International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-2, Issue-5)

Research Article

May 2013

Arabic numeral Recognition Using SVM Classifier

Gita Sinha
Assistant Professor CSE deptt.
National Institute of Technology Patna- 800005(India)

Dr. Jitendra kumar
Department of pharmacy
G D M H M C & H Patna, India.

Abstract—

In this paper we present a system called ORAN (Offline Recognition of Arabic Numerals). This system is based on a method called Zone based feature extraction techniques. In this paper we proposed a method for Offline Handwritten Arabic Numerals Recognition with the use of Classifier and Feature Extraction Techniques. Pre-processed, segmented and feature Extraction techniques are applied on every image. This paper present, three feature extraction techniques which is namely: image Centroid zone (ICZ), zone centroid zone (ZCZ) and hybrid feature extraction techniques Hybrid Feature Extraction Techniques is combination of ICZ+ZCZ. The support vector machine (SVM) which is based on statistical learning theory, with good generalization ability. Which is used as the classifier for Recognition of the numerals corresponding to these three types of features. According to experimental results, classification rates of 96.25%, 97.21% and 97.71% were obtained for Numerals respectively on the test sets gathered from various people with different educational background and different ages. The recognition rate of this method is 97.21% on handwritten bangle numeral database.

Keywords - Image centroid zone (ICZ), Zone centroid zone (ZCZ) and support vector machine (SVM).

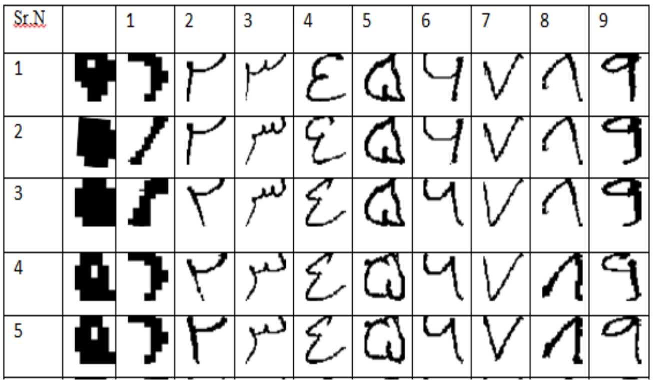
I. INTRODUCTION

Handwritten digits recognition has application like office automation, check verification, and postal address reading and sorting. While recognition of handwritten Latin digits has been extensively investigated using various techniques [1], too little work has been done on handwritten Arabic digits. Various symbol sets are used to represent numbers in the Hindu-Arabic numeral system, all of which evolved from the Brahmi numerals. West Arabic numerals developed in al-Andalus and the Maghreb. (There are two Typographic styles for rendering European numerals, known as lining figures and text figures). The Arabic–Indic or Eastern Arabic numerals, used with the Arabic script, developed primarily in what is now Iraq. A variant of the Eastern Arabic numerals used in the Persian and Urdu languages is shown as East Arabic-Indic. There is substantial variation in usage of glyphs for the Eastern Arabic-Indic digits, especially for the digits four, five, six, and seven. The Devanagari numerals used with Devanagari and related variants are grouped as Indian numerals. Samples of printed and handwritten Arabic digits are shown in figure 5.a and table 1

Western Arabic	0	1	2	3	4	5	6	7	8	9
Eastern Arabic	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Perso-Arabic variant	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹

Figure5.a Sample of Arabic numerals from 0-9

TABLE 1.1 HANDWRITTEN ARABIC NUMERAL


Sr.N
1
2
3
4
5
6
7
8
9
1
2
3
4
5

ORAN (Offline Recognition of Arabic Numerals) consist with various phases pre-processing, segmentation, feature extraction and classification shown in figer-1.

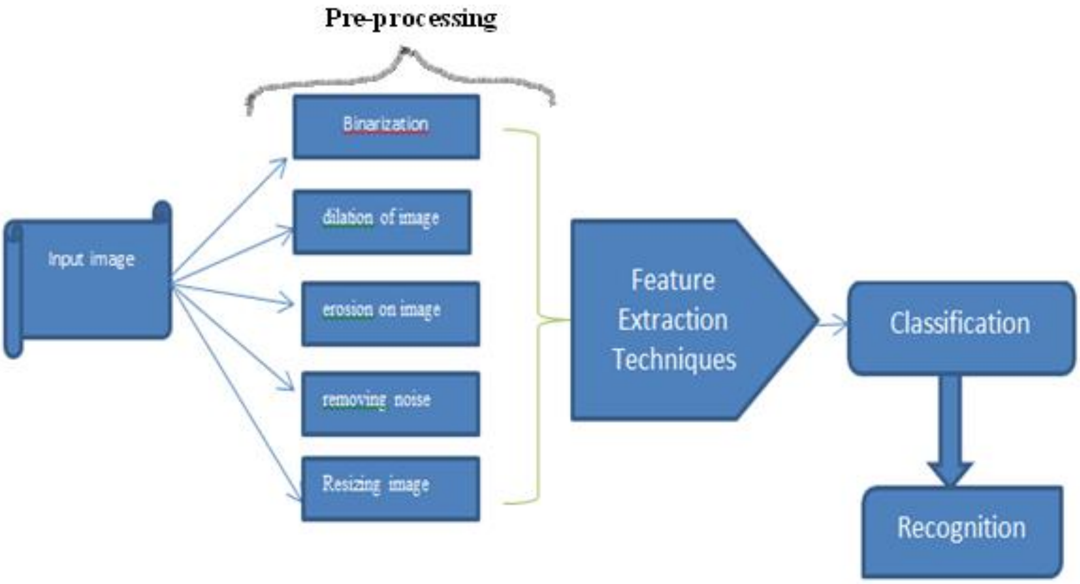


Fig. 1 phases of OCR

Although Several investigators have previously made, Chun Lei He Ching , Y. Suen [2] In this paper, errors are categorized based on different costs in misclassification. When a rejection measurement was applied, and the rejection threshold was adjusted to maintain the same error rate, both the recognition rate and reliability increased from 96.98% to 97.89% and from 99.08% to 99.28%, respectively. They have used SVM as a classifier and Radial Basis Function (RBF) as a kernel function. Faruq Al-Omari**,** Ph.D.[3] They have used the process involved extracting a feature vector to represent the handwritten sketch based on the “object” centroid and boundary points. A template vector was derived for each digit by taking the average feature vector **of** 30 handwritten sketches made by 30 different students. The test sketch is compared against all nine templates and a distance measure is performed to make the recognition. An overall hit ratio **of** 87.22% was achieved in the preliminary results. Saeed Mozaffari, Karim faez, Hamidreza Rashidy Kanan [4] In this paper proposed a new method for isolated handwritten Farsi/Arabic characters and numerals recognition using fractal codes and wavelet transform The support vector machine (SVM) which is based on statistical learning theory. According to experimental results, classification rates of 92.71% and 92% were obtained Abdelmalek Zidouri [5] they have used MCR (Minimum Covering Run) expression for document images. to describe the strokes of characters according to some extracted features. These are obtained alter a zoning scheme, where the baseline is detected and tile line of text divided into four zones. Reference prototypes for the system are built according to a structural description of characters in some model documents. By this method overcome the problem of segmentation. A recognition rate of more than 97% is achieved.

II. Feature Extraction

Feature Extraction Techniques is the most important step in any recognition system. Any recognition system’s performance depends on the quality of features as it depends on the classifier used to make the final classification decision [5]. For a better recognition system, the feature set extracted should capture properties of input image. Structural features tend to describe the visual topological and geometrical properties of the digit which make them similar to how humans distinguish the different classes. The feature extraction used in the proposed system is based on extracting different zone based features from the input image. First, the image is pre-processed to normalize the size of the images; this is done via resizing the bounding box of the digit to height h1 and width w 2. Second, the image is traversed to label each white pixel in the image with configuration as explained in the next section. 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.

A. 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). Figure 2 shows example of character image of size 32×32. First, centroid of image is computed. Then, image is divided into 16 equal zones each of size 8×8. ater, average distance from image Centroid to each pixel present in the image is computed

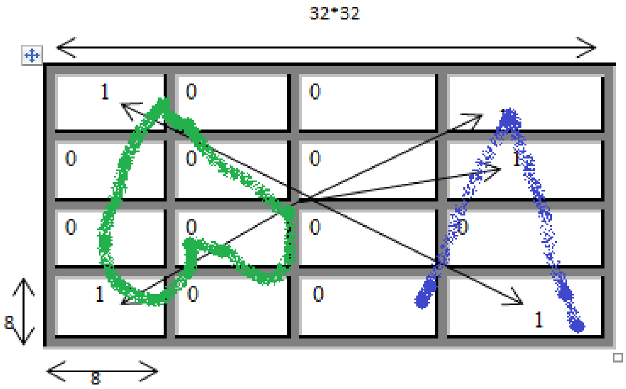


Fig-2 (ICZ) Image 32×32 and block 8×8.of handwritten Arabic numerals 5 (green) 8 (blue) .

B. 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 of four popular south Indian scripts as defined in [35]. Here, same method has been applied on few north Indian scripts. Algorithm 1 provides Image centroid zone (ICZ) based distance metric feature extraction system, while Algorithm 2 provides 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 shows example of character 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. Then, average distance from image centroid to each pixel present in the zone is calculated.

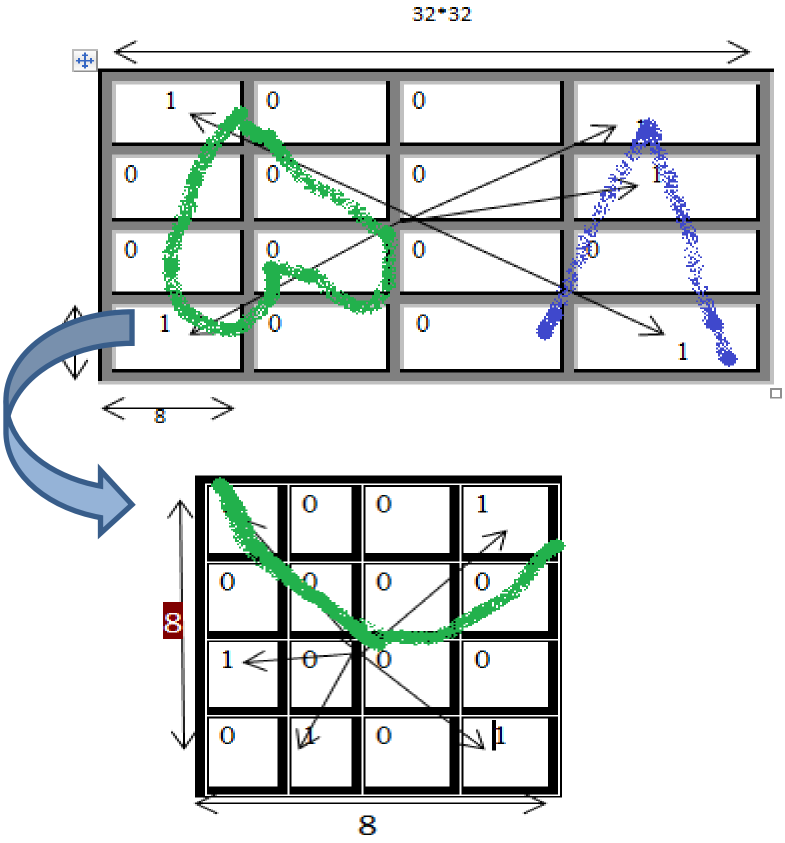
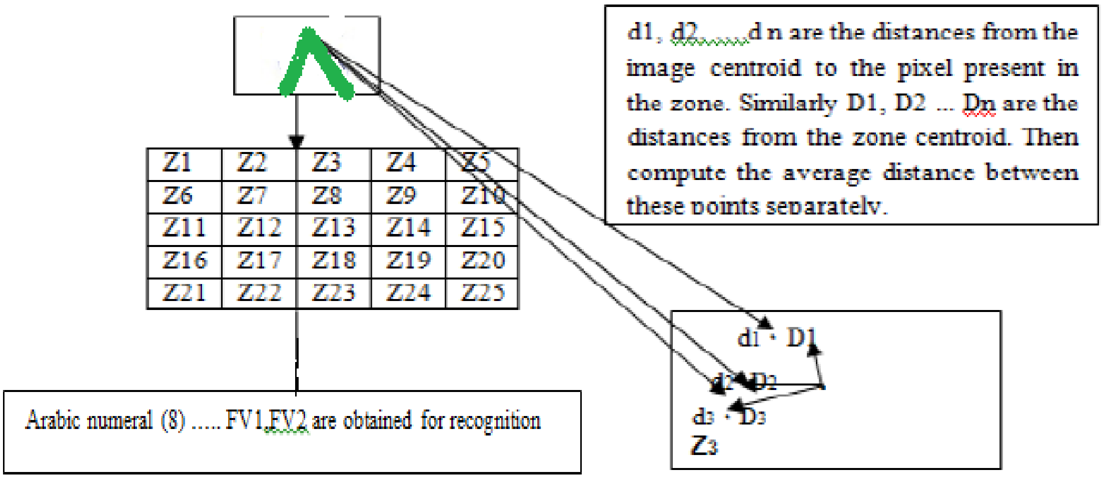


Fig 3 (ZCZ) Image 32×32 and block 8×8.of handwritten Arabic numeral.

Algorithm 1: Image Centroid Zone (ICZ) feature extraction method.
Input : Pre-processed Image (character/numeral)
Output : Extract the Features for Classification and Recognition

Method Begins
Step 1: Calculate centroid of input image.
Step 2: Division of input image in to 100 ×100 equal zones.
Step 3: Computation of the distance from the image Centroid to each pixel present in the zone.
Step 4: Repeats step 3 for the entire pixel present in the zone/boxes/grid.
Step 5: Average distance computed between these Points.
Step 6: Repeat this procedure sequentially for the entire zone present in the image.
Step 7: Obtaining 100 such feature for Classification and recognition process.
Algorithm 2: Zone Centroid and Zone (ZCZ) based feature extraction system.
Method Begins
Step 1: Division of input image in to **n** equal zones.
Step 2: Compute centroid of each zones.
Step 3: Compute the distance between the zone centroid to each pixel present in the zones.
Step 4: Repeat step 3 for the entire pixel present in the zone/box/grid.
Step 5: Computation of average distance between these points present in image.
Step 6: This procedure are sequentially repeat for the entire zone.
Step 7: Obtaining, 100 such features for classification and recognition.
Hybrid Algorithm 3: Hybrid feature extraction method is a combination of both of the algorithm **(**ICZ+ZCZ) defined above. This method provides 200 such features from each of the image.



d1, d2......d n are the distances from the image centroid to the pixel present in the zone. Similarly D1, D2 ... Dn are the distances from the zone centroid. Then compute the average distance between these points separatelv.

Z1	Z2	Z3	Z4	Z5
Z6	Z7	Z8	Z9	Z10
Z11	Z12	Z13	Z14	Z15
Z16	Z17	Z18	Z19	Z20
Z21	Z22	Z23	Z24	Z25

Arabic numeral (8) ..... FV1,FV2 are obtained for recognition

Fig 4 All procedure to extract feature.

In figure 4 image size is taken as 50×50. When this image is divided into equal zone of size 10×10 pixels, then total number of zones formed is 25. Similarly image size is taken as 100×100, and image is divided into equal zone of size 10×10 pixels then total number of zones will be 100. Likewise, when image size is 32×32, and lock size is 8×8, then 16 zones will be created.

III.Classification And Recognition

Support Vector Machines (SVMs).

Support Vector Machines (SVMs) are modern learning machines introduced by Vapnik and developed by Vapnik and other researchers [7]. For a two-class classification problem, assume that set of input vectors are given: xi ∈ Rd (i = 1,2, …N) with corresponding labels, yi ∈ {+1, -1}, (i = 1,2, …N). Here, +1 and -1 indicate the two classes and N is the number of samples. SVM maps the input vectors xi ∈ Rd into a high dimensional feature space φ (x) ∈ H and constructs an optimal separating hyper plane which maximizes the distance between the hyper plane and the nearest data points of each class in the space H. The mapping φ (.) is performed by a kernel function K (xi, yj) which defines an inner product in the space H. The decision function implemented by SVM can be written as:

$$f(x) = sgn \sum_{i=1}^{N} yi\, \alpha i.\, K(x, xi) + b\}$$

Where N is the number of the training samples, b the offset of the optimal hyper plane from the origin, and the coefficients αi are obtained by solving the convex quadratic programming problem:

The linear SVM can be extended to a non-linear classifier by using kernel functions like polynomial and Gaussian kernels [8]. We have use RBF kernel for the purpose of recognition during the phase of classification.

IV. Experimental Results And Comparative Analysis

For experimental results, we considered 6,000 samples per class for training from [9] [10][11] standard Arabic numeral dataset further. Experimental results are described in this subsections.

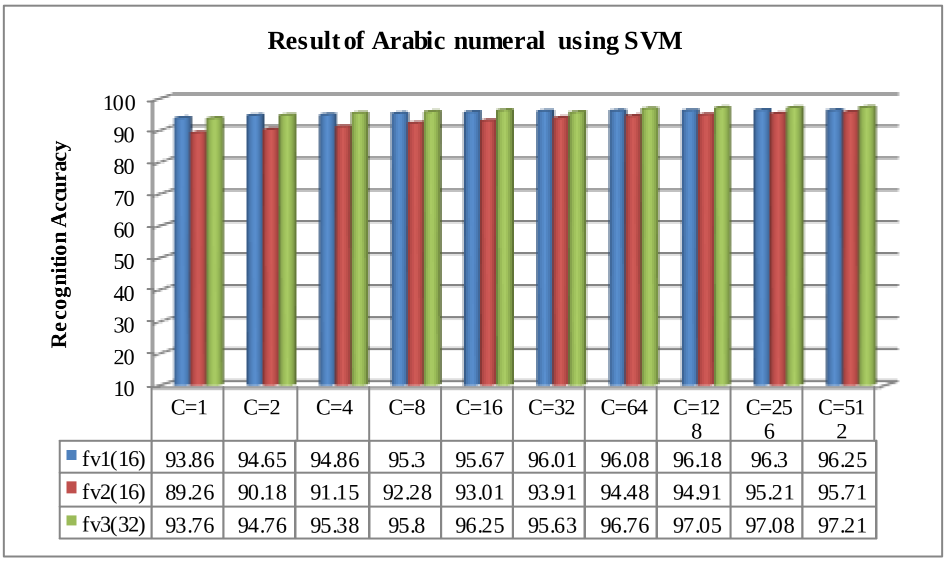
A. Experimental results

In table 3 we have applied three feature extraction technique: image centroid zone (ICZ), zone centroid zone (ZCZ) and hybrid feature extraction techniques which is combination of (ICZ+ZCZ) which makes fv1 have 16 feature vector , fv2 have 16 feature vectors and fv3 have 32 feature vector on Arabic numeral. Size of feature vector depends upon size of image we have taken. We have used SVM classifier with radial basic kernel (RBF) for the purpose of recognition. Recognition accuracy depends upon different value of parameter C and γ.we have used different value of gamma up to 8 and C=2^0, 2^1 …..2^9. Accuracy slightly increases corresponding to value of C.

Table 3.Arabic numeral accuracy with SVM on different value of parameter

Sr. no.	**SVM parameter**		**Feature vector's name and size**		
1	**C**	γ	**Fv1(ICZ)(16)**	**Fv2(ZCZ)(16)**	**Fv3(ICZ+ZCZ)(32)**
2	1	0.001	93.81%	89.26%	93.76%
3	2	0.08	94.65%	91.15%	94.79%
4	4	0.01	94.86%	91.20%	95.38%
5	8	0.8	95.3%	92.28%	95.8%
6	16	0.16	95.65%	93.01%	96.25%
7	32	2	96.01%	93.91%	96.63%
8	64	4	96.08%	94.48%	96.76%
9	128	8	**96.18%**	94.91%	97.05%
10	256	8	96.25%	95.71%	97.21%
11	512	8	**96.25%**	**96.18%**	**97.21%**

In figure 5 shows perform the experiment on changed value of C and fixed value of gamma. Fv3 produce the highest accuracy 97.21%. First highest accuracy of fv3 is 97.21% and second highest is 97.08%. Second highest accuracy 96.25% observed by Fv1. Fv2 provide first highest accuracy 95.71% and second highest accuracy is 95.21%. Recognition accuracy increase when value of C taken above 68096, but SVM only us allow to take value of c up to $2^{15.}$



	C=1	C=2	C=4	C=8	C=16	C=32	C=64	C=128	C=256	C=512
fv1(16)	93.86	94.65	94.86	95.3	95.67	96.01	96.08	96.18	96.3	96.25
fv2(16)	89.26	90.18	91.15	92.28	93.01	93.91	94.48	94.91	95.21	95.71
fv3(32)	93.76	94.76	95.38	95.8	96.25	95.63	96.76	97.05	97.08	97.21

Fig 5 Arabic numeral accuracy using SVM with different value of C and fixed value of gamma

B.Comparative Analysis

Table 2 Comparison with earlier approaches

Proposed by	Feature Extraction Techniques	Classification Techniques	Name of language	Year	RR
Chun Lei He et al. [12]	Gradient features	Support Vector Machines (SVM) with RBF	Arabic	2010	97.81%
SaeedMozaffari et al. [13]	moment features and wavelet features	Neural Networks and Hidden Markov	Arabic		92.71%

		Models			
Abdelmalek zidouri [14]	Topological, Structural features		Arabic	2004	97.50%
Muhammad Imran Razzak et al [15]	Directional features	Fuzzy rule-based classifier uses linguistic variables.	Urdu and Arabic	2009	96.30%
Our proposed work	Zone based techniques	SVM with RBF kernels	Arabic	2012	97.91%

V. Conclusion

In this paper we introduced a system that recognizes Arabic handwritten Digits with 97.91% efficiency. The system uses Different zone based features. The features are extracted from all zones which contain the image. SVM classifiers are then used, for each image recognition. Feature Vector three is used to achieve a final recognition accuracy of 97.91%.

VI. Acknowledgment

The author would like two thanks ISI Kolkata, university for provision of Arabic numeral dataset for experiment. So, we achieve level of accuracy.

References

[1] C. L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, “Handwritten digit recognition: benchmarking of state-of-the- art techniques,” Pattern Recognition, vol. 36, pp. 2271–2285, 2003.

[2]. Chun Lei He Ching Y. Suen “Error Reduction based on Error Categorization in Arabic Handwritten Numeral Recognition” 2010 IEEE DOI 10.1109/ICFHR.2010.125.

[3] Faruq Al-Omari, Ph.D. “Hand-Written Indian Numerals Recognition System Using Template Matching Approaches”.

[4]. Saeed Mozaffari, Karim faez, Hamidreza Rashidy Kanan “Feature Comparison between Fractal Codes and Wavelet Transform in Handwritten Alphanumeric Recognition Using SVM Classifier” 17th International Conference on Pattern Recognition (ICPR’04) 1051- 4651/04 IEEE

[5] Abdelmalek Zidouri “ORAN: A BASIS FOR AN ARABIC OCR SYSTEM” Proceedings of 2004 International Symposium on Intelligent Multimedia. Video and Speech Processing October 2[J.22.2004 Hong Kong

[6]. Sherif Abdel Azeem “Arabic Handwriting Recognition using Concavity Features and Classifier Fusion” 978-0-7695-4607-0/11 2011 IEEE

[7]. Sabri A. Mahmoud “Arabic (Indian) Handwritten Digits Recognition Using Gabor-based Features” 978-1-4244-3397-1/08 2008 IEEE.

[8]. Alireza Alaei, Umapada Pal, and P. Nagabhushan “Using Modified Contour Features and SVM Based Classifier for the Recognition of Persian/Arabic Handwritten Numerals” 978-0-7695-3520-3/09, 2009 IEEE DOI 10.1109/ICAPR.2009.14

[9] 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.

[10] 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.

[11] 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).

[12] Chun Lei He Louisa Lam Ching Y. Suen “Automatic Discrimination between Confusing Classes with Writing Styles Verification in Arabic Handwritten Numeral Recognition” 2010 International Conference on Pattern Recognition.

[13]. Saeed Mozaffari, Karim faez, Hamidreza Rashidy Kanan “Feature Comparison between Fractal Codes and Wavelet Transform in Handwritten Alphanumeric Recognition Using SVM Classifier” Proceedings of the 17th International Conference on Pattern Recognition (ICPR’04)1051-4651/04 IEEE

[14]. Abdelmalek Zidouri “ORAN: A BASIS FOR AN ARABIC OCR SYSTEM” Proceedings of 2004 International Symposium on Intelligent Multimedia.

[15] Muhammad Imran Razzak S. A. Hussain Muhammad Sher “Numeral Recognition for Urdu Script in Unconstrained Environment” 978- 1-4244-5632-1/092009 IEEE