An Approach towards Recognition of Size and Shape Independent Bangla Handwritten Numerals

Amitava Choudhury[1], Joydeep Mukherjee[2]

[1]School of Education Technology, Jadavpur University,
Kolkata
a.choudhury2013@gmail.com

[2] School of Education Technology, Jadavpur University,
Kolkata
joym761@gmail.com

Abstract: ***This paper presents an approach to recognize off-line Bangla numeral. Today there are many OCR used to recognize Bangla numeral. The recognition of handwritten character is still a challenging work in the field of pattern recognition. Numeral recognition in pattern recognition is the process to identify the given character according to the predefined character set. The difficulties of recognition of handwritten Bangla numeral are that they are different in shapes and sizes which are much curved in nature. . We try to establish a process to recognize such handwritten Bangla numerals having different shape and size. The input scanned image is first to be binarized. Then we have segmented all the ten digits of Bangla numerals to identify each and individual digit from a scanned image. We have used line segmentation to extract the feature from each numeral based on templates. A high correlation coefficient method provides a successful match between the test data and training data.***

Keywords: Hand written Bangla numerals, Pattern recognition, connected components, Templates matching.

1. Introduction

In India, due to the multifarious application potentials of Bangla characters' recognition, this has been a popular research topic for long. While going through the research work, we understood that there are lots of techniques on Bangla numeral recognition. Our proposed technique as presented here shows much more efficient and the technique is independent of size and shape of the Bangla numerals and less time consuming towards recognition.

Optical character recognition, in short OCR, is defined as an electronic translation into readable text, in which so ever form of text it may be – handwritten, typewritten or printed. There are many OCR available for different languages. In India, there are more than 18 languages and Bangla is one of the very popular languages. Extensive experimentations in this field has been carried out over the years, with that of handwritten characters as well as particular Bangla numerals [1]. Some of the available works include [2] for English, [3] for Chinese, [4] for Arabic, [5] for Korean, and [6] for Kanji script. It is seen that more than 200 million people in the world are used Bengali language to speak. Bangla language is very popular in the state of West Bengal in India and it is also the national language in Bangladesh. The numerals in bangali language are very curvy in nature with different shape and size. In this paper, a template matching method is proposed using correlation coefficient computation for recognition of hand written Bangla numerals. In section 2, we discuss about the brief description of literature survey. In section 3, we describe the proposed methodology. In section 4, we describe about the experimental results and finally in section 5, conclusion is drawn.

2. Brief Literature Survey

The literature in Bangla numeral is widely used and it based on two major areas of research, off-line and on-line systems.

The offline character recognition is presented in [7] to describe handwritten numeral recognition using template matching technique. In [8] C.Vasantha Lakshmi and other describe an approach to recognition of devnagari letter. R. V. Kulkarni and P.N. Vasambekar in [9] describe segmentation technique to recognize connected bangla digits.

Optical Character recognition (OCR) is an approach to carry out text recognition. There are several OCR methodology proposed to recognize Indian languages. BB Choudhury and Pal describe in [10] about banlga printed numeral. K.Roy, C. Chaudhuri, U.Pal and M. Kundu [11] describe an approach on the effect of varying training set sized on the recognition performance with handwritten bangla numerals.

M. Ziaratban, K. Faez, F. Faradji [12] describes an effective approach for character recognition named Template Matching. Using the technique extracts features by searching the selected templates in input images. The position of the best matched template is found and saved using the matching technique. The match amounts of templates can also be used as a feature.

3. Proposed work

In Bengali language there are ten digits to represent the number system consists of 0 to 9 as shown in Figure 1. Each digit is different from other in term of shape.

Figure 1: Ten Digits in Bengali

We divided the entire work into several steps as stated below

- Read the scanned input handwriting image
- Image preprocessing
- Feature Extraction

- Template Matching
- Classification

3.1 Input an image

Input image may be any handwritten scanned image, which contains Bangla numeric digits. The image may contain single or connected digit(s) and size of letter may differ from each other as show in Figure 2 and Figure 3.

Figure 2: Small shape single digit

Figure 3: Connected large digit

3.2 Image preprocessing

3.2.1 Noise Reduction

We use median filter technique to reduce noise. Median filtering is more successful than convolution if the object is to reduce noise as well as preserve edges. The noise in the images like tiny dots is eliminated by removing all region having less than 30 pixels. In Figure 4 the input image is mobile captured image with noise and Figure 5 shows the same image after noise removal.

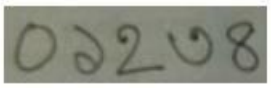

Figure 4: Input image

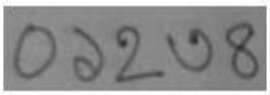

Figure 5: After noise removal

3.2.2 Binarization

Upon reducing noise from the image, the threshold is applied in order to convert the input image into a binary form. We have used Otsu's method for the process of binarization. Figure 6 shows the binarized image:

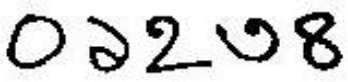

Figure 6: Binarized image

3.2.3 Segmentation

The image will be segmented into lines in two terms named first line and remaining line, such that each line happens to be an image of individuals. Further each line will be segmented into numbers depending on the spaces between the numerals. Now the numerals are segmented.

3.3 Feature Extraction

We have used the template matching technique to extract feature. It is very effective to use character reorganization where templates don't have such good features. The concept of templates matching technique is that it matches the similar feature of input image with the templates that store in dataset. We use different dataset consisting different Bangla digits as training database. In our approach, all the templates are matched with each segmented numeral. Handwritten numerals stored in training dataset in size of 32x32 pixels. The input image is matched with whole templates to find the absolute result using correlation method. If the input image is having same pixel definition of the trained dataset then successful recognition is done. We use all segmented characters are resized to a standard size of 32 x 32 pixels. The sample images are stored in white background and black font but while we use recognition we invert the image. A sample of dataset is shown in Figure 7. The dataset is collected from [13]

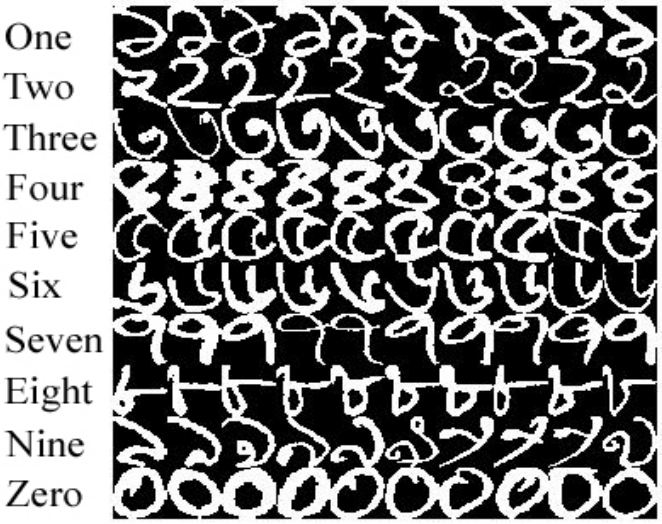

Figure 7: Bangla numeral dataset

3.4 Template Matching

Template matching is the technique in which pixel definition of presorted patterns are sought in an image. The templates are describing relationship between the regions. The dataset containing different set of numeral in Bangla Character and store them in a 10x10 matrix.

3.5 Classification

When an image is tested as input image, the system first loads the template. Resize the input image as same size of the template that is 32 x 32. Compute the numbers in the template file and count the connected components label. Compute correlation between the template and the input image. A perfect matching gives a correlation coefficient of 1.0. Then only the input image can be idententified by matching numeral from the database. In Figure 8 shows the steps of classification.

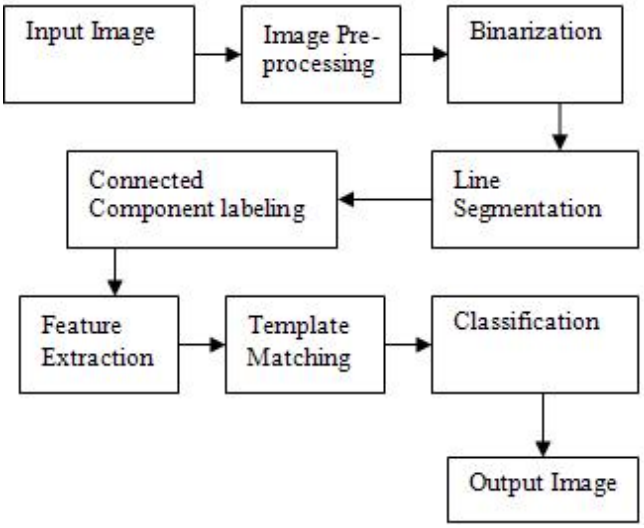

Figure 8: Recognizer components

4. Experimental Results

We have used Matlab 7.4.0 software for our implementation. The experiments were performed on many test images having different types of numerals in Bengali literature. First we read the scanned input image Figure 9 and then eliminate the noises from the scanned image. The preprocessed input image Figure 10 is segmented line by line by scanning and indicating the left-top edge and right-bottom edge of the line from the input image as shown in Figure 11 to Figure 14. The segmentation is extracting a line of numerals. All sides of these lines of numerals will exactly touch the line boundary of the bounding box. Then we segment on individual numerals from each segmented line. The scanned image is resized to 32x32 pixels. We collect hand written numeral for input from set of different people and a selected dataset. Figure 15 shows the output of the scanned input image.

Figure 9: Input image

Figure 10: Binarized numeral

Figure 11: Segmented zero

Figure 12: Segmented one

Figure 13: Segmented two

Figure 14: Segmented three

Figure 11 to 14 are the inverted segmented images.

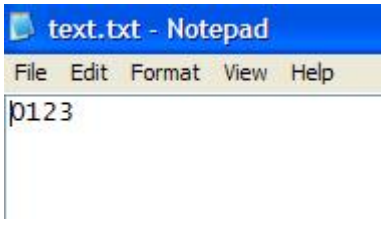

Figure 15: Output of scanned Bangla numeral as in shown Figure 10

5. Conclusion

In this paper we describe recognition of numerals in Bangla literature using template matching technique. No neural network is used for classifications. As correlation coefficient factor is the main feature in this paper so the training dataset must be large. A large dataset can improve the performance. The classification of handwritten data is shown in this paper. Bangla numeral containing 0 to 9 are tried to recognize with different shape and size. In this paper we proposed a technique to recognize Bangla numeral having different shape and size.

References

[1] C.Y. Suen, M. Berthod, and S. Mori, "Automatic Recognition of Hand printed Characters—The State of the Art," IEEE, 469-487, 1980.

[2] S.N. Srihari, E. Cohen, J.J. Hull, and L. Kuan, "A System to Locate and Recognize ZIP Codes in Handwritten Addresses," Int'l J. Research and Eng.-Postal Applications, 37-45, 1989.

[3] J. Tsukumo and H. Tanaka, "Classification of Handprinted Chinese Characters Using Nonlinear Normalization Methods," 168-171, 1988.

[4] A. Amin and H.B. Al-Sadoun, "Hand Printed Arabic Character Recognition System," 536-539, 1994.

[5] S.W. Lee and J.S. Park, "Nonlinear Shape Normalization Methods for the Recognition of Large-Set Handwritten Characters," Pattern Recognition, vol. 27, 895-902, 1994.

[6] H. Yamada, K. Yamamoto, and T. Saito, "A Non-Linear Normalization Method for Hand printed Kanji Character Recognition—Line Density Equalization," Pattern Recognition, vol. 23, 1023-1029, 1990.

[7] Farukh Al-omari,"Handwritten Indian Numeral Recognition System Using Template matches Approaches", IEEE, 0-7695-1165-1, 2001.

[8] C. Vadantha Lakhsmi, Ritu Jain,C Pathvardhan "Handwritten Devnagari Numeral Recognition with Higher Accuracy", IEEE, ICCIMA, 0-7695-3050-8/07,2007.

[9] R.V kulkarni, P.N. Vasambekar, "Overview of Segmentation Techniques for Handwritten Connected Digits", IEEE, 978-1-4244-8594-9,2010.

[10] B.B. Choudhury, U Pal. "An OCR system to read to Indian languages scripts: Bangla and Devanagari", 1011- 1015 (1997).

[11] K.Roy, C. Chaudhuri, U Pal, M.Kundu, " A Study on Effect of Varying Training set Sizes on Recognition Performance With Handwritten Bangla Numerals", IEEE, 0-7803-9503-4, 2005.

[12] M. Ziaratban, K. Faez, F. Faradji, "Language-Based Feature Extraction Using Template-Matching In Farsi/Arabic Handwritten Numeral Recognition", Ninth International Conference on Document Analysis and Recognition, 297 - 301, 2007.

[13] http://code.google.com/p/cmaterdb/.

Author Profile

Amitava Choudhury has received his B.Tech degree in Information Technology from West Bengal University of Technology. Presently he is perusing his M.Tech degree from School of Education Technology, Jadavpur University, Kolkata. He has started his career with Nimas College, Kolkata as IT- Lecturer since 2008. His research interests comprise of Handwritten Numeral Recognition, Character Recognition, and Pattern Recognition.

Joydeep Mukherjee obtained M.Tech degree from Jadavpur University. Currently he works as a Asst. Prof. in School of Education Technology, Jadavpur University, Kolkata. His research interests comprise of Digital image processing, Character Recognition, Pattern Recognition.