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

Comparative Result on Bangla, Arabic and Gurumukhi Numerals Recognition

Gita sinha[1], Ashif Habibi[2], Dr. Pankaj Kumar Chaudhary[3]

[1]Assistant Professor, Department of Computer Science & Technology, Women's Institute of Technology, L.N.M.U.Darbhanga 846004

[2]Assistant Professor, Department of Computer Science & Technology, Women's Institute of Technology, L.N.M.U.Darbhanga 846004

[3]Assistant professor, Department of Mathematics, Women's Institute of Technology, L.N.M.U.Darbhanga 846004

Abstract: ***In this paper we present the recognition accuracy of Bangla, Arabic and Gurumukhi numerals and compare their result using different parameters like gamma and alpha. We describe two feature extraction techniques and two classifier for recognition of offline numerals. We have got more accurate result on these experiments using SVM classifier. Images are divided in 8*8, 16*16, 32*32 according image size feature vector is calculated. Highest recognition on Bangla numerals produces 98.45%, Arabic 97.21% and 99.73% on Gurumukhi numerals.***

Keywords: SVM-support vector machine, AN-Arabic numerals, BN-Bangla Numerals and Gurumukhi Numerals, ICZ-image centroid zone, ZCZ- zone centroid zone

1. Introduction

These are several steps of OCR namely pre-processing, segmentation, feature extraction and classification which describe latter on this paper. There are many languages and script in India but not much research has been done for these handwritten numerals recognition. Recognition of handwritten numerals has been a popular research area for many years because of its various application potentials. these are the potential application of OCR postal automation, shopping mall bank cheque reading, automatic data entry etc. several pieces of work have been done towards handwritten recognition of Roman, Japanese, Chinese and Arabic scripts and various approaches have been proposed by the researchers towards handwritten numerals recognition . Although In this paper, we propose a system based on Support Vector Machines (SVM) and for the recognition of off-line handwritten AN, BN, GN.
printed Bangla Character started in early 1990s [2] and till mid-1990s. no significant work has been reported . Recently, several pieces of work on Bangla have been published [3]. Among the earlier pieces of work, some of the efforts on Bangla character recognition are due to Ray and Chatterjee[4].

2. Related Work

In 2010 chin lei he at all [5] *an automated writing style detection process is effectively implemented in the pair-wise verification of samples in these two classes. As a result, the recognition results have improved significantly with a reduction by 25% of previous errors. With rejection, when the LDA (Linear Discriminant Analysis) measurement rejection threshold is adjusted to maintain the same error rate, the recognition rate* ***Arabic*** *numerals increases from 96.87% to 97.81%.*

In 2008 Mohammed Moshiul Hoque [6] have proposed a method **use unique fuzzy rule base for Bangla numerals considering various writing style and got more than 80% recognition accuracy. Every numeral is segmented and several features are extracted for each segment.**

In 2000 ZHAO Bin at all [6] proposed different classes of SVM and got different rate of recognition accuracy on numerals. Recognition rate is 91.98, 91.47 and 90.98 accordingly class1, class2, class 3. They achieves good recognition rate in Methodl with the test speed is fast. Method2's speed is much slow. Method3 is the fastest while sacrifices recognition rate

In 2012 Rekha Anoop at all [7] present paper on Gurmukhi numeral. They have use Projection Histograms for feature extraction and SVM for classifier , which produce the highest accuracy 99.2% during Gurmukhi numeral recognition.

3. Phases of OCR

3.1 Digitization

Document is scanned by the process digitization and an original electronic documents are representation by bitmap image, which is produced by this process. digital image,is produced by digitization produces the which is fed to the pre-processing phase as an input.

3.2 Preprocessing

Skew detection/correction, skeletonization, and noise reduction/removal are perfomed in the phase of Preprocessing. Tilt in the bit mapped image of the scanned paper for OCR is refers to Skewness. when decreasing the line width of text from many pixels to single pixel skeletonization is used. Noise removal is used to remove unwanted bit pattern which does not play any significant role in document.

3.3 Segmentation

Segmentation is process of partitioning the digital image into small segments. There are three types of segmentation words, line and characters.

3.4. Feature Extraction

Feature extraction need to reducing the amount of resources and effort required to describe a large set of data. When the input data for an algorithm is very large to be processed and it may be redundant or the repetitiveness of images presented as pixels, then it can be transformed into a reduced set of features also called a feature vector.

3.5. Classification

Decision making is performed in the process of Classification of an OCR engine, which accept output produced by the features extraction phase as an input for making the class memberships in pattern recognition system.

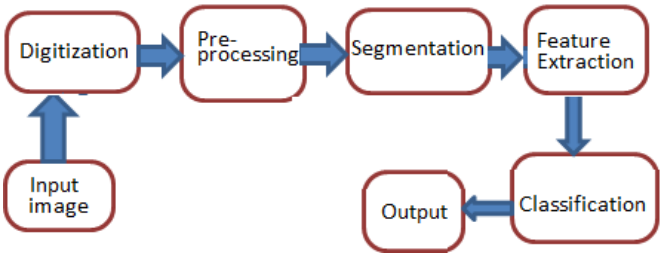

Figure 1: Phases of OCR

3.6 Challenges of Handwritten Numeral Recognition

Numeral Recognition has following challenges [2]:

1) Variability of writing style, both between different writers and between separate examples from the same writer overtime.
2) Similarity of some characters.
3) Low quality of text images
4) Unavoidable presence of background noise and various kinds of distortions.

4. Feature Extraction Techniques

In current experiment two types of features namely image centroid zone and zone centroid zone has been used. 200 feature vectors have been formed using combinations of both basic features. These two methods produce the ease of implementation and good quality recognition. Algorithm has been defined in the next section in details. Feature extraction method is explained below.

4.1 Image Centroid Zone

At first we have computed centroid of an image (numeral/character) Then given image has been further divided into 100×100 equal size of 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. Each image are thus produced 100 feature vectors .Empty zones are assumed to be zero. This process is repeated for all zones present in image (numeral/ character). Figure 2 shows example of Arabic numeral image of size 32×32. First, centroid of image is computed. Then, image is partitioned into 16 equal zones each of size 8×8. Later, average distance from image centroid to each pixel present in the image is computed.

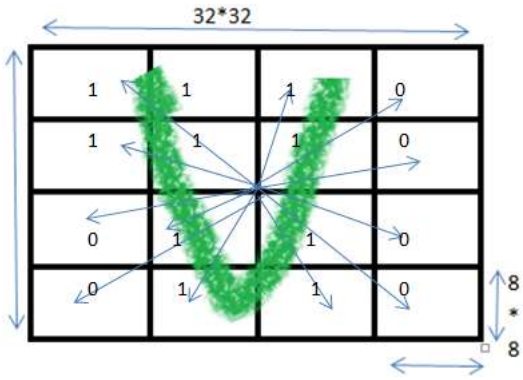

Figure 2: Arabic numeral image seven

4.2 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. empty zones 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 of four popular south Indian scripts as defined in [8]. Here, same method has been applied on few north Indian scripts. Algorithm 1 provides steps for Image centroid zone (ICZ) based distance metric feature extraction system, while Algorithm 2 provides steps for Zone Centroid Zone (ZCZ) based Distance metric feature extraction system. Further, Algorithm 3 provides the combination of both (ICZ+ZCZ) feature extraction systems. The following algorithms illustrate the working procedure of feature extraction methods as depicted in figure 3.

Figure 3.3 shows example of numeral image for size 32×32. In this figure image has been divided into 16 equal zones, each of size 8×8. Centroid of each zone in image has been computed. After that, average distance from image centroid to each pixel present in the zone is computed

Algorithm for Image Centroid Zone (ICZ) feature extraction technique.

Give Pre-processed numeral Image for input and produce Extracted the Features for Classification and Recognition as output.
Step 1: Centroid of input image has been computed.
Step 2: Input image is Divided into 100 ×100 equal zones.
Step 3: calculate the distance from the image Centroid to each pixel present
in the zone.
Step 4: For the entire pixel present in the zone/boxes/grid repeats step 3.
Step 5: computed average distance between these Points.

Step 6: This procedure is sequentially repeated for the entire zone present in the image.
Step 7: Achieved 100 such feature for Classification and efficient recognition
Ends.

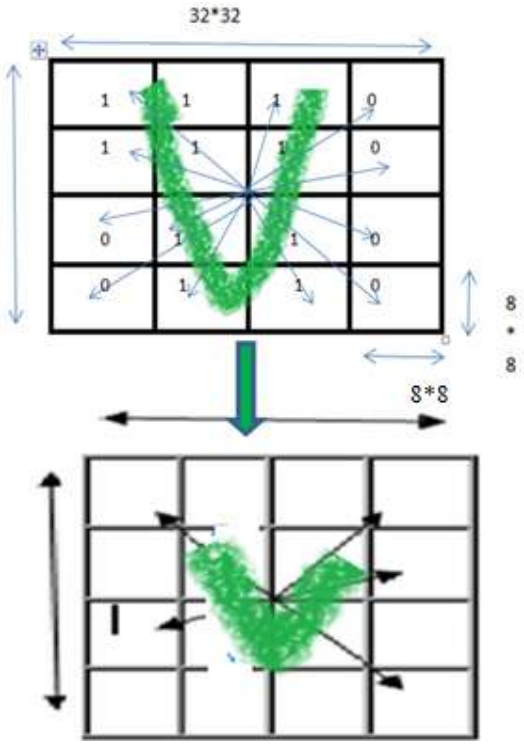

Figure 3: Arabic numeral image "seven"

Algorithm 2: Zone Centroid and Zone based method for feature extraction system.
Method Begins
Step 1: divide input image in to **n** equal zones.
Step 2: Compute centroid of each zones of an image.
Step 3: Computation of distance between the zone centroid to each pixel present in the zones.
Step 4: Repetition of step 3 for the entire pixel present in the zone/box/grid.
Step 5: Compute the average distance between these points present in image.
Step 6: This procedure are sequentially repeated for the entire zone.
Step 7: Achieved, 100 such features for classification and recognition.
Ends.

4.3 Hybrid Algorithm 3

Hybrid feature extraction method is a combination of both of the algorithm **(**ICZ + ZCZ) defined above. This method produced 200 such features from each of the image.

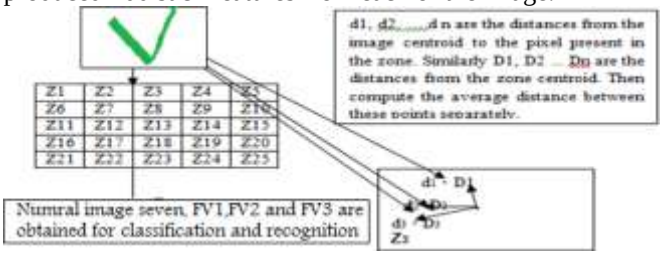

Figure 4: Hybrid feature extraction method for Arabic numeral image "seven"

4.4 Database for implementation

Table 1: Database for implementation

Database	Language	Availability
1500	Gurmukhi Numerals	Manually
12000	Bangla Numerals	[9][10][11]-[14]
6000	Arabic Numerals	[9][10][11]-[14]

For category (GN), data has been manually created by personal efforts. There are total of 1500 Gurmukhi number, while another (BN, AN) have been received from[9][10][11]-[14]. There are total of 1500 Gurmukhi number, 12000 Bangla dataset available at[9][10][11]-[14], 6000 Arabic numerals available at [9][10][11]-[14]. In order to test the efficiency of our recognition system approach, following database has been defined in table 1.

5. Recognition Accuracy on Different Numerals Image

Handwritten numeral recognition has been tested on local database and also standard database. Experiments have been performed and the analyzed. Results have been shown in figure 5 Highest accuracy have been defined in all the experimental results. In addition, parameter, feature vector and image size have not been mentioned. All these parameters has been mentioned in table-2.

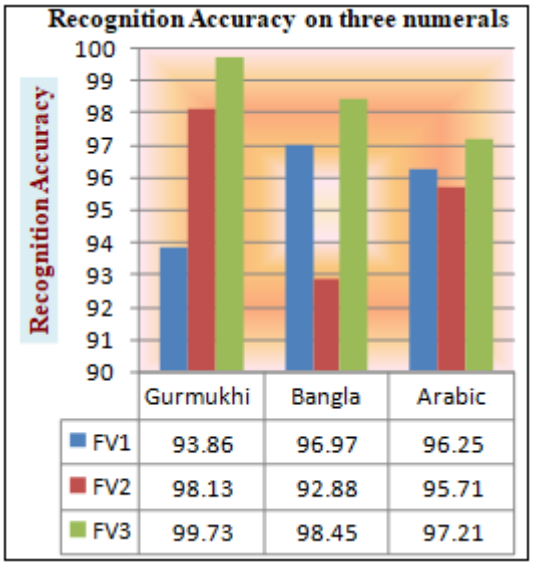

Figure 5: Optimized Recognition accuracy on Gurmukhi, Arabic and Bangla Numerals using SVM

Arabic Numeral Recognition			
11	8	.8	95.3%
12	64	4	96.76%
13	4	.01	95.38%
14	**128**	**8**	**97.21%**
15	**512**	**8**	**97.21%**

Table 2: Comparative Result on Gurmukhi , Bangla and Arabic Numeral Recognition with SVM Classifier

Gurmukhi numerals			
S. No.	Value of gamma(γ) for RBF	Value of C	Recognition Accuracy
1	4	4	99.60%
2	16	0.08	90.40%
3	32	8	99.66
4	4	1	**99.73%**
5	16	0.001	35.46%
Bangla numerals			
6	0.02	8	97.24%
7	0.008	16	95.56%
8	**0.001**	**32**	**98.45%**
9	0.0016	64	97.22%
10	0.016	128	97.67%

Five-fold cross validation has been used to validate result obtained. Zone based feature extraction techniques consisting 100 feature vectors and SVM classifier with RBF kernel has been used to achieve 99.73% accuracy which is the highest obtained result in Gurmukhi numerals , 98.45% is the highest obtained result in Bangla numerals and 97.21% is the highest obtained result in Arabic numerals . Table 2 depicts various experiments on different numeral image and parameters. The first method consisting of 100 feature vectors, second method also consists of 100 feature vectors and third method is combination of both of the method, so it consists of 200 feature vectors. Feature vector depends upon size of image and block which is not mentioned in table.

6. Conclusion

In this paper we used a handwritten Gurmukhi, Bangla and Arabic numerals recognition system that contains the different feature vector in three feature extraction technique. in the features extraction phase, ICZ, ZCZ and the Hybrid method(ICZ+ZCZ) has been used. Support vector machine (SVM) has been used in classification phases. Our goal is to compare the performances of these numbers. The experimental results that we have obtained prove that hybrid feature extraction techniques are more accurate than ICZ and ZCZ in this recognition.

References

[1] *Umapada Pal et al. "*Accuracy Improvement of Devnagari Character Recognition Combining SVM and MQDF*".*
[2] B Chatterjee and A.K Roy, *On the classification of hand- printed Bengali numeral characters, Proceedings of the Symposium on Microwaves and Communication*, IIT Kharagpur, India (1983). [SD-008]
[3] B.B Chaudhuri and U Pal, A. complete printed Bangla OCR system. *Pattern Recognition* **31** (1998), pp. 531–549. [SD- 008]
[4] K Ray and B Chatterjee, Design of a nearest neighbor classifier system for Bengali characterrecognition *J. Inst. Electron. Telecom.Eng.*, **30** (1984), pp. 226–229. [SD-008]
[5] Chun lei he at all , " Automatic discrimination between confusing class with writing style verification in Arabic numerals recognition " 2010 International conference on pattern recognition.
[6] ZHAO Bin at all "**Support Vector Machine and its Application in Handwritten Numeral Recognition"** 0-7695-0750-6/00 , 2000 IEEE
[7] Rekha Anoop at all "**Offline Handwritten Gurmukhi Character and Numeral Recognition using Different Feature Sets and Classifiers**" IJERA vol. 2 Issue 3,May- June 2012.
[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] 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.
[10] 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.
[11] 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.
[12] 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 I
[13] 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.*
[14] 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).