An Approach to Recognize Bangla Digits from Digital Image

Abdul Kadar Muhammad Masum[1], Mohammad Shahjalal[2], Md. Iqbal Hasan Sarker[3], Md. Faisal Faruque[4]

[1]Assistant Professor, [2,3,4]Lecturer
[1.2]International Islamic University Chittagong, Bangladesh, [3]Chittagong University of Engineering & Technology
[4]University of Information Technology & Sciences

Summary
Pattern recognition is important in converting digital document image into electronic document. Today there are some OCR softwares to read the documents from scanned papers. We have designed a process to recognize the Bangla digits from digital images. Bangla digits are discrete when they are used in documents to write numbers. Each character has its own shape and design. We have used these characteristics of Bangla digits to recognize them from digital images. We have applied different morphological operation and segmentation techniques to separate each digit or character from the image. Then we have just drawn a single bit outline of the character, and generated a sequence based on the traversing path of the border outline. Then we applied a longest common subsequence operation between the generated sequence of the character and the pre-defined sequence of the 10 digits and detected the unknown Bangla digit. This is a sequence-based method.

Key words:
Digital image, Image processing, Bangla digit, LCS (Longest Common Subsequence), DFS (Depth First Search).

1. Introduction

The conversion of paper-based document information to electronic format is important for automated document delivery, journal distribution, document preservation, and other document related applications. In biomedical documents, especially documents in molecular biology, biochemistry, pharmacology, drug development, chemical analysis, etc. We have developed an idea to recognize a Bengali digit from an image of document. Our process generates a special sequence on the shape of the characters or objects. And analyzing the sequence we can decide that which digit it is.

2. Literature review

2.1. Pattern Recognition

Pattern recognition is a sub-topic of machine learning. It can be defined as "the act of taking in raw data and taking an action based on the category of the data" [1].

Most research in pattern recognition is about methods for supervised learning and unsupervised learning. Pattern recognition aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space.

A complete pattern recognition system consists of a sensor that gathers the observations to be classified or described; a feature extraction mechanism that computes numeric or symbolic information from the observations; and a classification or description scheme that does the actual job of classifying or describing observations, relying on the extracted features.

The classification or description scheme is usually based on the availability of a set of patterns that have already been classified or described. This set of patterns is termed the training set and the resulting learning strategy is characterized as supervised learning. Learning can also be unsupervised, in the sense that the system is not given an a priori labeling of patterns, instead it establishes the classes itself based on the statistical regularities of the patterns.

The classification or description scheme usually uses one of the following approaches: statistical (or decision theoretic), syntactic (or structural). Statistical pattern recognition is based on statistical characterizations of patterns, assuming that the patterns are generated by a probabilistic system. Structural pattern recognition is based on the structural interrelationships of features. A wide range of algorithms can be applied for pattern recognition, from very simple Bayesian classifiers to much more powerful neural networks.[2]

An intriguing problem in pattern recognition yet to be solved is the relationship between the problem to be solved (data to be classified) and the performance of various pattern recognition algorithms (classifiers). Van Dar Walt and Barnard investigated very specific artificial data sets to determine conditions under which certain classifiers perform better and worse than other.

Holographic associative memory is another type of pattern matching scheme where a target small patterns can be searched from a large set of learned patterns based on cognitive meta-weight.

Manuscript received June 5, 2011
Manuscript revised June 20, 2011

Typical applications are automatic speech recognition, classification of text into several categories (e.g. spam/non-spam email messages), the automatic recognition of handwritten postal codes on postal envelopes, or the automatic recognition of images of human faces. The last two examples form the subtopic image analysis of pattern recognition that deals with digital images as input to pattern recognition systems.
Pattern recognition is more complex when templates are used to generate variants. For example, in English, sentences often follow the "N-VP" (noun - verb phrase) pattern, but some knowledge of the English language is required to detect the pattern. Pattern recognition is studied in many fields, including psychology, ethology, and computer science.

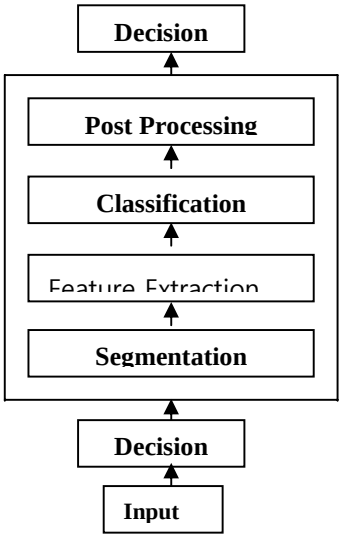

Fig. 1: Pattern recognition system.

2.2. Template Matching

Template matching is defined as the ideal representation of the object (or pattern) to be identified or classified within the image. The maximum correlation matrix value denotes where the template match with the image much properly than other. Moving the template to every position in the image and evaluating the degree of similarity at each correlation or position.
There are many disadvantages that associate with the template matching they are as follows:

- Presence of the shadows in the images
- Noise in analog or gray scale images
- Sampling and quantization noise
- Improper illuminations in the images

Also, the template matching is heavily relies on background illumination and discrimination which are the most varying factors in the images hence it is not a feasible option to go for. [3]

3. Proposed Approach of Bangla Digits Recognition

We have designed our approach for recognizing Bangla digits in the following steps

3.1. Proposed Model of Bangla Digit Recognition

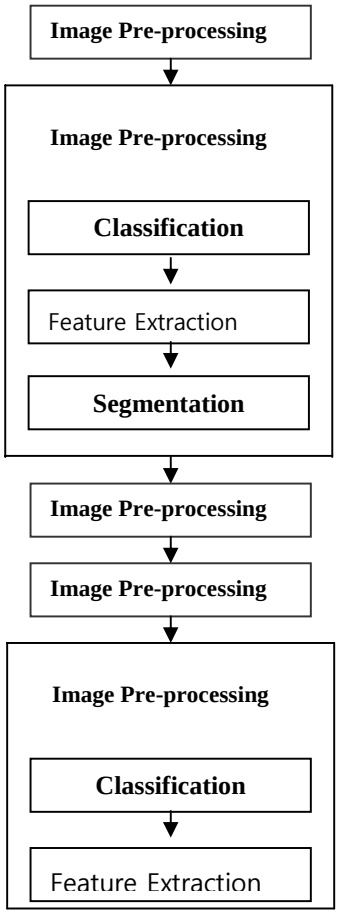

Fig.2 Pictorial diagram of proposed algorithm

3.2. Proposed Algorithm

Our total work can be separated into following parts.
1. Input an Image
2. Image Pre-processing
3. Bordering
4. Sequence Generation
5. Making Decision

3.2.1Input Image

Input image may be any digital scanned image, which contains Bangla numeric digits.

Fig3: Input Image

3.2.2 Image Preprocessing

Noise Reduction: We have applied mean filtering and high pass filtering in our process to reduce the noises of the image. [4]

Fig.4: Noise Reduced

Binarization: After reducing the noises from the image apply threshold operation to convert it into binary image.

Fig.5: Binary image of the input

Segmentation: In this step we have segment this image into numbers and digits. The image will be segmented into horizontal lines. Then each line is an individual image. Again each line will be segmented into numbers or integers depending on the spaces between the digits. Now the numbers are segmented. The digits will be separated from the numbers. And further operations will be done on the digits.

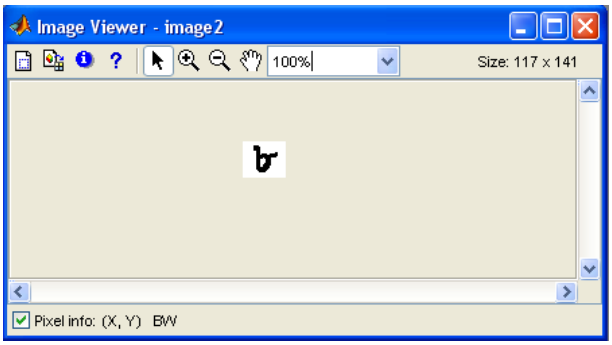

Fig. 6: Character segment

Resizing: This process is designed for specific size of fonts. So if a large font is given then this font must be resized into specific size. Again if this image is small then this should be resized in large size.

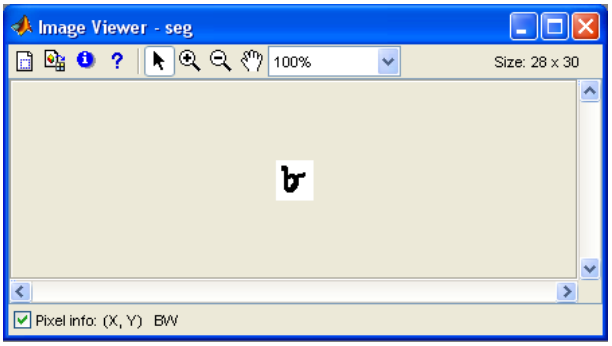

Fig.7: Resized character

Padding: Now the resized Digit will be padded. That is there will be an extra border added to the image. So the image will be increased by the size of two rows and two columns. This is necessary to avoid the boundary of the image while checking the eight neighbors of each pixel.

3.2.3 Bordering

In this stage a single pixel border of the digit will be generated. To generate single pixel border we have run morphological edge detection operation by using Sobel [5] operator. The found border is several pixel thick. Then we applied the iterative thinning operation till one pixel border is not found. As each digit has different shape so, the borders will also be of different shape. And this border must be a connected graph.

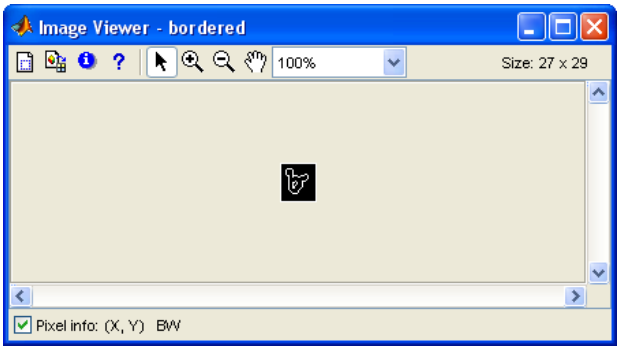

Fig.8: Bordeing

3.2.4Sequence Generation

A sequence will be generated on the basis of the border. To generate the sequence we have decided to run the Depth First Search (DFS) [6]. We always start from the bottom left most point of the border. It starts traverse in the left top direction and stops in the previous point of starting point.

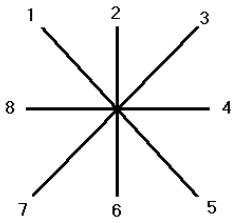

Fig.9: Sequencing Direction

1 8 8 1 8 1 1 1 2 1 2 2 2 ....

Fig.10: Generating sequence

3.2.4Making Decision

Now it is ready to take decision which digit it is. Steps are discussed below.

3.2.4.1Compare Sequence

We have defined some standard sequences for the ten Bengali digits (০,১,২,৩,৪,৫,৬,৭,৮,৯). To defining the sequences we have generated three sequences for large, medium and small font of each digit. To take decision we preferred to run the longest common subsequence (LCS) [7] on the new generated sequence and the pre-defined sequences.

১: 1 2 3 3 3 2 1 1 1 2 ...
২: 1 8 1 1 8 1 8 8 1 2 ...
৩: 1 1 1 1 2 1 1 2 2 2 ...
৪: 1 1 2 2 3 3 2 1 1 3 ...
৫: 1 8 8 1 8 1 1 1 2 1 ...
৬: 1 2 1 1 1 2 2 2 2 3 ...
৭: 1 2 2 2 1 8 1 1 1 2 ...
৮: 1 2 3 2 2 1 3 1 1 2 ...
৯: 1 1 2 2 3 3 3 1 1 1 ...
০: 1 1 8 1 2 1 1 2 3 3 ...

Fig.11: Pre-defined Sequences

LCS gives the maximum subsequence that refines the order of the sequence for that special digit. So when LCS runs through the pre-defined sequences it finds out the common sequence with all the predefined sequences. As LCS gives the length of subsequence so there is a matter of order of sequence. So longest subsequence matches with the sequence of that digit. So, after comparing with all the sequences we can decide the digit based on the length of LCS.

4. Performance Evaluation

There are given achieved result on basis of testing with different sized fonts

4.1. Testing with different sized fonts

We have tested our implemented system with various sized digits and found the following results.

Size	০	১	২	৩	৪	৫	৬	৭	৮	৯
9	N	Y	N	Y	Y	N	Y	Y	N	N
10	Y	Y	N	Y	Y	Y	Y	Y	Y	Y
15	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
20	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
40	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
200	Y	Y	N	Y	Y	Y	Y	Y	Y	N

4.2 Result

We have tested our program with different Bangla digits of different font and size. It can successfully recognize most of the digits from the image. It has some limitations those are discussed here.

5. Advantages for proposed technique

- Here we don't have used any Neural Network.
- No need to train the program to recognize the digits.
- Simple Algorithms.
- Easy to Implement.

6. Conclusion

Generally digit recognition techniques are used in optical character recognition software. Usually character recognition techniques use Artificial Intelligence and Neural Network to performing the recognition. Here we discussed a system which didn't use any Neural Network and Artificial Intelligence. So, our proposed system will work faster with limited database, where Neural Network takes a larger Database.

7. Future Works

The technique we have described here can be applied to recognize any discrete character. Just we have to set a well defined sequence of the characters in the program. English alphabets, numerals and Arabic numerals are also discrete when they are written in documents. It can be used in optical character recognition for Bangla & Arabic numbers and English texts and numbers also. Accuracy can be increased by taking different sequences from different sized character. Performance can also be increased by doing the image pre-processing efficiently.

References

[1] C.M. van der Walt and E. Barnard,"Data characteristics that determine classifier perfromance", in Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa, pp.160-165, 2006. Available [Online] http://www.patternrecognition.co.za
[2] Text line segmentation and word recognition in a system for general writer independent handwriting recognition Marti, U.-V.; Bunke, H.
[3] Wikipedia: http://en.wikipedia.org/wiki/Morphological_image_processing
[4] R. Gonzalez and R. Woods Digital Image Processing, Addison-Wesley Publishing Company, 1992
[5] http://www.cee.hw.ac.uk/hipr/html/thin.html
[6] ww3.algorithmdesign.net/handouts/dfs.pdf
[7] LCS 161: Design and Analysis of Algorithms Lecture notes for February 29, 1996

Abdul Kadar Muhammad Masum is a graduate of B.Sc and MSc in Computer Science and Engineering from International Islamic University Chittagong, and United International University, Dhaka, Bangladesh, respectively. He has been serving as a faculty member in IIUC since 2005. Now he is an Assistant Professor of Department of Business Administration, International Islamic University Chittagong, Bangladesh. Moreover, he has mentionable research works on Data Mining, E-commerce and E-governance and has a great interest in MIS field. He has also experience in research paper presentation in many International and National Conferences.

Mohammad Shahjalal has received his B. Sc. degree in Electrical and Electronic Engineering from Chittagong University of Engineering and Technology in 2009. Now he is serving as a lecturer in the department of Electrical and Electronic Engineering of International Islamic University Chittagong since February 28, 2010. His research interests focus on Artificial Intelligence, image processing, wireless communication, wireless networking, digital signal processing.

Md. Iqbal Hasan Sarker has received his B.Sc degree in Computer Science & Engineering from Chittagong University of Engineering & Technology (CUET) in 2009. Now he is serving as a Lecturer in Chittagong University of Engineering & Technology (CUET) in the department of Computer Science & Engineering since September 19, 2010. His research interests focus on Artificial Intelligence, Digital Image Processing and Natural Language Processing, Wireless Communication.

Md. Faisal Faruque has received his B.Sc degree in Computer Science & Engineering from International Islamic University Chittagong (IIUC) in 2007. Now he is serving as a Lecturer in University of Information Technology & Sciences (UITS) in the department of Computer Science & Engineering since February 18, 2008. His research interests focus on Artificial Intelligence, Digital Image Processing and Natural Language Processing.