International Journal of Research (IJR) Vol-1, Issue-7, August 2014 **ISSN 2348-6848**

Handwritten Urdu Character Recognition Using Zernike MI's Feature Extraction and Support Vector Machine Classifier

Devendra Singh Kaushal[1], Yunus Khan[2] & Dr. Sunita Varma[3]

[1]*Department of CSE*
Jawaharlal Institute of Technology Borawan Khargone M.P. India
devendrasinghkaushal@gmail.com

[2]*Department of CSE*
Jawaharlal Institute of Technology Borawan Khargone M.P. India
callyunuskhan@gmail.com

[3]*Department of CTA*
Shree G.S. Institute of Technology and Science Indore M.P. India
Sunita.varma19@gmail.com

Abstract:

This paper present handwritten Urdu Character recognition technique based on Zernike invariants and SVM as classifier. Automatic recognition of Handwritten Urdu numerals based on Zernike Moments Invariant (MI) features and SVM classifier is described. The hybrid approach of Zernike moment invariants has been adopted for feature extraction. The technique is independent of basic transformation and other variations in handwritten numerals.. Thus, overall 22 features corresponding to each numeral proceed for classification using Support Vector Machine (SVM) classifier. The success rate of the method is found to be 96.29%.

Keywords: SVM, Zernike Feature Extraction, Recognition Rate, Moment Invariants, Handwritten Urdu, Character Recognition, Zernike MI's Feature Extraction, Vector Machine Classifier, Zernike Moments Invariant

I. INTRODUCTION:

Handwritten recognition has always been a challenging task in pattern recognition. Many systems and classification algorithms have been proposed in the past years on handwritten character/numeral recognition in various languages like English, Persian, Arabian, Devanagari and, Urdu scripts also. Researchers had been worked on Handwritten Urdu characters by applying different techniques, but very less work has been performed on Handwritten Urdu numerals. So, this research work has been conducted on Handwritten Urdu numeral.

Recognition of Handwritten Urdu Numerals/ Characters is a complicated task due to the unconstrained shape variations, different writing style and different kinds of noise. Also, handwriting depends much on the writer and because we do not always write the same digit in exactly the same way, building a general recognition system that would recognize any digit with good reliability in every application is not possible. In recent time, specialists have made use of different techniques such as Modified Discrimination Function (MQDF) , Multilayer Perceptron (MLP), Principal Component Analysis (PCA) , K-Nearest Neighbor , Wavelet-based multi resolution , Quadratic classifier have been applied to solve this problem. These recognition systems are produces the recognition rate between 89% and 99.04%. To achieve these accuracies the researchers used various feature extraction techniques also. They considered large number of features for recognition.

A common task in Machine Learning is to classify the data using training and testing. Support Vector Machine (SVM) is one of the better classifier among all Machine Learning algorithms for pattern

recognition. So, linear SVM is chosen as classifier for getting the better recognition rate in our research work for classification. In order to provide a basis for classification SVM implicitly map the training data into high dimensional feature space. Hyper plane is then constructed in this feature space which maximizes the margining of separation between the plane and those points lying near to it. The plane so constructed can then be used as a basis for classifying vector of uncertain type.

In recent time, specialists have made use of different types of tools for feature extraction. A few reports include on fuzzy features, Hu's invariants moment features. However, different suggested approaches on recognition of Urdu Handwritten numerals have not yet been achieved satisfactory success rate. In numeral recognition problem, the description phase plays a fundamental role, since it defines the set of

Properties which are considered essential for characterizing the pattern. Moments & function of moments have been utilized as pattern feature in number of application. Hu first introduced Zernike moment invariants in 1961, based on the theory of algebraic invariants. Using non-linear combinations of geometric moments, a set of moment invariants has been derived; these moments are invariant under image translation, scaling, rotation & reflection. A number of papers describing application of invariant moment with its types (e.g. complex moments, rotational moments, hu's moments, Legendre moments etc) have been published. Few reports are on comparative study of Fourier Descriptors and Zernike Moment Invariants (MIs). They showed comparatively better results with MIs. A comparison is also made in with affine moment invariants [AMI]. A. G. Mamistvalov presented the proof of generalized fundamental theorem of moment invariants for n-dimensional pattern recognition. He has formulated correct fundamental theorem of Moment Invariants. Using these moments, the conceptual mathematical theory of recognition of geometric figures, solids and their n-dimensional generalization is worked out.

Handwritten Numeral Recognition system typically involved two steps: feature extraction in which the patterns are represented by a set of features and classification in which decision rules for separating pattern classes are defined. Recognition of Handwritten Urdu Numerals is a complicated task due to the unconstrained shape variations, different writing style and different kinds of noise that break the strokes primitives in the character or change their topology.

II. URDU SCRIPT

In India there are twelve scripts and Urdu is one of the popular Indian scripts. Here we describe some properties of the Urdu script that are useful for building the OCR system. The modern Urdu alphabet consists of 39 basic characters. These characters are shown in Fig.1. Urdu has 10 numerals and the numerals are shown in Fig.2. Like other Indian scripts in Urdu also two or more characters may combine and create a complex shape called *compound characters*. As a result, the total number of characters to be recognized is very large. Thus, OCR development for Urdu is more difficult than any European language script having a smaller number of characters.

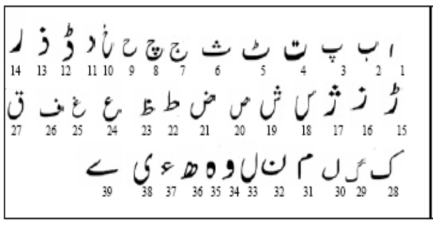

Figure 1: Urdu Characters

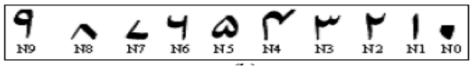

Figure 2: Urdu Numerals

III. SYSTEM METHODOLOGY

Identifying correct features is major part in handwriting recognition system. Feature extraction is essential for efficient data representation and for further processing. Also, high recognition performance can be obtained by selecting suitable feature extraction method. Computational complexity of a classification problem can also be reduced if suitable features are selected. Features vary from one script to another script and the

method that gives better result for a particular script cannot be applied for other scripts. Also, there is no standard method for computing features of a script. It is worth to note that features must vary to a reasonable extent and must be available in different users' cursive handwriting. Also, these features should be measurable through algorithms.

This research work extracts the features from numeral images using Zernike Moment Invariant, Blur Invariants and Affine Moment Invariant Feature Extraction Method. Both methods are described in literature work.+

ZERNIKE MOMENT BASED FEATURE EXTRACTION

Zernike moments are complex number by which an image is mapped on to a set of two-dimensional complex Zernike polynomials. The magnitude of Zernike moments is used as a rotation invariant feature to represent a character image patterns. Zernike moments are a class of orthogonal moments and have been shown effective in terms of image representation. The orthogonal property of Zernike polynomials enables the contribution of each moment to be unique and independent of information in an image. A Zernike moment does the mapping of an image onto a set of complex Zernike polynomials. These Zernike polynomials are orthogonal to each other and have characteristics to represent data with no redundancy and able to handle overlapping of information between the moments . Due to these characteristics, Zernike moments have been utilized as feature sets in applications such as pattern recognition and content-based image retrieval . These specific aspects and properties of Zernike moment are supposed to found to extract the features of compound handwritten characters. Teague + has introduced the use of Zernike moments to overcome the shortcomings of information redundancy due to geometric moments.

The Zernike moment were first proposed in 1934 by Zernike. Their moment formulation appears to be one of the most popular, outperforming the alternatives (in terms of noise resilience, information redundancy and reconstruction capability). Complex Zernike moments are constructed using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc (x2+y2) _ 1.They are expressed as A_{pq}. Two dimensional Zernike moments:

$$A_{mn} = \frac{m+1}{\pi} \int_x \int_y f(x,y)[V_{mn}(x,y)]^* dx\ dy$$
$$where\ x^2 + y^2 \leq 1\ \bigtriangledown_y\ (x_i + h, y_j + k) \quad (1)$$

where m = 0; 1; 2; :::;1 and defines the order, f(x; y) is the function being described and * denotes the complex conjugate. While n is an integer (that can be positive or negative) depicting the angular dependence, or rotation, subject to the conditions:

$$m - |n| = even, |n| \leq m \quad (2)$$

m □ jnj = even; jnj _ m (2) and A_mn = Am;□n is true. The Zernike polynomials [20] Vmn(x; y)Vmn(x; y) Zernike polynomial expressed in polar coordinates are:

$$V_{mn}(r,\theta) = R_{mn}(r)exp(jn\theta) \quad (3)$$

Vmn(r; _) = Rmn(r)exp(jn_) (3) where (r; _) are defined over the unit disc, j =p□1 and Rmn(r) and is the orthogonal radial polynomial, defined as Rmn(r) Orthogonal radial polynomial:

$$R_{mn}(r) = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s\ F(m,n,s,r) \quad (4)$$

Where:

$$F(m,n,s,r) = \frac{(m-s)!}{s!\left(\frac{m+|n|}{2}-s\right)!\left(\frac{m-|n|}{2}-s\right)!} r^{m-2s} \quad (5)$$

where Rmn(r) = Rm;□n(r) and it must be noted that if the conditions in Eq. 2 are not met, then Rmn(r) = 0. The first six orthogonal radial polynomials are:

$$R_{00}(r) = 1 \quad R_{11}(r) = r$$
$$R_{20}(r) = 2r^2 - 1 \quad R_{22}(r) = r^2$$
$$R_{31}(r) = 3r^3 - 2r \quad R_{33}(r) = r^3 \quad (6)$$

So for a discrete image, if Pxy is the current pixel then Eq.

$$A_{mn} = \frac{(m+1)}{\pi} \sum_x \sum_y P_{xy} \left[V_{mn}(x,y)\right]^* \quad where\ x^2 + y^2 \leq 1 \qquad (7)$$

To calculate the Zernike moments, the image (or region of interest) is first mapped to the unit disc using polar coordinates, where the centre of the image is the origin of the unit disc. Those pixels falling outside the unit disc are not used in the calculation. The coordinates are then described by the length of the vector from the origin to the coordinate point, r, and the angle from the x axis to the vector r:r Polar co-ordinate radius, _:_ Polar co-ordinate angle, by convention measured from the positive x axis in a counter clockwise direction. The mapping from Cartesian to polar coordinates is:

$$x = r\cos\theta \quad y = r\sin\theta \qquad (8)$$

$$r = \sqrt{x^2 + y^2} \quad \theta = \tan^{-1}\left(\frac{y}{x}\right) \qquad (9)$$

After performing preprocessing on each image, each image is ready for extracting features. In preprocessing, each image is binarized. This binarized image is complemented and we get image having white color numeral. From each complemented 8 features is extracted using Zernike Moment Invariant method. Also, 14 more features are extracted from thinned image as well as Blur Invariants of the same numeral using Moment Invariant method.

IV SUPPOT VECTOR MACHINE CLASSIFIER

The objective of recognition is to interpret a sequence of numerals taken from the test set. Any new numeral that is to be recognized is preprocessed first. Features extracted from this numeral are sent to the Classifier. An SVM is basically defined for two-class problem and it finds the optimal hyper-plane which maximizes the distance, the margin, between the nearest examples of both classes, named support vectors (SVs). The SVM (binary classifier) is applied to multiclass numeral recognition problem by using one-versus-rest type method. The problem now is a 10-class problem with 10 equal to the number of segments in total. The SVM is trained with the training samples using linear kernel.

Classifier performs its function in two phases; Training and Testing. After preprocessing and Feature Extraction process, Training is performed by considering the feature vectors which are stored in the form of matrices. Result of training is used for testing the numerals

The sample image of Urdu numeral is given as input. Support vector machine is primarily the classified method that performs classification task by constructing hyper plane in multidimensional space that separate cases of different class labels. To construct optimal hyper plane, SVM employs an iterative training algorithm, which is use to minimize an error function. The support vector machine classifier is optimizing an error function that minimizes the misclassification on the training set. Input images of each digit have 210 samples. The overall recognition rate is 96.29%. The algorithm for recognition phase is as given below.

V. RESULT ANALYSIS

A result of Phase II was satisfactory but still I found scope of improvement in the program. So, Phase III was come into existence. In this phase, I decided to join Zernike Moment Invariant technique, Hu's Invariants and Blur Moment Invariant technique. Eight features from complemented image and eight features from thinned image were computed. So, total 16 features were extracted using Zernike Moment Invariant method. Six more features from thinned image are extracted using Affine Moment Invariant technique and added with the 16 features. Hence, total 22 features are extracted from each image. After executing SVM Training and Recognition program following results were obtained.

By combining the result of all three datasets, new system provides 96.29% recognition rate. Overall Recognition rate obtained from 2100 samples of Urdu numerals is depicted. Obtained recognition rate in this research work is found to be higher than all other techniques. While Support Vector Machine performs

training on datasets, these feature vectors are extracted and stored in matrix form which is also called as feature matrix. The performance using linear SVM classification method rate is presented in table 1.

VI.CONCLUSION AND FUTURE WORK

This research work deals with the recognition of Handwritten Urdu numeral by applying Support Vector Machine techniques of Machine Learning approach. Proposed methodology is implemented a program for preparation of database. So, we no need to create database manually which required more efforts than automated system. This automated program is named as Automated Numeral Extraction and Segmentation Program (ANESP). Some part of preprocessing is accomplished through this program. Proposed system reduces the manual work of dataset preparation. This system provides better dataset with clean images by using automated process. This methodology provides more efficient and accurate results than any other existing systems.

As a part of future work, recognition rate need to be tested by increasing datasets. This work implements linear kernel function of SVM. By application of other kernel functions such as Radial Basis Function (RBF), Polynomial Kernel function, Sigmoidal function, this accuracy of Urdu Handwritten numeral recognition can be further increased.

Dataset	**Total sampl**	**Recognize samples**	**Recognit n Rate (%**
DATASET 1	800	718	89.75%
DATASET 2	700	694	99.14%
DATASET 3	600	600	100.00%
Average Recognition Rate (%)			96.29%

Table 1

REFERENCES

[1] Project proposal (2011) "Software Requirement, Design, & Testing documents, Development of Robust Document Image Understanding System for Documents in Indian Scripts Phase II" ,Sponsored By, Ministry of Communication & Information Technology, Govt. of India.

[2] Pal U. and Anirban Sarkar (2003) 7th International Confer-ence on Document Analysis and Recognition, 1183-1187.

[3] Lorigo L.M., Govindaraju V. (2006) IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(5).

[4] Hoque S., Sirlantzis K., Fairhurst M.C. (2003) 7th International Conference on Document Analysis and Recognition, 2, 834-838.

[5] Husain S.A. and Amin S.H. (2002) International Multi Topic IEEE INMIC.

[6] Nain N., Laxmi V., Bhadviya B. (2007) 3rd International IEEE Conference, 821- 825

[7] Syed Afaq Hussain and Syed Hassan Amin (2002) Interna-tional Multi Topic IEEE INMIC.

[8] Zahra A. Shah and Farah Saleem (2002) International Multi Topic IEEE INMIC.

[9] Liana M. Lorigo, Venugopal Govindaraju (2006) IEEE Trans. Pattern Anal. Mach. Intell. 28(5), 712-724.

[10] Maged Mohamed Fahmy and Maged Mohamed (2001) Stud-ies in Informatics and Control, 10(2).

[11] Mandana Kairanifar and Adnan Amin (1999) 5th International Conference on Document Analysis and Recognition, 213.

[12] C.M. Naim (1999) Introductory Urdu - Volume I, Book Pub-lished by National Council for Promotion of Urdu Language.

[13] Zaheer Ahmad, Jehanzeb Khan Orakzai, InamShamsher and Awais Adnan (2007) World Academy of Science Engineering and Technology, 26.

[14] Gurpreet Singh Lehal (2010) 23rd International Conference on Computational Linguistics.

[15] Gheith A. Abandah and Mohammed Z. Khedher (2009) Inter-national Journal of Computer Processing of Languages, 22(1), 1-25

[16] Ramteke R.J., Mehrotra S.C. (2008) International Journal of Computer Processing of Oriental Languages.

[17] Ramteke R.J., Mehrotra S.C. (2006) IEEE Conference on Cybernetics and Intelligent Systems, 1-6.

[18] Abdulbari Ahmed Ali, Ramteke R.J. (2011) International Jour-nal of Machine Intelligence, 3(3), 116-120.