IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 9 (September. 2013), ||V2|| PP 50-56

Automatic Recognition of Offline Handwritten Urdu Digits In Unconstrained Environment Using Daubechies Wavelet Transforms

Imtiyaz Ahmed Ansari, Dr. R. Y. Borse
[1](*Research Scholar, JJT University, Rajasthan, India*[1])
[2](*Dept of Electronic Science, University of Pune, Maharashtra, India*)

Abstract: - This paper presents an optical character recognition system for the handwritten Urdu Digits. A lot of work has been done in recognition of characters and numerals of various languages like Devanagari, English, Chinese, and Arabic etc. But in case of handwritten Urdu Digits very less work has been reported. Different Daubechies Wavelet transforms are used in this work for feature extraction. Also zonal densities of different zones of an image have been used in the feature set. In this work, 200 samples of each digit have been used. The back propagation neural network has been used for classification. An average recognition accuracy of 92.07% has been achieved.

Keywords: - *Handwritten Digits, Daubechies, Wavelet Transforms, Feature Extraction, Zonal Densities*

I. INTRODUCTION

The goal of Optical Character Recognition (OCR) is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of OCR involves several steps including segmentation, feature extraction, and classification.

A few examples of OCR applications are listed here. The most common for use OCR is the first item listed below, people often wish to convert text documents to some sort of digital representation. Character recognition is a process of converting an image of a handwritten or printed text in to a computer editable format.

1. People wish to scan in a document and have the text of that document available in a word processor.
2. Recognizing license plate numbers
3. Post Office needs to recognize zip-codes
4. Facial feature recognition (airport security) – Is this person a bad-guy?
5. Speech recognition – Translate acoustic waveforms into text.
6. A Submarine wishes to classify underwater sounds – A whale? A Russian sub? A friendly ship?

There are many external and internal problems which are present in an OCR system for handwritten characters. The external problems are related to the variation in the shapes of characters and writing styles of different writers. There is a possibility of wrong recognition due to similarity between different characters. The internal problems in an OCR system are related to the distortion in the character images during scanning of images, addition of noise during image acquisition and degraded and broken characters images.

These problems lead to reduction in the accuracy of the offline handwritten character recognition. Same problems are there in case of handwritten Urdu Digits. The ten Urdu language digits are shown in Figure-1. Recognition of Handwritten Urdu digits is a complicated task due to the cursive and unconstrained shape variations, different writing style and different kinds of noise that break the strokes primitives in the character or change their topology.

The digit writing stroke, length, width, orientation and other geometrical features of a digit changed while writing the same digit. In Figure-2 variation in handwriting of same writer are shown with respect to every attempt.

English Digit	Urdu Digit	English Digit	Urdu Digit
0	۰	5	۵
1	۱	6	۶
2	۲	7	۷
3	۳	8	۸
4	۴	9	۹

Fig. 1: English and Urdu Digits with their equialence

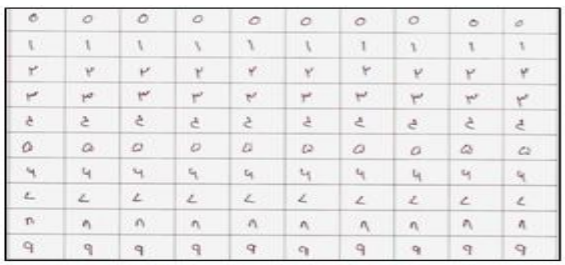

Fig.2: Handwritten Urdu Digits Samples of one person

The problem of recognizing characters in images has been extensively studied in the last few decades, mostly in the context of scanned documents and books [1, 2]. Handwriting recognition has also been widely addressed by both academia and industry [3]. As a result of focused research over many years, automated systems can now perform many of these tasks with accuracy rivalling human beings.

The handwritten recognition of Urdu Digits finds application in the field of automatic sorting of letters in postal services based on the postal codes, in automatic processing of various handwritten forms in various government departments and institutes, in digitization of old manuscripts etc.

S. A. Husain et al.[4] presented a method for recognition of online Cursive Urdu hand written Nastaliq Script. The system wss currently trained for 250 ligatures. The Recognition rate of base ligatures was 93% and of the secondary strokes was 98%. Sabri A. Mahmoud and Sameh M. Awaida [5] had presented a system for automatic independent writer off-line handwritten Arabic (Indian) numeral recognition, based on a quasi-multi-resolution approach to feature extraction using SVM. The database had 44 writers with 48 samples of each digit samples. The recognition results of SVM were compared to those of the HMM classifier. The achieved average recognition rates were 99.83% and 99.00% using, respectively, the SVM and HMM classifiers. Yuval Netzer et al. [6] had presented the problem of recognizing digits in a real application using unsupervised feature learning methods: reading house numbers from street level photos. To this end, we introduce a new benchmark dataset for research use containing over 600,000 labeled digits cropped from Street View images. Mohamed Abaynarh et al.[7] had used Legendre moments features for recognizing Amazighe characters. The system showed good performance (97%) on a database of 7524 handwritten Amazighe characters. Mohamed Abaynarh et al. had [8] presented unconstrained handwritten Amazighe character recognition based upon orthogonal moments and neural networks classifier. The result shows that if the number of hidden nodes increases the number of epochs (iterations) taken to recognize the handwritten character is also increases. The proposed system extracts moments features from character images and accuracy achieved is 97.46%.

Mostofa Kamal Nasir and Mohammad Shorif Uddin [9] developed a method that was based on preprocessing, k-means clustering, Bayesian theorem and SVM. Number of sample digits was 300.Success rate achieved was 99.33%. Ban N. Dhanoon and Huda H. Ali [10] proposed a method for handwritten numerals recognition. The method was simply depends on determining number of terminal points and its positions for each digit in its different shapes, that represent the main feature for recognition. The proposed method was based on structural primitives such as curve, line, point type etc. in a manner similar to that in which human beings describe characters geometrically. Anilkumar N. Holambe et al.[11] developed a system for extracting feature of handwritten and ISM printed characters of devanagri script. Sobel and Robert operator were used for extracting Gradient feature of the devanagari script . In this System computing gradient used was 8,12,16,32 directions and getting different feature vectors. Vikas J Dongre et al. [12] has given a review of various techniques used for feature extraction and classification of Devnagari character recognition. The various feature extraction techniques like Fourier transforms, wavelets, zoning, projections etc has been discussed. Raju G. [13] has proposed an OCR system for Malayalam characters. The proposed feature extraction method has used different wavelet filters and MLP network has been used as classifier. An average recognition rate of 81.3% has been achieved. M Abdul Rahiman et al. [14] has also proposed a Malayalam OCR system. The proposed system has used Daubechies wavelet (db4) for feature extraction and neural networks for recognition. The system has been given an accuracy of 92%. G S Lehal et al. [15] has given an OCR system for printed Gurmukhi script. The feature extraction has been done using the structural features and binary classifier trees and nearest neighbour classifier has been used. It has been found that an accuracy of 96.6% has been obtained. Syed. Afaq Husain and Syed. Hassan[16] had present a paper for the off-line recognition of cursive Urdu Text. The methodology had been developed for the Noori Nastaliq Script [17]. Word (Ligature) based identification had been adopted instead of character based identification. A multi-tier holistic approach had been utilized to recognize ligatures from a pre-defined ligature set. The system was currently trained for a small number of ligatures. Abdurazzag Ali ABURAS and Salem M. A. REHIEL [18] were proposed a new structure of off line OCR system which uses the wavelet image compression 40x40 bitmap image as input this produces a decomposition vector for each character. These vectors can uniquely represent the corresponding characters. Haar Wavelet transform used. The wavelet of level three of details was applied since higher level of details did not give better results. Average accuracy achieved was average 80%.

II. PROPOSED RECOGNITION SYSTEM FOR URDU DIGITS

According to Tou and Gonzalez[19], "The principal function of a pattern recognition system is to yield decisions concerning the class membership of the patterns with which it is confronted." In the context of an OCR system, the recognizer is confronted with a sequence feature patterns from which it must determine the character classes. If we model the character classes by their estimated means, we can use a distance measure for classification. The class to which a test character is assigned is that with the minimum distance. There are two steps in building a classifier: training and testing see Figure-3.

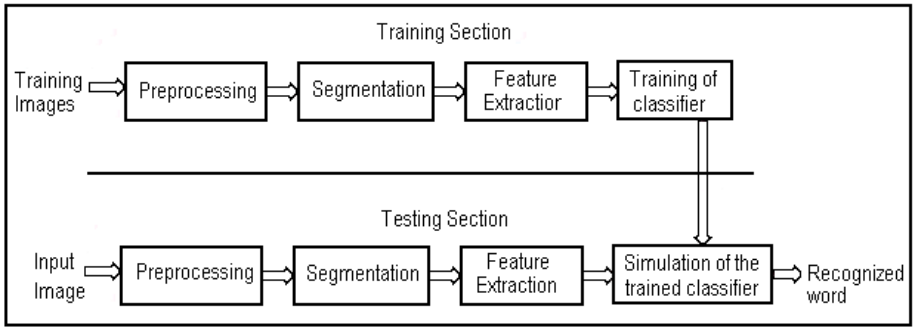

Fig. 3: Steps in building a classifier: training and testing

These steps can be broken down further into sub-steps.

1. Training
a) Pre-processing – Processes the data so it is in a suitable form.
b) Segmentation – dividing image to segments
c) Feature extraction – Reduce the amount of data by extracting relevant information, Usually results in a vector of scalar values. (We also need to NORMALIZE the features for distance measurements!)
d) Model Estimation/ Training Classifiers – from the finite set of feature vectors, need to estimate a model (usually statistical) for each class of the training data.

2. Testing
a) Pre-processing
b) Segmentation
c) Feature extraction – (both same as above)
d) Simulation/Classification of trained Classifier – Compare feature vectors to the various models and find the closest match. One can use a distance measure.

Handwritten OCR system for Urdu Digits consists of several stages see Figure-4. The stages for recognition process of handwritten Urdu digits are given below:
1. Image acquisition
2. Pre-processing
3. Feature extraction using different wavelet transforms
4. Classification using BP neural network
5. Recognized character

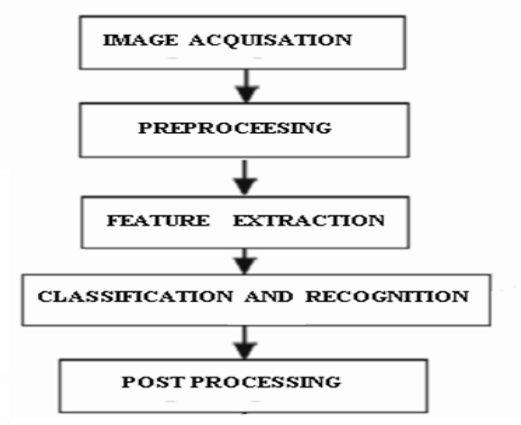

Fig. 4: Block Diagram of Handwritten Urdu Digit Recognition System

2.1. Image Acquisition

The handwritten Urdu Digit samples are taken from different writers. There are total 2150 samples which have been used in the proposed recognition system. 2000 samples have been used for training of neural network and 150 samples have been used for testing. These samples are taken by scanning the handwritten Digits at 400 dpi. Some samples of Urdu samples out of 2000 have been shown below in Figure 5 and samples out of 150 have been shown below in Figure-6.

Fig. 5: Urdu Digit sample for training

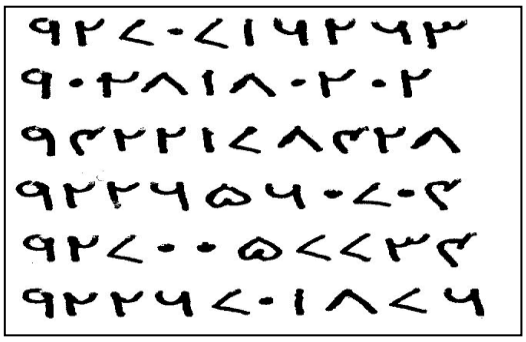

Fig. 6: Urdu Digit sample for testing

2.2. Pre-Processing

In the pre-processing stage the Recognition system has given a raw scanned colour image then following operations has been performed on it:

1. Conversion of colour image in to grey image.
2. Median filtering is performed to the image to remove noise [20].
3. The image then converted in to the binary image using thresholding.
4. The binary character image is normalized to 64*64.

2.3. Feature Extraction

Wavelets are localized basis functions which are translated and dilated versions of some fixed mother wavelet. The decomposition of the image into different frequency bands is obtained by successive low-pass and high-pass filtering of the signal and down-sampling the coefficients after each filtering. In this work Daubechies (db1, db2, …, Bb10) discrete wavelet transforms are used[21, 22].

The feature extraction is done by using the following algorithm:

For each pre-processed image following steps have been repeated:

a. Number of black pixels along each row of the binarized image has been counted to form a 32 sized vector.
b. The 1D wavelet transform on row count vector (two levels) has been applied.
c. Then the approximation (low frequency or average) coefficients have been directly taken as feature values.
d. Number of black pixels along each column has been counted to form a 64 sized vector.
e. The 1D wavelet transform on column count vector (three levels) has been applied.
f. Then the approximation coefficients have been directly taken as next feature values.
g. Divide each 64*64 image in to 16 zones of size 8*16.
h. Then find the mean zonal densities of these 16 zones. Out of 16 low level image with lesser texture image is ignored and 15 images are considered for next step.
i. Take these as the next 15 values of feature vector.
j. Take aspect ratio as the last feature element of the feature vector.

Above explained steps are repeated with different wavelet filters viz. db1, db2, ….., db10 After the feature extraction has been done, the feature vectors lengths are summarized in the Table-I:

Wavelet Filter	length of feature vector
db1	64
db2	66
db3	68
db4	70
db5	72
db6	74
db7	76
db8	78
db9	80
db10	82

Table-I: LENGTH OF FEATURE VECTORS

2.4. Classification using BP neural network

The back propagation neural network has been used for classification of the Urdu digits. Back Propagation Neural Network (BPNN), is a Multilayer Neural Network which is based upon back propagation

algorithm for training. This neural network is based upon extended gradient-descent based Delta learning rule, commonly known as Back Propagation rule. The basic architecture of a back propagation neural network has been shown in Figure 7.

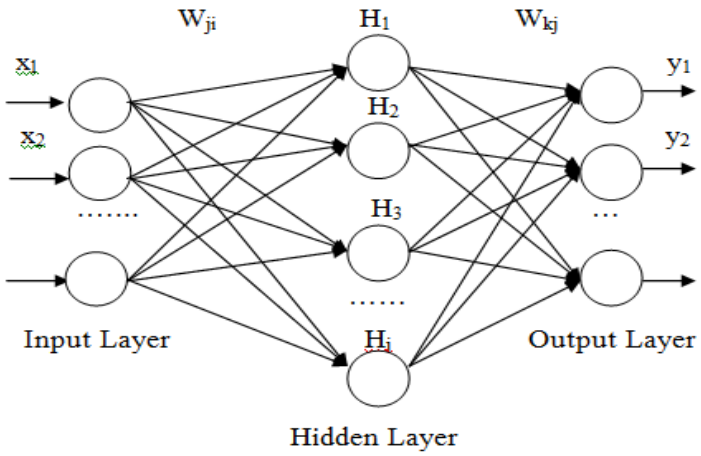

Fig.7: Back Propagation Neural Network Architecture

In this network, error signal between desired output and actual output is being propagated in backward direction from output to hidden layer and then to input layer in order to train the network. In this network input nodes equal to number of feature vector elements, 25 hidden layer with 51 input nodes and 1 output nodes are used. The testing input is fed into the input layer, and the feed forward network will generate results based on its knowledge from trained network.

The introduction of the paper should explain the nature of the problem, previous work, purpose, and the contribution of the paper. The contents of each section may be provided to understand easily about the paper.

III. RESULT AND DISCUSSION

In this work, various Daubechies Wavelet Transforms e.g. db1 to db10 have been used to extract the wavelet coefficients. Then a feature vector has been obtained by combining the wavelet coefficients, zonal densities and aspect ratio which is given as input to the BPN network. The outcomes have been summarized in Table-II. The values of average recognition accuracy using different Daubechies wavelets are 92.07% and average recognition time is 1.42.

Table-II: Comparisons of Recognition Accuracy using different Daubechies Wavelets

Wavelet Filter	%Recognition Rate	Recognition Time (in sec)
db1	82.2%	1.0247
db2	95.4%	1.1015
db3	92.85%	2.0003
db4	91.55%	1.4659
db5	76.15%	1.501
db6	95.8%	1.1163
db7	98%	1.4792
db8	98%	1.5082
db9	95.55%	1.5725
db10	95.25%	1.4754
Average	92.07%	1.4245

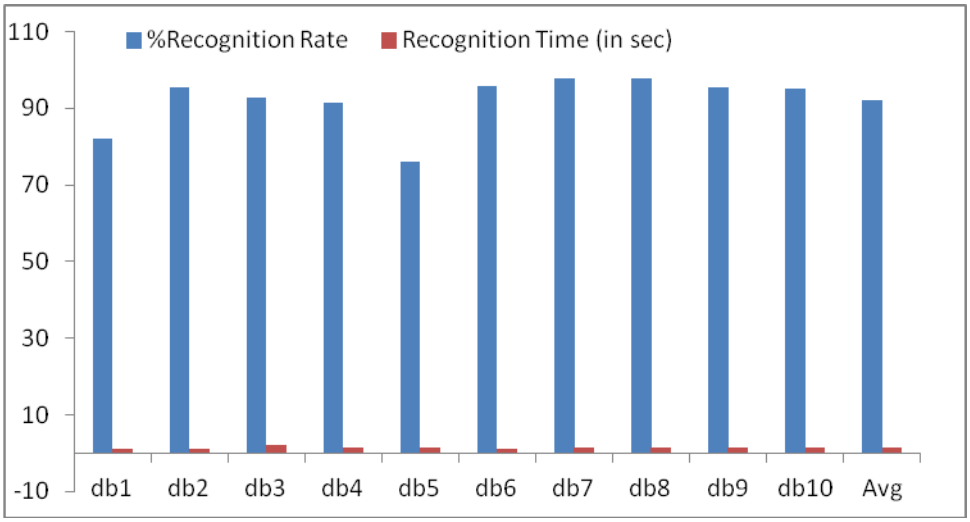

Fig. 8: Recognition rate and recognition time using different Daubechies wavelets

IV. CONCLUSION

In this recognition system an average recognition rate of 92.07% has been obtained. It has been found that db7 and db8 wavelets have given the highest recognition accuracy. The db1 wavelet has the least recognition time 1.02 seconds. As the size and quality of database is major factor influencing OCR systems, so relatively large database can be used in the future work. This will help to enhance the recognition accuracy. By adding some more features can also be helpful to enhance the recognition accuracy. It has been observed in this work that certain digits have confused with other digits during recognition process. It has been found that for numeral 6 recognition accuracy is least. The confusion matrix for each Urdu Digit using db7 is shown below:

Table- III: The Confusion Matrix for each Urdu Digit using db7

Urdu Digit	Recognition as per Urdu Digit									
	0	1	2	3	4	5	6	7	8	9
0	99	1	0	0	0	0	0	0	0	0
1	0	97	2	0.5	0.5	0	0	0	0	0
2	0	0.5	98	1.5	0	0	0	0	0	0
3	0	0	1	98	1	0	0	0	0	0
4	0	0	0	0.5	99.5	0	0	0	0	0
5	0	0	0	0	0.5	98.5	0	0.5	0	0.5
6	0	0	0	0	0	2.5	96.5	0.5	0.5	0
7	0	0	0	0	0	0.5	0.5	99	0	0
8	0	0	0	0	0	0	0.5	1	98.5	0
9	0	0	0	0	0	0	0.5	1	1	95

V. ACKNOWLEDGEMENTS

The authors gratefully acknowledge moral support from Management, M. G. Vidyamandir, Nasik, Principal, M. S. G. College, Malegaon providing facilities to work in the premises.

REFERENCES

[1] G. Nagy. At the frontiers of OCR, Proceedings of IEEE, 80 (7), 1992, 1093–1100.

[2] S. Mori, C. Y. Suen, and K. Yamamoto, Historical review of OCR research and development, Proceedings of IEEE, 80(7), 1992, 1029–1058.

[3]] R. Plamondon and S. N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. I EEE Trans. Pattern Anal. Mach. Intell, 22(1), 2000, 63–84.

[4] S. A. Husain, Asma Sajjad, Fareeha Anwar, "*Online Urdu Character Recognition System*", MVA 2007 *IAPR Conference on Machine Vision Applications, Tokyo*, JAPAN, May 16-18, 2007,

[5] Sabri A. Mahmoud and Sameh M. Awaida" RECOGNITION OF OFF-LINE HANDWRITTEN RABIC (INDIAN) NUMERALS USING MULTI-SCALE FEATURES AND SUPPORT VECTOR MACHINES VS. HIDDEN MARKOV MODELS", The Arabian Journal for Science and Engineering, Volume 34, Number 2B, October 2009.

[6] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning", research.google.com/pubs/archive/37648.pdf

[7] Mohamed Abaynarh, Hakim Elfadili, Khalid Zenkouar, and Lahbib Zenkouar, "Neural Network Classifiers for Off-line Optical Handwritten Amazighe Character Recognition", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.6, June 2012.

[8] Mohamed Abaynarh, Hakim Elfadili and Lahbib Zenkouar , " Handwritten Characters Classification Using Neural Networks and Moments Features", International Journal of Modern Engineering Research (IJMER) Vol.2, Issue.5, ISSN: 2249-6645, Sep-Oct. 2012, pp-3572-3577.

[9] Mostofa Kamal Nasir1 and Mohammad Shorif Uddin, "Hand Written Bangla Numerals Recognition for Automated Postal System", IOSR Journal of Computer Engineering (IOSR-JCE)), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 6, Jan. - Feb. 2013, PP 43-48.

[10] Ban N. Dhanoon and Huda H. Ali, "Handwritten Hindi Numerals Recognition", International Journal of Innovation and Applied Studies, ISSN 2028-9324 Vol. 3 No. 1, May 2013, pp. 310-317.

[11] Anilkumar N. Holambe, "Printed and Handwritten Character & Number Recognition of Devanagari Script using Gradient Features", International Journal of Computer Applications (0975 – 8887), Volume 2 – No.9, June 2010.

[12] Vikas J Dungre et al., A Review of Research on Devnagari Character Recognition, International Journal of Computer Applications, Vol. 12, No.2, November 2010

[13] Raju G., ―Wavelet Transform and Projection Profiles in Handwritten Character Recognition – A Performance Analysis , IEEE, 2008, pp. 309-314.

[14] M Abdul Rahiman, M S Rajasree, ―OCR for Malayalam Script Using Neural Networks, IEEE, 2009.

[15] G S Lehal and Chandan Singh, "*A Gurmukhi Script Recognition System", Proceedings of the International Conference on Pattern Recognition* (ICPR'00), 2000.

[16] Syed. Afaq Husain and Syed. Hassan Amin, "A Multi-tier Holistic approach for Urdu Nastaliq Recognition", IEEE INMIC, Karachi, Dec. 2002.

[17] Ahmad Mirza Jamil, "Noori Nastaliq, Computerized Urdu Calligraphy", (Elite Publishers), 1982.

[18] Abdurazzag Ali ABURAS and Salem M. A. REHIEL, "*Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression*", ARISER Vol. 3 No. 4, 2007, 123-135.

[19] J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, (Addison-Wesley Publishing Company, Inc., Reading, Massachusetts), 1974.

[20] Lim, Jae S., Two-Dimensional Signal and Image Processing, Englewood Cliffs, NJ, (Prentice Hall), 1990, pp. 469-476.

[21] Daubechies, I., *Ten lectures on wavelets, CBMS-NSF conference series in applied mathematics. SIAM Ed*. 1992.

[22] Mallat, S., "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Pattern Anal. and Machine Intell, vol. 11, no. 7, 1989, pp. 674–693.