International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Performance Analysis on Bangla and Arabic Numeral Recognition

Gita Sinha[1], Md. Ashif Habibi[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Women's Institute of Technology, L.N.M.U., Darbhanga, Bihar, India

[2]Assistant Professor, Department of Computer Science and Engineering, Women's Institute of Technology, L.N.M.U., Darbhanga, Bihar, India

Abstract: ***Handwritten numeral recognition is a challenging work at this time. In this paper we deal with accuracy comparison among Arabic numerals and Bangla numeral recognition. This paper composed with four main phases pre-processing, feature extraction, classification and result comparison with both numerals. After pre-processing image are divided into different equal parts and then applied feature extraction techniques. Numerical feature extracted contain more details about image that provide improved recognition accuracy. These numerals recognition is carried out by SVM classifier.***

Keywords: Handwritten and printed numeral recognition, features extraction, pre-processing, support vector machine, image centroid zone and zone centroid zone, handwritten numeral recognition

1. Introduction

For the last three decades, the recognition of handwritten and printed letters has played a crucial and highly beneficial role in tedious and hard tasks such as postal sorting, bank check reading, order form processing, robotics. Several numerals recognition schemes have been proposed in the literature. They are all concerned with finding ways of how to differentiate between the different numerals. They are also interested in finding techniques for better numerals classification. Methods proposed in this context include dynamic programming, hidden Markov modeling, neural networks, support vector machines, k nearest neighbors, etc [1]. The problem of handwritten numeral recognition is addressed under the present work in respect to handwritten Arabic numerals. Arabic is spoken throughout the Arab World and the fifth most popular language in the world slightly before Portuguese and Bengali.

The automatic recognition of digits on scanned images has wide commercial importance. It has applications in OCR systems, automatic pin code recognition, cheque reading, collecting data from filled in forms. Though there is some commercially available software, mainly for printed character recognition of some languages, But the success yet to be extended for handwritten characters. Such technique is to facilitate smoother interaction between man and machine the technique of Handwritten.

Arabic numeral recognition can contribute tremendously to the development of a complete OCR system. Therefore OCR of handwritten numerals is still an active area of research.

The past work on OCR of handwritten alphabets and numerals has mostly found to concentrate on Roman script related to English and some other European languages, and scripts related to Asian languages like Chinese [2], Korean, Japanese. Among Indian scripts, Devnagari, Tamil, Oriya and Bangle have started to receive attention for OCR related research in recent years. Compared to these, Arabic is one of the major languages in the world. It is spoken in a large area including North Africa, most of the Arabian Peninsula and other parts of the Middle East. About 500 million peoples speak in this language. Rank wise it is the fifth most popular language in the world. Popularity wise Portugese and Bengali slightly trail behind Arabic. Arabic isthe official language of around 24 countries like Algeria, Baharain, Egypt etc. and also the national language of Mali, Senegal, Somali. More over Arabic is the liturgical language of Islam. It is sometimes difficult to translate Islamic concepts, and concepts specific to Arab culture, without using original Arabic terminology. The Arabic script has been adapted to such diverse languages as Persian (Farsi), Turkish, Spanish, Urdu, and Swahili. In spite of that, OCR of handwritten Arabic script including numerals has not so far received sufficient attention. Majority of the past work related to OCR of Arabic script was done with the printed characters [2].

Arabic words and characters within the words are written from right to left. But Arabic number is written from left to right.

2. Bangla and Arabic Numeral Dataset

The present Bangla database consists of 12000 samples of handwritten numerals. The experimental result of Bangla script is defined in next section. The few samples from this database shown in following table1.12000 Bangla numerals and6000 arabic numerals dataset have been received from [2] [3] [4]-[6]

Table 1: Numeral Dataset for Implementation

Bangla Numerals Sample	Arabic Numerals Sample
	

3. Pre-processing

Numerical recognition is nothing but the conversion of handwritten or typed numerical images into machine understandable form. For this the document is first scanned using a regular scanner. Before performing the feature extraction directly, first the scanned data/image is passed through some pre-processing steps (shown in the flow chart below), these steps are as follows:

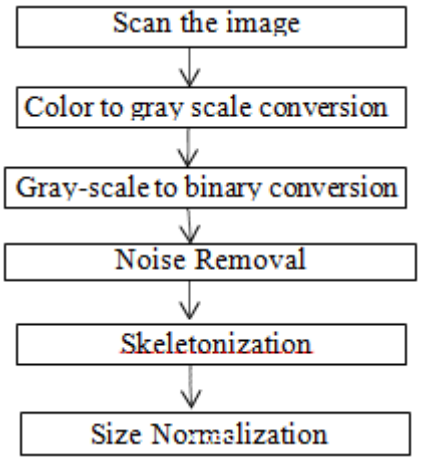

Figure 1: Pre-processing Steps

The procedure done before pro-processing by correcting image from different errors is Pre-processing. This has to be done before image enhancement. In HNR, typical Pre-processing operations include Binarization, Contour smoothing, Noise reduction, skew detection and Skeletonization of a digital image so that subsequent algorithms along the road to final classification can be made simple and more accurate.

4. Segmentation

Segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as superpixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. There are three types of segmentation Line segmentation, Word segmentation and Character segmentation.

5. Feature extraction

When the input data of an algorithm is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called *feature extraction*. If the features extraction methods are carefully chosen it is expected that Recognition is more accurate.

Post-processing:- In post-processing we want to replace the input sequence of characters with another sequence of characters that is graphically similar and form the likeable sentence of the given language. We also reduce the number of errors in text. Semantic relations between words can be used to aid selection from alternative candidate words output from an Optical Character Recognition (OCR) system in order to improve the overall recognition rate. One method of automatically identifying the semantic relations between words is by using an existing knowledge source [1].

6. Classification

Classification aims at creating a map from the input data to a corresponding known output in the training phase. The constructed map, called classifier, is then used to predict new input instances. Many classification techniques have been developed such as linear discriminate functions, *K*-nearest neighbors, Bayesian decision theory, neural networks, committee machine, Support Vector Machine (SVM) and so on. However, all those techniques train and classify data considering only the physical features. Techniques used for classification is defined in next section.

7. Feature Extraction

Following listed features have been used for current experiment. Two types of features namely image centroid zone, and zone centroid zone. 200 feature vectors have been formed using combinations of both basic features. These methods provide the ease of implementation and good quality recognition. Step-by-step algorithm has been defined in the next section. In the next section, these algorithms have been defined. The following paragraph explains the details about feature extraction method.

7.1 Image Centroid Zone

The centroid of image (numeral/character) has been computed. The given image has been further divided into 100×100 equal zones where size of each zone is (10×10). Then, the average distance from image centroid to each pixel present in the zones/block has been computed. 100 feature vectors of each image are thus obtained. Zones which are empty are assumed to be zero. This procedure is repeated for all zones present in image (numeral/character).

7.1.1 Image Centroid Zone (ICZ)

.

Algorithm:

1) Calculate centroid of input image.
2) Divide the image into equal zones.
3) After zone formation, calculate the distance between centroid and each numeral pixel in the zone.
4) Now calculate the average distance corresponding to that zone.
5) Repeat procedure 3) and 4) for each zone.
6) In this way we obtain the number of features equal to the number of zones created in second step.

Figure 2 shows example of Arabic numeral image of size 32×32. First, centroid of image is computed. Then, image is divided into 16 equal zones each of size 8×8. Later, average distance from image centroid to each pixel present in the image is computed.

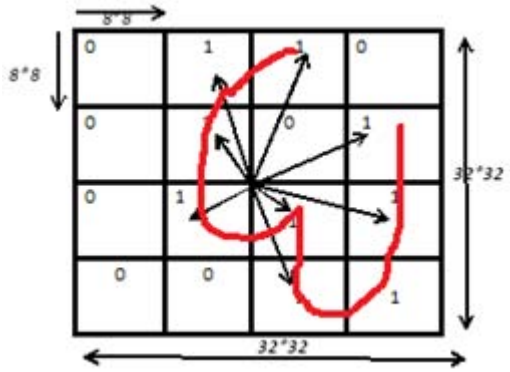

Figure 2: (ICZ) Image 32×32 and block 8×8.ofhandwritten Arabic numeral image "four"

Zone Centroid Zone

In ZCZ, image is divided into 100×100 equal zones and centroid of each zone is calculated. Followed by computation of average distance of zone centroid to each pixel present in zone. Zones which are empty are assumed to be zero. This procedure is repeated for all pixels present in each zone.

Efficient zone based feature extraction algorithm has been used for handwritten numeral recognition. Image centroid zone (ICZ) based distance metric feature extraction system, while Zone Centroid Zone (ZCZ) based Distance metric feature extraction system have been used to extract features of the images. Further, another Algorithm that is the combination of both (ICZ+ZCZ) feature extraction systems. The following algorithms illustrate the working procedure of feature extraction methods as.

7.1.2 Zone Centroid Zone (ZCZ) Algorithm:

1) Divide the image into equal zones.
2) Calculate centroid for each zone.
3) After zone formation, calculate the distance between the zone centroid and each numeral pixel in the zone.
4) Now calculate the average distance corresponding to that zone.
5) Repeat procedure 3) and 4) for each zone.
6) In this way we obtain the number of features equal to the number of zones created in first step.

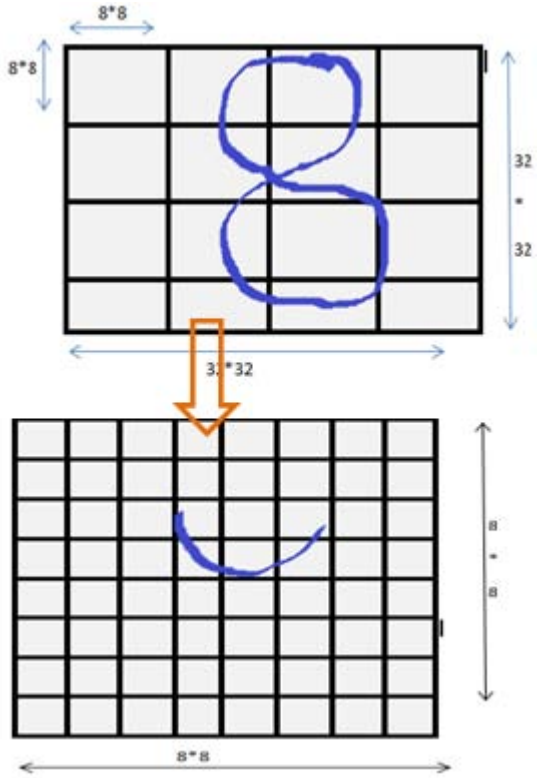

Figure 3: (ZCZ) Image 32×32 and block 8×8.ofhandwritten Bangla numeral image "four"

Figure 3 shows example of Bangla Numeral image for size 32×32. In this figure image has been divided into 64 equal zones, each of size 8×8. Centroid of each zone in image has been computed. Then, average distance from image centroid to each pixel present in the zone is calculated. FV1 is achieved by ICZ, FV2 is achieved by ZCZ and FV3 is combination of ICZ and ZCZ.

8. Result

8.1 Comparative Analysis

Figure-4 shows the recognition rate of Arabic numerals and Bangla numerals, using different feature extraction techniques. There are three feature vector that is FV1,FV2 and FV3.

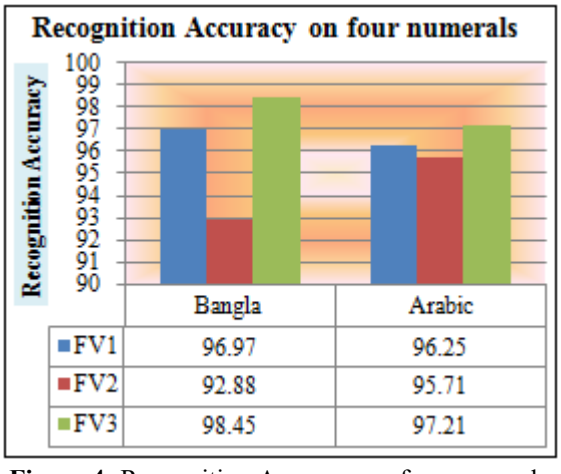

Figure 4: Recognition Accuracy on four numerals

Table 2 shows that the results obtained on Bangla Numerals are good, with the SVM Classifier where the recognition rate achieved 98.45%. Dividing the image into different equal parts has provided more details for the Images which have become more discriminate and provide Feature vector. A well designed SVM classifier is used to achieving good results.

Table 2: Shows Overall accuracy

Language	Dataset	Feature Extraction Techniques	Classifier	Accuracy
Arabic Numerals	5000	Zonal based	SVM	97.21%
Bangla Numerals	12000			98.45%

Table 3 shows recognition accuracy using SVM classifier with different value ofγ and C on Arabic and bangla numeral image. The highest accuracy using bangla Numerals is 98.45% And Highest accuracy of Arabic numerals is 97.21%

Table 3: Recognition accuracy using SVM

Sr. No.		Arabic			Bangla		
	Feature vector	γ	C	Recognition accuracy	γ	C	Recognition accuracy
1.	Fv1	.001	1	93.81%	0.001	1	96.97%
2.	Fv2	.001	1	89.26%	0.002	2	92.88%
3.	Fv3	.001	1	93.76%	0.01	4	98.39%
4.	Fv1	.08	2	94.65%	0.02	8	97.24%
5.	Fv2	.08	2	91.15%	0.008	16	95.56%
6.	Fv3	.08	2	94.79%	0.001	32	98.45%
7.	Fv1	4	64	96.08%	0.0016	64	97.22%
8.	Fv2	4	64	94.48%	0.016	128	97.67%
9.	Fv3	2	32	96.63%	0.1	256	98.45%
10.	Fv3	8	128	97.05%	0.1	512	97.22%
11.	Fv3	8	256	**97.21%**	0.016	512	**98.45%**

9. Summary

In this method the feature set which is extracted using ICZ & ZCZ feature extraction technique is applied to SVM for classification. The data base used for both the techniques is different. Also the images used for testing both the system are different, than we compare the result obtained from both the methods.

References

[1] Ouafae EL Melhaoui et al "Arabic Numerals Recognition based on an improved Version of the Loci Characteristic" International Journal of Computer Applications (0975 – 8887) Volume 24– No.1, June 2011

[2] N. Das, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application," Applied Soft Computing, vol. 12, pp. 1592-1606, 2012.

[3] N. Das, J. M. Reddy, R. Sarkar, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "A statistical–topological feature combination for recognition of handwritten numerals," Applied Soft Computing, vol. 12, pp. 2486-2495, 2012.

[4] N. Das, K. Acharya, R. Sarkar, S. Basu, M. Kundu, and M. Nasipuri, "A Novel GA-SVM Based Multistage Approach for Recognition of Handwritten Bangla Compound Characters," *Proceedings of the International Conference on Information Systems Design and Intelligent Applications vol. 132, pp. 145-152 2012.*

[5] N. Das, S. Basu, R. Sarkar, M. Kundu, M. Nasipuri, and D. K. Basu, "Handwritten Bangla Compound character recognition: Potential challenges and probable solution," *in 4th Indian International Conference on Artificial Intelligence, Bangalore, pp. 1901-1913 2009.*

[6] N. Das, S. Basu, R. Sarkar, M. Kundu, M. Nasipuri, and D. K. Basu, "An Improved Feature Descriptor for Recognition of Handwritten Bangla Alphabet," *in International conference on Signal and Image Processing, Mysore, India, pp. 451-454.2009.*

[7] N. Das, K. Acharya, R. Sarkar, S. Basu, M. Kundu, and M. Nasipuri, "A Benchmark Data Base of Isolated Bangla Handwritten Compound Characters," IJDAR(Revised version communicated).

[8] S.V. Rajashekararadhya, "efficient zone based feature extraction Algorithm for Hand written numeral Recognition of four popular south Indian Scripts," *Journal of theoretical and Applied Information Technology, 2008.*

[9] Nibaran Das et al. "Handwritten Arabic Numeral Recognition using a Multi Layer Perceptron" Proc. National Conference on Recent Trends in Information Systems (2006) 200-203.

[10] S. L. Mhetre et al. "Comparative study of two methods for Handwritten Devanagari numeral recognition",Issue6 *(Nov.-Dec. 2013), PP 49-53www.iosrjournals.org.*