Design of a view based approach for Bengali Character recognition

Sumana Barman[1], Amit Kumar Samanta[2], Tai-hoon Kim[3] and Debnath Bhattacharyya[3]

[1]Heritage Institute of Technology
Kolkata-700107, India
sumanabarman@gmail.com

[2]IBM India Pvt. Ltd
Kolkata, India
amitkumar.samanta@gmail.com

[3]Hannam University
Daejeon, Republic of Korea
tahoonn@empal.com, debnathb@gmail.com

Abstract

This paper presents a method to use View based approach in Bangla Optical Character Recognition (OCR) system providing reduced data set to the ANN classification engine rather than the traditional OCR methods. It describes how Bangla characters are processed, trained and then recognized with the use of a Backpropagation Artificial neural network. This is the first published account of using a segmentation-free optical character recognition system for Bangla using a view based approach. The methodology presented here assumes that the OCR pre-processor has presented the input images to the classification engine described here. The size and the font face used to render the characters are also significant in both training and classification. The images are first converted into greyscale and then to binary images; these images are then scaled to a fit a pre-determined area with a fixed but significant number of pixels. The feature vectors are then formed extracting the characteristics points, which in this case is simply a series of 0s and 1s of fixed length. Finally, a Artificial neural network is chosen for the training and classification process. Although the steps are simple, and the simplest network is chosen for the training and recognition process.

1. Introduction

The term 'Natural language processing' (NLP) is normally used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language. Natural language understanding is sometimes referred to as an AI-complete problem, because natural language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it. The definition of "understanding" is one of the major problems in natural language processing. The goal of the Natural Language Processing (NLP) group is to design and build software that will analyze, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing another person. NLP has some sub problems. Pattern Recognition is one of them.

The primary goal of pattern recognition is supervised or unsupervised classification. Among the various frameworks in which pattern recognition has been traditionally formulated, the

statistical approach has been most intensively studied and used in practice. More recently, neural network techniques and methods imported from statistical learning theory have been receiving increasing attention. The design of a recognition system requires careful attention to the following issues: definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, selection of training and test samples, and performance evaluation.

Automatic (machine) recognition, description, classification, and grouping of patterns are important problems in a variety of engineering and scientific disciplines such as biology, psychology, medicine, marketing, computer vision, artificial intelligence, aid remote sensing. But what is a pattern? Watanabe [7] defines a pattern "as opposite of a chaos; it is an entity, vaguely defined, that could be given a name." For example, a pattern could be a fingerprint image, a handwritten cursive word, a human face, or a speech signal. Given a pattern, its recognition/classification may consist of one of the following two tasks [7]: 1) supervised classification in which the input pattern is identified as a member of a predefined class, 2) unsupervised classification *(e.g.,* clustering) in which the pattern is assigned to a hitherto unknown class. Note that the recognition problem here is being posed as a classification or categorization task.

Interest in the area of pattern recognition has been renewed recently due to **e**merging applications which are not only challenging but also computationally more demanding**.** These applications include data mining (identifying a "pattern," e.g., correlation, or an outlier in millions of multidimensional patterns), document classification (efficiently searching text documents), financial forecasting, organization and retrieval of multimedia databases, and biometrics (personal identification based on various physical attributes such as face and fingerprints).

Pattern recognition in image processing encompasses several areas of research, viz., face recognition, signature recognition, text recognition, and fingerprint recognition. High accuracy text recognition or optical character recognition (OCR) is a challenging task for scripts of languages. The OCR research for the English script has matured. Commercial software is available for reading printed English text. However, for the majority of other scripts such as Arabic and Indian, OCR is still an active domain of research. For English and Kanji scripts, good progress has been made towards the recognition of printed scripts, and the focus nowadays is on the recognition of handwritten characters. OCR research for different Indian languages is still at a nascent stage. There has been limited research on recognition of Oriya, Tamil, Devanagari and Bengali.

Word recognition is the ability of a computer to receive intelligible input. The image of the written text may be sensed “off-line” form a piece of paper by optical scanning (optical character recognition). Alternatively, the movements of the pen tip may be sensed “on-line”.

With the development of digitizing tablets and micro computers, offline handwriting recognition has become an area of active research since last decade. This became a need because machines are getting smaller in size and keyboards are becoming more difficult to use in these smaller device. Recognition of printed character is itself a challenging problem since there is a variation of the same character due to change of fonts or introduction of noise. Recognition of Bengali character is a subject of special interest for us and some works have been done in this area.

There is a proliferation of on-line recognizers as compared to off-line recognizers. There are two main reasons for this disparity. On-line recognizers are easier to built, because the order of the pen-strokes is known, as well as timing information and also direction information of writing may be extract whereas Off-line recognizer are built to difficult.

But work on offline Bangla word recognition is very few. Many techniques are available for offline recognition of English, Arabic, Japanese and Chinese characters but there are only a few pieces of work available towards Indian characters although India is a multi-lingual and multi-script country. Also for offline printed word recognition very few works are there in Indian languages. For e.g. in Tamil language some works are there. Some works are available in English, Japanese and in Chinese.

The goal of this Project is to develop a system for new view based approach for recognition of offline printed Bengali word or character. First the approach I had used for recognition is without segmentation. Some morphological operations like thinning and thickening is needless here.

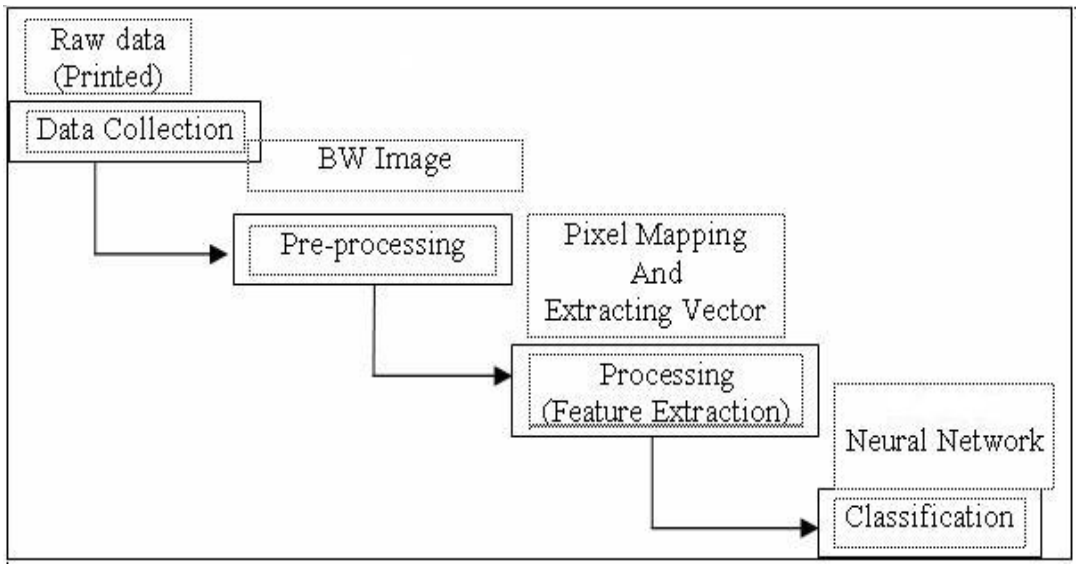

Figure 1. Diagram of Phases

There are twelve scripts in India and in most of these scripts the number of alphabets (basic and compound characters) is more than 250, which makes keyboard design and subsequent data entry a difficult job. Hence, offline recognition of such scripts has a commercial demand. Although a number of studies have been done for offline recognition of a few printed Indian scripts like Devnagari, Bengali, Gurumukhi, Oriya, etc. with commercial level accuracy, but to the best of our knowledge no system is commercially available for offline word recognition of Bengali script. In this project I propose a scheme for offline Bengali word recognition based on view based approach of word or characters.

Segmentation of Bengali word is very difficult with compared to English because of its shape variability of the characters as well as larger number of character classes. Automatic Word Recognition has been classified into two categories based on the presentation of the data into the system i.e. offline and online. In offline the scanned or printed text is fed to the system in digital image format for recognition. In online recognition approaches, the user writes using a digital device. The digitized samples are fed to the system as strokes as the sequence of 2D points i.e. with the x & y coordinate values. This project has been designed strictly keeping in mind all the documents are collected from various sources. Thus the scope of the project is limited to Bangla scripts only. The major challenges are here to recognize the offline printed text.

The problem has different phases (Figure 1):

- Printed Bangla character is taken for raw data
- Preprocessing (matra removal, binarization and scaling)
- Pixels are grabbed and mapped into specific area and a vector is extracted from the image containing the Bangla word or character
- Neural Network is used for classification

1.1. Optical Character Recognition

Script segmentation is an important primary task for any Optical Character Recognition (OCR) software. Through script segmentation a big image of some printed document is fragmented into a number of small pieces which are then used for pattern matching to determine the expected sequence of characters. In the implementation of Bangla OCR, the script segmentation may also play a vital role. But, for accurate and proper segmentation it is necessary to identify the properties of Bangla script as well as the exceptions.

1.2. Overview of Bangla scripts for OCR

Bangla, the second most popular language in India and the fifth most popular language in the world, is an ancient Indo-Aryans language. About 200 million people in the eastern part of Indian subcontinent speak in this language. Bangla script alphabets are used in texts of Bangla, Assamese and Manipuri languages. Also, Bangla is the national language of Bangladesh. The alphabet of the modern Bangla script consists of 11 vowels and 39 consonants. These characters are called as basic characters. Writing style in Bangla is from left to right and the concept of upper/lower case is absent in this script. It can be seen that most of the characters of Bangla have a horizontal line (Matra) at the upper part. From a statistical analysis we notice that the probability that a Bangla word will have horizontal line is 0.994. In Bangla script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right, both left and right, or bottom of the consonant. These modified shapes are called modified characters. A consonant or a vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters also exists in Bangla. There are about 280 compound characters in Bangla. To get an idea of Bangla basic characters and their variability in handwriting, a set of handwritten Bangla basic characters are shown in Figure 2.

Raghu Bengali Ekushey Godhuli

Figure 2. Variability of Bengali fonts.

Though Bangla is a very rich and old language, the matter of regret is its computerization has not yet gone much far. In fact, dedicated research and development for Bangla computerization has just been started from the last decade. In computerization of any

language, one of the vital tasks is to develop an efficient and effective Optical Character Recognition (OCR) system for the respected language. In order to store million pages of paper documents into electronic form, OCR is the key tool. Otherwise, if those are entered by typing manually, the efficiency, effectiveness and correctness will drastically fall down. As a result, all the effort will go in vain.

In any language, there are two types of written document. One is hand-written document and the other is printed document. Most of the characters of Bangla hand-written words are touching which is the prime bottleneck of this kind of documents. Besides, the shapes of Bangla hand-written characters are extremely diversified. Doing proper segmentation of these scripts is really tough. Hence, creating an efficient and useful OCR system for Bangla hand-written scripts is really a great challenge for computer scientists. On the other hand, the printed Bangla documents are much simpler than that of the previous one; because, in printed scripts variations in style of Bangla characters are limited. In case of printed documents, there are two basic classifications that are computer composed documents and type-machine composed documents. The primary alphabet of Bangla script is quite large compared to the alphabet sets of English and other western languages. It comprises of 11 vowels, 39 consonants and 10 numerals. The total number of symbols is approximately 300. Besides this huge quantity of symbols, there are various types writing style of those. All these aspects have thrown a great challenge to the researchers in developing a comprehensive OCR for Bangla printed scripts. For both on-line and off-line OCR, recognizing the Bangla scripts is really tough. Though, some sophisticated research and development has been done on recognition of handwritten Bangla numerals, but very few research works have been found on overall Bangla OCR.

2. Previous work

Various strategies have been proposed by different authors. Multi font character recognition scheme suggested by Kahan and Pavlidis[1]. Roy and, Chatterjee[29] presented a nearest neighbour classifier for Bengali characters cniploying features extracted by a string connectivity criterion. Abhijit Datta and Santanu Chaudhuri [30] suggestcd a curvature based feature extraction strategy for both printed and handwritten Bengali characters. B.B. Chaudhuri and U.Pal [4] combined primitive analysis with template matching to detect compound Bcngali characters. Most of the works on Bengali character are recognition of isolated characters. A very few deal with a complete OCR for printed document in Bengali. Mahmud,Rihan and Rahman[5] have used the chain code method of image representation.

Thinning of the character image is needless when chain code representation is used. Angshul

Majumdar[6] has used ‘a novel feature extraction scheme based on the digital curvelet transform’. The curvelet transform has been heavily utilized in various areas of image processing. The curvelet coefficients of an original image as well as its morphologically altered versions are used to train separate k–nearest neighbour classifiers. Segmentation in Bangla text is very common and there has been particular interest over the last decade. Segmentation of Bangla text is itself a challenging problem since there is a variation of the same character due to change of fonts or introduction of noise. Segmentation of Bangla character is a subject of special interest for us and many works have been done in this area. Various strategies have been proposed by different authors. A two phase approach is applied by Mahmud et. al. [2] in order to overcome the common problems related to the segmentation of printed Bangla characters. Their approach contains the text digitization and noise cleaning and skew detection and correction.

Segmentation of handwritten words into characters is one of the important components in handwritten text OCR. Bishnu and Chaudhuri [9] put forward a method for the segmentation of handwritten Bangla (an Indo- Bangladeshi language) text into characters. Based on certain characteristics of Bangla writing methods, different zones across the height of the word are detected. These zones provide certain structural information about the constituent characters of the respective word. In Bangla handwritten texts often there is overlap between rectangular hulls of successive

characters. As such the characters are seldom vertically separable. So, they have proposed a method of recursive contour following in one of the zones across the height of the word to find out the extents within which the main portion of the character lies. If the successive characters are not touching in the zone of contour following, the algorithm gives fairly good results.

A new approach to segment and recognize Printed Bangla Text using Characteristic functions and Hamming network was proposed by Mehedi et. al. [10]. The main difficulties in printed Bangla text recognition are the separation of lines, words and individual characters. A new algorithm has been proposed to detect and separate text lines, words and characters from printed Bangla text. The algorithm uses a set of characteristic functions for segmenting upper portion of some characters and characters that come under the Base line. It also uses a combination of Flood-fill and Boundary-fill algorithm for segmenting some characters that cannot be segmented using traditional approach. It is well-recognized that it is difficult to segment individual characters from handwritten words without the support from recognition and context analysis. One common characteristic of all the existing handwritten word recognition algorithms is that the character segmentation process is closely coupled with the recognition process. YI LU and Shridhar [14] represents three major portions, hand printed word segmentation, handwritten numeral segmentation and cursive word segmentation.

The higher recognition rates for isolated characters vs. those obtained for words and connected character strings well illustrate this fact. Casey and Lecolinet [15] have focused on this. Their aim is to provide an appreciation for the range of techniques that have been developed, rather than to simply list sources. Segmentation methods are listed under four main headings. What may be termed the “classical” approach consists of methods that partition the input image into sub images, which are then classified. The operation of attempting to decompose the image into classifiable units is called “dissection.” The second class of methods avoids dissection, and segments the image either explicitly, by classification of pre specified windows, or implicitly by classification of subsets of spatial features collected from the image as a whole. The third strategy is a hybrid of the first two, employing dissection together with recombination rules to define potential segments, but using classification to select from the range of admissible segmentation possibilities offered by these.

To take care of variability involved in the writing style of different individuals Pal and Datta [3] propose a robust scheme to segment unconstrained handwritten Bangla texts into lines, words and characters. For line segmentation, at first, they divide the text into vertical stripes. Stripe width of a document is computed by statistical analysis of the text height in the document. Next they determine horizontal histogram of these stripes and the relationship of the minimal values of the histograms is used to segment text lines. Based on vertical projection profile lines are segmented into words. Segmentation of characters from handwritten word is very tricky as the characters are seldom vertically separable. They use a concept based on water reservoir principle for the purpose.

One of the important reasons for poor recognition rate in optical character recognition (OCR) system is the error in character segmentation. Existence of touching characters in the scanned documents is a major problem to design an effective character segmentation procedure. A new technique is presented by Utpal and Chaudhuri [16] for identification and segmentation of touching characters. The technique is based on fuzzy multifactor analysis. A predictive algorithm is developed for effectively selecting possible cut columns for segmenting the touching characters. The proposed method has been applied to printed documents in Devnagari and Bangla: the two most popular scripts of the Indian sub-continent. The results obtained from a test-set of considerable size show that a reasonable improvement in recognition rate can be achieved with a modest increase in computations. But, it shows poor performance for documents like

a. old books: print and paper quality inferior due to aging;
b. copied materials: documents like photocopies or faxed documents, where print quality is inferior to the original;
c. newspapers: generally printed on low-quality paper, etc.

Casey and Lecolinet [l7] describes three strategies for segmentation namely (i) classical approach, (ii) recognition based segmentation approach, and (iii) holistic method. Among these in (i) image portion having character like properties are separated. In (ii) the system searches for component that match classes in the alphabet. In holistic method an entire component is considered for recognition. All

other methods are combination of the above three. In the classical approach to character segmentation, there exist quite a few studies. The contour analysis method [18, 19] does a local extreme analysis of the upper contour of the word image. The local minima that are deep enough from the adjacent local maxima are the possible segmentation points.

Segmentation points are often shifted horizontally to the right or left to obtain characters separated by vertical lines. Lecolinet and Crettez [20] analyzes the upper profile for segmentation. Some post processing is done to find some overshadowed segmentation points. In run length analysis [21] proposed by Bozinovic and Srihari, horizontal stretches of single-runs shorter than a threshold determined by average stroke width are detected and termed as single-run stretch. The single-run stretches having well defined peaks and valley points in the upper contour are removed. The remaining ones are split at the middle point.

The bounding box method proposed by Kimura et al[18,19] splits a word image into horizontally overlapping zones. A connected component analysis is applied to detect the boxes enclosing each connected component. The boxes are usually disjoint. Yanikoglu and Sandon[22] uses linear programming to segment off-line cursive handwriting. The successive segmentation points are found by evaluating a cost function at each point along the base line. The algorithm is guided partly by the global characteristics of the handwriting. In the recognition based segmentation, multiple sequences of overlapping components are generated as segmentation hypothesis without regard to its contents. The best hypothesis givesthe desired recognition. Casey and Lecolinet [23] scan the word from left to right for generation and verification of the segmentation hypothesis.

Kovalevsky [24] proposes a parallel method in which a lattice of possible feature to letter combinations is generated. An optimal path through the lattice gives the final recognition. Some works in the area of holistic approach have used Dynamic Programming techniques

[25, 26] to optimally align features like ascender, descender, directional strokes, loops etc. with features of lexicon word. Leroux [27] proposes a scheme using different features derived from the contour of the word images. Hidden Markov Models (HMM) of first and second order were used by Kundu et al [28] for handwriting recognition. In HMM, a word is represented as a matrix of transition probabilities of feature occurrences.

Md. Abdul Hasnat[8] represent the training and recognition mechanism of a Hidden Markov Model(HMM) based multi-font Optical Character Recognition (OCR) system for Bengali character. In our approach, the central idea is to separate the HMM model for each segmented character or word. The system uses HTK toolkit for data preparation, model training and recognition. The Features of each trained character are calculated by applying the Discrete Cosine Transform (DCT) to each pixel value of the character image where the image is divided into several frames according to its size. The extracted features of each frame are used as discrete probability distributions which will be given as input parameters to each HMM model. In the case of recognition, a model for each separated character or word is built up using the same approach. This model is given to the HTK toolkit to perform the recognition using the Viterbi Decoding method. The experimental results show significant performance over models using neural network based training and recognition systems.

3. Properties of Different Bangla Scripts

Bangla scripts are moderately complex patterns. Unlike simple juxtaposition in Roman scripts, each word in Bangla scripts is composed of several characters joined by a horizontal line (called ‘Maatra’ or head-line) at the top. Of-ten there may be different composite characters and vowel and consonant signs (‘Kaar’ and ‘Falaa’ symbols). This makes the development of an OCR for Bangla printed scripts a highly challenging task. There are some basic features or properties of any Bangla printed script.

i. Writing style of Bangla is from left to right.
ii. The concept of upper and lower case (as in English) is absent here.
iii. Among the characters, the vowels often take modified shapes in a word. Such characters are called modifiers (in Bangla ‘Kaar’).

iv. In a single syllable of a word, several consonant characters may combine to form a compound character that partly retains the shape of the constituent characters (e.g. Na + Da, Ka + Ta, Va + Ra-falaa, Na + Daa + Ra-falaa shown in Table 2).
v. Except very few characters and symbols(e.g. Ae, Oy, O, Ow, Kha, Ga, Ungo, Nio etc), almost all Bangla alphabets and symbols have a horizontal line at the upper part called 'maatra'. Some are shown in Figure 3a.
vi. In a word, the characters with 'maatra' remain connected together through their maatra' and other characters and symbols (e.g. Khondota, Bishorgo, Ungo, Ae, Oy etc) remain isolated in the word. Some are shown in Figure 3b.

Vowel and Consonant modifiers are possible (called 'Falaa'). These are shown respectively in Table 1a and Table 1b.

Table 1a. Bangla vowels and their modified forms.

Vowel	Corresponding Vowel Modifier
আ	া
ই	ি
ঈ	ী
উ	ু
ঊ	ূ
ঋ	ৃ
এ	ে
ঐ	ৈ
ও	ো
ঔ	ৌ

Table 1b. Bangla consonants and their modified forms.

Consonant	Corresponding Consonant Modifier
য	্য
র	্র
র	্র
হ	্ট

Table 2. Bangla compound characters and their modified forms.

Compound Character	Formation of the Character
ন্ড	ন + ড
ক্ট	ক + ট
ভ্র	ভ + ্র
ন্দ্র	ন + দ + ্র

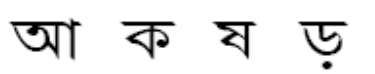

Figure 3a. Some alphabets with 'maatra' or headline.

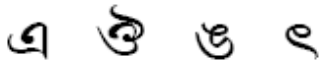

Figure 3b. Some alphabets without ‘maatra’.

vii. Each syllable in a Bangla word can be divided into three horizontal layers (shown in Figure 4). These are –
   a. Upper Layer containing the upper-extended potion of some alphabets and symbols (e.g. Oy, Uu, Ta, Tha, Chandra-Bindu etc). It starts from the top most abstract line of the syllable and runs till the ‘maatra’. It covers about upper 20% of the whole syllable.
   b. Middle Layer containing the major part of the alphabet or symbol. It begins from just below the ‘maatra’ and ends to an abstract base line. It covers almost
   c. 80% of the whole syllable.
   d. Lower Layer containing the lower extended portion of some alphabets and symbols (e.g. Ra,Uuu, Uu-Kar, Ree-Kar, Hashanta etc). It is situated between the base line of the middle layer and the bottom most abstract line of the syllable. It also covers approximately lower 20% of the whole syllable.

Figure 4. Three layers of Bangla scripts.

viii. Several characters including some vowel and consonant modifiers, punctuations etc have vertical stokes, too [16].
ix. All the basic alphabets, compound characters and numerals have almost same width. Whereas, the modifiers and punctuations vary in their width and height.
x. Most of the characters of Bangla alphabet set have the property of intersection of two lines in different positions as shown in Fig.3. Many characters have one or more corner or sharp angle property. Some characters carry isolated dot along with

Figure 5. Intersection points of some characters.

In the computer composed scripts, it is observed that around 50% characters become partially overlapped with one another. It implies that some alphabets and symbols in a word often enter into the region of their neighbour alphabets or symbols.

On the other hand, in the type-machine composed scripts, less than 10% of the total characters partially overlap with one another. Thus, the characters in a single word usually do not go into the region of their neighbour.

4. Our work

4.1. Pre-processing

Most of the classification techniques assume that the data is given in a predetermined form, which satisfy certain requirements as to quality, size, invariance, etc. However, these characteristics are commonly not satisfied by off-line printed data. The low quality of the data is due basically to the combination of some facts. One is the addition of noise during scanning. Skew generation at the time of scanning. Quality of the printed document is bad.

To overcome these problems, we use pre-processing, which involves matra removal, scaling and binarization, and/or extraction of the data. Pre-processing eliminates noise, reduces the amount of redundant information and facilitates encoding of raw data into feature vectors.

(I) Noise and Data reduction (II) Matra Removal (III) Scaling (IV) Binarization. Figure 6 shows the Block Diagram of the scheme.

4.2. Data Collection

We have collected five different samples of Bengali Characters by using various Bengali fonts like Avro keyboard, Tanmatra etc. for our input purpose. Then we have done matra removal. And we did also scaling. Now very common thing is image binarization. Then we have kept the characters separately within a folder.

4.3. Design Methodology

Here we have proposed a new system for character recognition. The system is based on the view-based approach. The system does not need thinning of analyzed character. The characteristic vectors taken from both top and bottom views. Here we are considering only two views that is top and bottom among four. The obtained characteristics vectors are used for classification with Artificial Neural Networks. The input of the experiment is a set of common bangle characters.

The main idea in this method is based on the view-based algorithm. The essential ideas of the view-based recognition system were presented in [11-12]. In case of characters only two of four views are analyzed, the upper and lower views. The most significant characteristic points are extracted from each image to form the feature vector describing the tested word or character. The features vectors are obtained from these images are the basis for further classification.

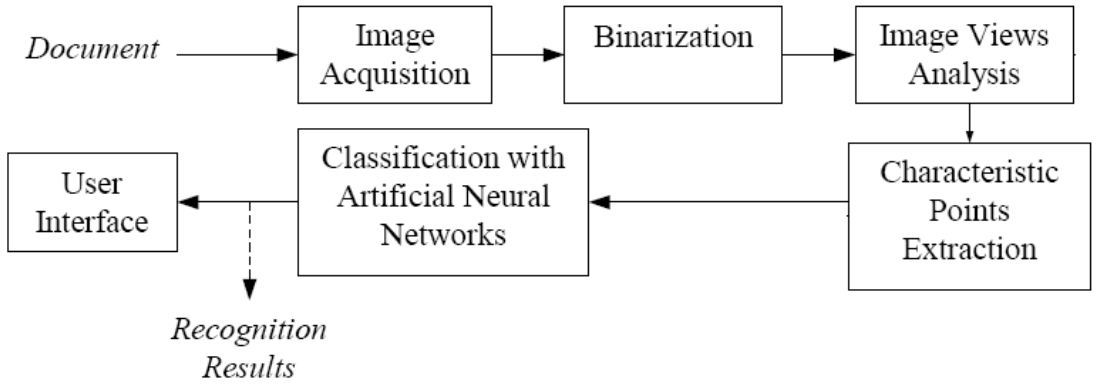

Figure 6. Block diagram

At first it was used for recognition of single characters. Then we will apply it to recognize whole words. This method is based on the fact, that for correct image recognition we usually need only partial information about its shape – its silhouette or contour.

Two "views" of each character are examined to extract from them a characteristic vector, which describes the given character. The view is a set of points that plot one of two projections of the object (top or bottom) – it consists of pixels belonging to the contour of a character and having extreme values of *y* coordinate – maximal for top, and minimal for bottom view (Figure 7).

Figure 7. Two views of sample characters

Both of the essential conventional stages of segmentation and thinning in the image processing techniques are unnecessary here. Only the shape of the character is analyzed.

Next, characteristic points are marked out on the surface of each view to describe the shape of that view. The method of selecting these points and their amount may vary. In our experiments 3 points are taken for each view of a character.

To find the characteristic points, one needs to divide the image vertically into a number of identical segments equal to the number of points we want to obtain. Next, we find the position of the highest and the lowest pixel in each segment – these are the points of top and bottom views (Figure 8).

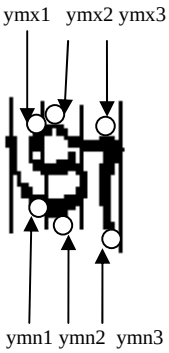

Figure 8. Characteristic points.

The next step is the calculation of *y* coordinates for the selected points. Thus we obtain two 3-element characteristic vectors describing the given character. Next, these two vectors together with their two values describing the aspect ratio of the picture (width by height) are transformed into a one 8-element vector, which describes the given character to be the base for further analysis. It is also possible to directly use this vector in the classification process.

5. Result and discussion

The experimental evaluation of the above technique was carried out using isolated bangla characters. The data was collected from different Bengali fonts like Avro keyboard, Tanmatra etc. We have collected five data samples of all bangla characters. Among these five samples three data samples have been used for training purpose and two data samples have been used for testing dataset. From the

experiment it was found that the overall accuracy of the proposed scheme was around **74.166%**. Some problems occur due to invariant size of bangla characters. We have checked that **accuracy** of **recognition** is around **75%**,not very high due to less number of training data set in Artificial Neural Network (i.e only 3 samples). Table 3 shows the percentage of recognition and error.

Table 3. Recognition results on Bangla Character.

Data	Correct Recognition rate	Error rate
120 Characters	74.166%	25.834%

Our goal is to test how our system performs in noticeably different conditions than the typical character recognition system deals with. Many other methods concentrate on finding a large number of characteristic data and a large number of examples for each class. Another way is followed in order to limit unnecessary growth of data and to show how our system performs with the reduced data set dimension. Because we are handling with the characteristics vector rather than the whole image matrix which will take a lot of space and therefore execution time will be more. Once we get the characteristics vector the image is of no more use. From then we will handle with only the characteristics vectors. The size of this vector is so small with compared to the original image matrix. And the execution time will be less while we are working with the characteristics vectors. If we work with an image of a piece of printed paper that will be a very large image. In that case if we work with the whole image no memory will be able to execute this program within a reasonable time. And also no cache memory will be there to provide such large space.

5.1. Bangla Basic Character Recognition Implementation

The experiments were carried out in Matlab 6.1, 1.4 GHz processor, with 512 MB RAM. The basic Image Processing operations were performed using Matlab's Image Processing Toolbox. Five Bengal fonts were used here. Different fonts are collected from different software like iLeap, Tanmatra etc. The entire data set consisted of 60 characters for each of the fonts and each of the 5 fonts, making a total of 300 samples.

6. Future Scope

In computerization of any language, one of the vital tasks is to develop an efficient and effective Optical Character Recognition (OCR) system for the respected language. In order to store million pages of paper documents into electronic form, OCR is the key tool. Otherwise, if those are entered by typing manually, the efficiency, effectiveness and correctness will drastically fall down. As a result, all the effort will go in vain. For Bangla, there is no good OCR solution till now. But, our government has huge quantities of Bangla paper documents that are so important that those should be stored for a long period of time. To do so, making electronic copies of those documents are unparalleled and it can be done by using a high-quality Bangla OCR system. But, to implement an OCR the foremost step in the recognition process is the script segmentation of the document image. Since, the written form of Bangla documents is more complex than that of many other languages, Bangla script segmentation is of great importance for creating a Bangla OCR system.

Our goal is to develop such system to improve the system efficiency. Till now we have seen that almost all methods handle with the whole image. Here we will use only characteristics vector instead of the whole image. View based approach has not been used for Bengali Character Recognition. First time we are going to use this approach for Bangla Character recognition.

7. Conclusion

The advantage of the proposed system is its efficiency. Because we do not need the whole image for execution ; characteristics vectors would be executed. We can handle with a large image i.e. a large scanned document. Here thinning is not required. We will apply this method to a word without segmentation. Disadvantage of this proposed system is that it is size dependent. We need matra removal. We have considered only basic characters not the compound characters. But we have future goals to apply this classification to recognize compound characters.

References

[1] S. Kahan and T.Pavlidis, "Recognition of printed characters of any font and size", *IEEE Trans. Pattern Anal.* Arid Mach.lnteN. 9,274-288, 1987.
[2] S.M. Milky Mahmud, Nazib Shahrier, A.S.M Delowar Hossain, Md. Tareque Mohmud Chowdhury, Md.Abdus Sattar,", *An Efficient Segmentation Scheme for the Recognition of Printed Bangla characters*", Proceedings of ICCIT, 2003, pp 283-286.
[3] U. Pal and Sagarika Datta, "*Segmentation of Bangla Unconstrained Handwritten Text*", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03).
[4]. B.B.Chaudhuri and U.Pal , "A Complete Printed Bangla OCR System", *Pattern Recognition* Vol-31, 531-549 ,1997. Graphics and Image Processing, NCCIS ,1997.
[5] J.U. Mahmud, M.F. Raihan and C.M. Rahman, "A Complete OCR System for continuous Bengali Character", TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region, 15-17 Oct. 2003
[6]Angshul Majumdar,"Bangla Basic Character Recognition using Digital Curvlet Transform", Journal of Pattern Recognition Research 1 (2007) 17-26
[7] *S.* Wntanabe, *Pltitruiz* Kccopiitiu!!: Htrnwr *n d M w h i c o l* . New York: Wiley, 1985.
[8] Md. Abul Hasnat" Research Report on Bangla OCR Training and Testing Methods" Working Papers 2004-2007
[9] A. Bisnu and B. B. Chaudhuri, "*Segmentation of Bangla handwritten text into characters by recursive contour following"*, Proc. 5th ICDAR, pp. 402-405, 1999.
[10] Md. Al Mehedi Hasan, Md. Abdul Alim, Md. Wahedul Islam "*A New Approach to Bangla Text Extraction and Recognition From Textual Image*", Proceedings of ICCIT, 2005.
[11] Rybnik, M., Chebira, A., Madani, K., Saeed, K., Tabedzki, M., Adamski, M.: A Hybrid Neural-Based Information-Processing Approach Combining a View-Based Feature Extractor and a Treelike Intelligent Classifier. *CISIM – Computer Information Systems and Industrial Management Applications.* WSFiZ Press, Bialystok 2003, pp. 66-73.
[12] Saeed, K., Tabedzki, M.: A New Hybrid System for Recognition of Handwritten-Script. *COMPUTING – International Scientific Journal of Computing.* Institute of Computer Information Technologies, Volume 3, Issue 1, Ternopil 2004, pp. 50-57.332 Advances in Information Processing and Protection
[13] Saeed, K.: A New Approach in Image Classification. *Proc. 5th International Conference on Digital Signal Processing and its Applications – DSPA'03*. Moscow 2003. Vol. 1, pp. 49-52.
[14] YI LU and M. Shridhar, "*Character Segmentation in Handwritten Words "*, Pattern Recognition, Vol. 29, No. 1, pp. 77- 96, 1996.
[15] Richard G. Casey and Eric Lecolinet, "*A Survey of Methods and Strategies in Character Segmentation*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18 No. 7, July1996.
[16] Utpal Garain and Bidyut B. Chaudhuri , "*Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis*", IEEE Transactions on Systema, MAN, and Cybernetics—Part C: Applications and Reviews, Vol. 32, No. 4, November 2002.
[17] Richard G. Casey and Eric Lecolinet, "*A Survey of Methods and Strategies in Character Segmentation*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18 No. 7, July1996.
[18] F. Kimura, M. Shridhar, Z. Chen, "*Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words*", Proc. of 2nd ICDAR, 1993.
[19] F. Kimura, S. Tsuruoka, M. Shridhar, Z. Chen, "*Context directed handwritten word recognition for postal service applications*", Proc. of 51h Advanced Technology Conference, 1992.
[20] E. Lecolinet and J. Crettez, "*A grapheme based segmentation technique for cursive script recognition*", Proc. of 1, ICDAR,1991.
[21] R.M. Bozinovic and S.N. Srihari, "*Off line cursive script word recognition*", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.1 1, No.], Jan. 1989.

[22] B. Yanikoglu and P.A. Sandon, "*Segmentation of Off-Line Cursive Handwriting using Linear Programming*", Pattern Recognition, Vol.3 1, No. 12, Dec. 1998
[23] R.G. Casey and G. Nagy, "*Recursive segmentation and classification of composite patterns*", Proc. 61h Intl. Conf. Pattern Recognition, 1982.
[24] V.A. Kovalevsky, "*Character Readers and Pattern Recognition*", Washington D.C.; Spartan Books, 1968.
[25] B. Plessis, A. Siscu, E. Menu and J.W.V. Moreau, "*Isolated handwritten word recognition for contextual address reading*", Proc. USPS 51h Advanced Technology Conference, Nov. 1992.
[26] T. Paquet and Y. Lecourtier, "*Handwritten recognition: Application on bank cheques*", Proc. of l, ICDAR, 1991.
[27] M. Leroux, J.C. Salome and J. Badard, "*Recognition of cursive script words in a small lexicon*", Proc. of 1, ICDAR, 1991.
[28] A. Kundu, Yang He and P. Barl, "*Recognition of handwritten word: first and second order HMM based approach*", Pattern Recognition, Vol.22, No.3, 1989.
[29]. A.K.Roy and B.Chatterjee, "Design of nearest neighbour classifier for Bengali character recognition", *J.IEEE* 30.1984.
[30]. Abhijit Dutta and Santanu Chaudhury, "Bengali Alpha- Numeric Character Recognition Using Curvature Features", *Pattern Recognition* Vol-26, 1707-1 720 ,1993.
[31] Tabedzki, M., Saeed, K.: A View-Based Töeplitz-Matrix-Supported System for Word Recognition without Segmentation. *ISDA'2006.* October 16-18, 2006, China.
[32] Khalid Saeed, Marek Tabedzki "New Experiments on Word Recognition Without Segmentation" conference proceedings of ACS-CISIM 2007 under the title "A Hybrid Word-Recognition System."
[33]R. C. Gonzalez and R. E Woods, *Digital Image Processing,* Addison-Wesley Publishing Company, 1992, pp. 7-9, 413-414.