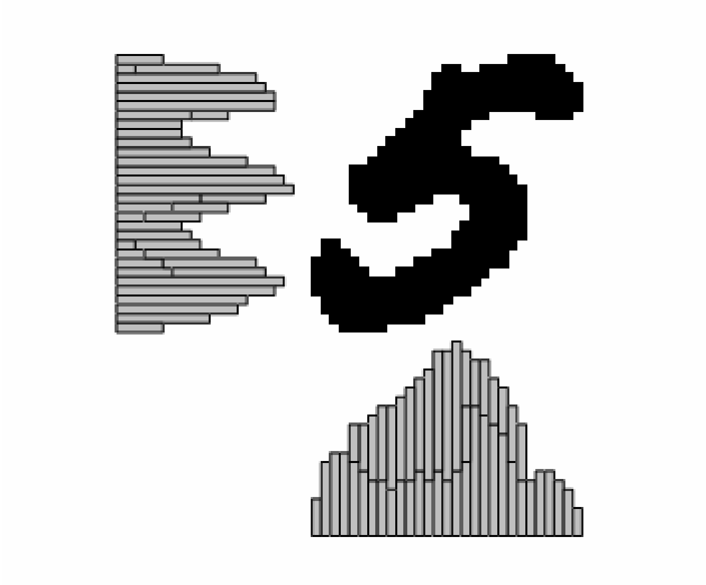


Figure5: Horizontal and Vertical projection histograms.

2.4 Zoning

The commercial OCR system CALERA uses zoning on binary characters. The system was designed to recognize machine printed character of almost any non-decorative font, possibly severely degraded ones. Both contour extraction and thinning proved to be unreliable for self touching characters. The zoning method was used to compute the percentage of black pixels in each zone. Additional features were needed to improve the performance. Zoning by the far the most easiest and efficient method to implement after Template matching.

2.5 Geometric Moment Invariants

A binary image can be considered a special case of a gray scale image with Z(x,y) =1 print pixels and Z(x,y) =0 for background pixels .By summing the N print pixels only it can be written as

$$M_{pq} = \Sigma(x_i)^p(y_j)^q \quad \ldots\ldots(4)$$

We can figure out the values for different characters using geometric moment for Binary images and then differentiate each other with the value. For characters that are not too elongated a fast algorithm for computing the moments based on the character contour exists giving the same values

Chapter 3

Factors for selecting one particular method

Before, selecting a specific feature extraction method the first consideration is where the system will operate. What kind of input characters would be recognized by the system?

Another consideration is whether the system would work for gray scale images or binary images. Also the requirement of throughput opposed to recognition requirements and also hardware availability.

Often, a single feature extraction method alone is not sufficient to obtain good discrimination power. An obvious solution is to combine features from different feature extraction methods. If a statistical classical classifier to be used and a large training set are available, discriminant analysis can be used to select the features with highest discriminative power. The statistical properties of such feature vectors need to be explored. The main disadvantages of using gray scale images are the memory requirement that is why binary images come handy.

CHAPTER 4
Methodology and Architecture

The Bangla OCR system works roughly as follows: Given a binary image of a document page, after skew correction, page segmentation can be conducted manually. After selecting a text region and determining whether the text is horizontally or vertically aligned, line segmentation based on projection. Analysis is conducted automatically. Once a character line is located, the OCR problem becomes similar to the automatic speech recognition (ASR) problem. Similar to what have been used successfully in the ASR area, we construct our character recognizer by adopting a powerful statistical pattern recognition paradigm and a strategy of dynamic programming search over a structural network representation of character image models and other possible linguistic knowledge sources. By considering the characteristics of printed Bangla documents, an innovative confidence-guided integrated search technique has been developed for the recognition of the whole character line. However, for this thesis most of the steps in pre processing into a single character for each character lines have been done manually. The emphasis was given solely into the character recognition.

CHAPTER 4

Data Collection

In order to achieve high recognition accuracy for documents of different types and with different quality, it is critical to collect a sufficient amount of representative training data which follow as vigorously as possible the sample distribution of the testing data to be recognized. The original documents are from varied sources, such as newspapers, magazines, journals, books, printed lists of characters generated from many popular font libraries available on the market. Apart from that different handwritten characters were taken into consideration. The document quality and font sizes vary widely among the various sources. The characters then were all taken into 16 by 16 sizes for the test. This was done using the best available photo editor in the market.

CHAPTER 6

The Test Results

After acquiring the data of twenty different Bengali characters a script written in Matlab was used to extract the features using zoning method. In this case we used 4 by 4 sizes of zones acting as the features for individual characters. For each character this zone would produce different illumination level depending on the number of pixels switched on or off. This is the basic behind this technique. However, all the characters were of same size and illumination level before the test for recognition. The test results are given below :-

Alphabets	Average Illumination per Zone
ক	3.9375
খ	3.6875
র	2.9375
প	2.4375
য়	4.0625
ভ	4.6824
ঔ	3.0625
	1.5625
ট	2.5555
ম	2.7500
ষ	3.6250
ফ	3.3125
চ	2.3125
অ	3.0000
আ	3.1478
শ	4.0890

These are the average illumination level in per zone for different characters .If we compare each character and with a threshold of 0.111 are good enough to distinguish each of them at least from the above group of characters. This test shows a pretty bright picture of using zoning method for character recognition of Bengali characters .However, a more comprehensive test for different illumination level was later done with similar results. The preprocessing steps may be automated to build the ultimate OCR system for Bengali characters. One key observation was that this character recognition system is not illumination invariant and rotation invariant. Although with the features a reconstruction is possible of the original characters.

Conclusion

A new and more efficient feature extraction method can be build using zoning and template matching together. Template matching can be used first then we can use zoning to make error rate less and the overall process more efficient. This a two stage recognizer with template matching is the first level significantly grouping Bengali characters into similar classes. Afterwards from each class zoning would be the method to select the exact character. Overall the results of this experiment were mixed. On the one hand, the initial results certainly are not of commercial quality. When only a couple of pixels differed between the unknown character and the reference, the results were fairly good, but larger differences often made the algorithm unable to correctly identify the unknown character. On the other hand, the low success rate is not indicative of the general algorithm, just the current implementation. There are many possible changes that could vastly improve the method's recognition abilities. With a few of these changes implemented, the mistakes of the algorithm would indeed be very similar to the types of mistakes humans would make. Thus general algorithm holds promise as a character recognizer that identifies characters in a manner similar to the way that humans identify characters.

References

1. Bahl L.R.; Jelinek, F.; Mercer, R.L. (1988). "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, No. 2. pp.179-190.

2. Baird, H. ; Thompson, K. (1990). "Reading Chess", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 12, No. 6, pp.552-558.

3. Baird, H. (1988). "Feature Identification for Hybrid Structural/Statistical Pattern Classification", Computer Vision, Graphics & Image Processing, Vol. 42, pp. 318-333.

4. Cover, T.M. ; Hart, P.E. (1967). "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory Vol IT-13, pp.21-27.

5. Duda, R.D. ; Hart P.E. (1973) "Pattern Classification and Scene Analysis", New York: Willey.

6. Ferri, F. ; Vidal, E. (1992). "Small Sample Size Effects in the Use of Editing Techniques", International Conference on Pattern Recognition (ICPR 92). The Hague. pp. 607-610.

7. Fisher, J.L. ; Hinds, S.C. ; D'Amato D.P. (1990). "A Rule-Based System for Document Image Segmentation", International Conference of Pattern Recognition 90 . pp.567-572.

8. Fletcher, L.A.; Kasturi, R. (1988). "A Robust Algorithm for Text String Separation

from Mixed Text/Graphics Images", IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 12, No. 6, pp.910-918.