Zone based Method to Classify Isolated Malayalam Handwritten Characters using Hu-Invariant Moments and Neural Networks

Paulose Raj
P.G. STUDENT
Department of Information Technology
Bannari Amman Institute of Technology
Erode, Tamil Nadu, India

Amitabh Wahi
PROFESSOR
Department of Information Technology
Bannari Amman Institute of Technology
Erode, Tamil Nadu, India

ABSTRACT
Handwritten Character Recognition of Indian languages have been a demanding task in image processing and pattern recognition. Structural complexity and likeness in the characters also increases the complexity in the classification of characters. In this study, Malayalam, a south-Indian language investigated for recognition of its characters using Hu-invariant moments.Moments applied to the preprocessed image after zoning the image. The image divided horizontally, vertically and diagonally to which moments are applied. Feed-forward backpropagation neural network used for classification of characters with two hidden layers. A better Recognition rate of 93.7 percentagenoted.

General Terms
Pattern Recognition, Malayalam Handwritten Character Recogntion, Image Zoning, Image Moments, Hu-invariant Moments, Neural Networks.

Keywords
Malayalam characters; Offline character recognition; Hu-Invariant Moment; Neural network

1. INTRODUCTION
Character recognition is the method of recognition of characters into machine-understandable format. They classified as offline and online character recognition. Online character recognition is the recognition of characters from the pressure sensitive surface like digital tablet, Personal digital assistant (PDA). Real time information is used to process for feature extraction such as order of strokes, stroke angle etc. Whereas offline character recognition is the method of recognition of characters from the scanned, image or paper. Isolated handwritten character recognition is the process of recognition of isolated words of the language. Handwritten Character Recognition (HCR) is a popular research area capable of put use in application such as various aspects of banking, government, legal documents. It considered as a reading aid for blind people [9].

In India twenty-two scheduled languages are there namely (in the alphabetical order) Asamese, Bengali, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithei, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telungu and Urudhu. The notion of Upper-case letters and lower letters is not present in Indian languages. Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Mathili use Devanagri script to write. Panjabi languages use a Gurumukhi script to write. Most of the Indian scripts originated from very old Brahmi, which are phonetic in nature and complex writing shapes. Among them, except Urdu, written from left to right. Basic language set consist of vowels and consonants. Indian languages contain strokes, loops, curve, and arcs unlike of English character set where only the strokes are the major feature set.

2. MALAYALAM SCRIPT
Malayalam is the language used by peoples in Kerala, a south-Indian state in India. Like other Indian languages, Malayalam follows a writing procedure that is partially alphabetic and partially-syllable based. Malayalam language uses Brahmic script to write. Malayalam follows both old and new scripts for writing character set. Government of Kerala made some renovations in the Malayalam character set for easy printing. The major difference between old and new script is that old script adds up various characters whereas, new script keep apart the characters with a distinct character.

Left to right is the nature of writing and reading of Malayalam Characters. Malayalam characters uni-case in nature, that is, it does not have case distinction. Though left to right is the direction of writing and reading, some vowels attached to the left of a consonant letter that logically follows. Vatteluttu is the first written form of Malayalam, derived from an ancient script of Tamil. Grantha is the base for modern Malayalam language, which used to write Sanskrit.

Brahmi script is the base for the evolution of both Vatteluttu and Grantha. Total number of character set cannot be correctly predicted due to the usage of the Old and New script. The similarity of the shapes and complexity in the script of character also adds up the difficulty in the detection of the Malayalam characters. Handwritten character decreases the detection rate due to the irregularity and discrepancies in the characters fed to the system.

2.1 Vowels
Vowels are two types in the Malayalam. They are Independent Vowels and Dependent Vowels. They knew as Svaram and Svarksharangal in Language. The vowels used in the Present Malayalam script categorized in figure 1.

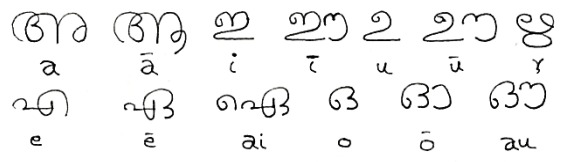

Fig 1: Vowels used in Malayalam Language (Handwritten)

2.2 Consonants

Vyanjanam or Vyanjanaksharangal is the name given to the consonants in the Malayalam Character. A consonant letter, regardless of its name, does not epitomize a wholesome consonant, but indicates a consonant and a short vowel. The consonants used in Modern Malayalam script characterized in figure 2.

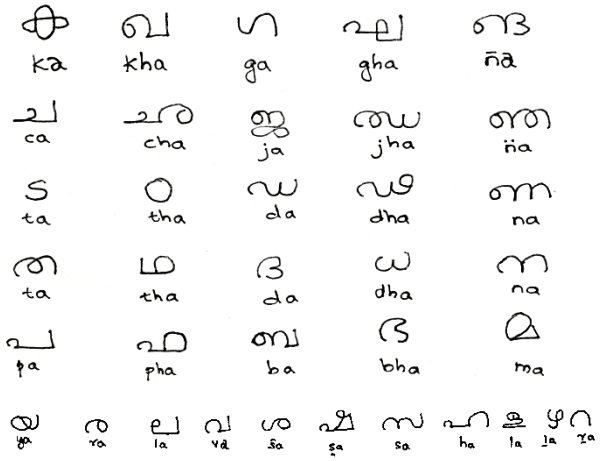

Fig 2:Consonantsused in Malayalam Language (Handwritten)

Special diacritic virama used to cancel the short inherent vowel. Method of usage of diacritic virama and several numbers of vowel diacritics of one individual consonant /ga/ given in the figure 3.

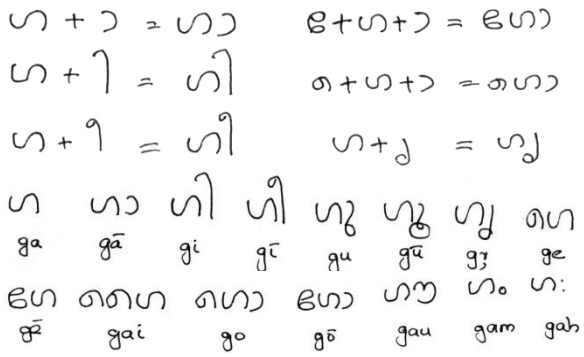

Fig 3:Usage of special diacritic viramaon character /ga/ (Handwritten)

2.3 Conjunct Consonants

Conjunct Consonant characters are far and wide used in the ancient lipi writings rarely used in the modern scripting too. These characters shaped by a combination or mixture of two or more characters and imperative in the language as they convey more information to the reader. In a Handwriting schema, correct method of defining these characters still not done.

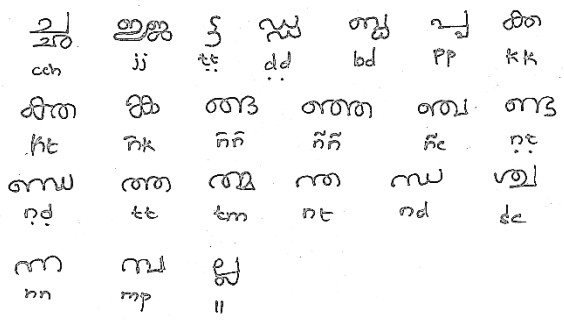

Fig 4:Conjunct Consonants used in Malayalam Language (Handwritten)

3. DEVELOPED METHODOLOGIES

Raju G. [1, 2, and 6] carried out a performance analysis of twelve different wavelet filters as a feature extraction purpose. Aspect ratio also considered as one of the feature. The classifier used was the MLP network results a recognition rate of 81.3%. Bindu Philip [3] studied Segmentation and Classification of Malayalam Characters through Characterization using Dominant Singular Values. Lajish V.L. explored methods using State-Space Map (SSM), State-Space Point Distribution (SSPD) parameters [4], the fuzzy-zoning and normalized vector distance measures [5]. Bindu Philip et al [7, 8, 9] proposed Malayalam OCR System uses Column-Stochastic Image Matrix Approach. A recognition rate of 97.08 % with an error rate of 2.92% obtained. The work extended for Bilingual OCR for English-Malayalam characters and a text aloud system for blind people. Binu P. Chacko [10] suggested character recognition method based on Discrete Curve Evolution Based Skeleton Pruning, which reduces spurious branch in the skeletonized images. M. Abdul Rahiman proposes an OCR for Printed Malayalam Character Recognition using Back-propagation Neural Networks [11]. Another method proposed is using HLH intensity patterns. The work is extended in recognition of combinational Malayalam characters and for a Bilingual OCR system for English and Malayalam [13, 14, 16]. Binu P. Chacko [15] suggested Pre-Processing and Post-Processing methodologies in Edge Detection for Malayalam Character Recognition. Bindu S Moni investigated Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition [18]. The work widened by using Run length count for the same [17]. Jomy John [19] made a study on offline OCR based on Chain Code histogram. Abdul Rahiman M proposed a Vertical & Horizontal Line Positional Analyzer Algorithm [20] for the recognition, a Method based on the vertical and horizontal line positions.

4. HU-INVARIANT MOMENTS AND IMAGE ZONING

Classification of the characters depends on the efficiency of the features extracted from the character image. Hu-invariant moment [21, 22] gives a notion of the global character shape information. Seven distributed parameters of the image evaluated. These Moments never change the values under translation, scaling and rotation [23, 24, 25, and 26] because they give the measures of the pixel spreading around the center of gravity thus results a general character silhouette.

Unvarying moments demarcated as

$$\mu_{ij} = \sum_{A} \sum_{B} (A - A')^{i} (B - B')^{j} f(A, B) \quad (1)$$

fori, j = 0,1,2,….

A' and B' are moments calculated from the geometrical moments m_{ij} as follows:

$$A' = m_{10} / m_{00} \quad (2)$$

$$B' = m_{01} / m_{00} \quad (3)$$

$$m_{ij} = \sum_{A} \sum_{B} A^{i} B^{j} f(A, B) \quad (4)$$

The standardized central moment to shape and size of order (i +j) defined as

$$\eta_{ij} = \mu_{ij} / \mu_{00}^{\kappa} \quad (5)$$

fori, j = 0,1,2,…

where $\kappa = \frac{i+j}{2} + 1$ (6)

for (i +j) = 2,3,…

A conventional of seven moment invariants derived from above equations (7)

$$\Omega_1 = \eta_{20} + \eta_{02}$$

$$\Omega_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$\Omega_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

$$\Omega_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$

$$\Omega_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

$$\Omega_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$

$$\Omega_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$

In this work, each segmented character resized as 128 x 128-pixel image. The image matrix f (A, B) processed to obtain the character with white edge color and black background color. The expressions given above used to extract features. In the experiment, the invariant moments evaluated by taking the log of the absolute value of the moment of each of the sample images of the characters.

To improve the detection rate further, the image zoning method applied to the characters [22]. The image used for the training further divided horizontally, vertically, and diagonally, keeping the size of image 128 x 128. Hu Invariant moments applied for these zones. For analysis purposes diagonal zoning also found (both left and right diagonal) Hu invariant moments of these images combined with the previous set and comparison made.

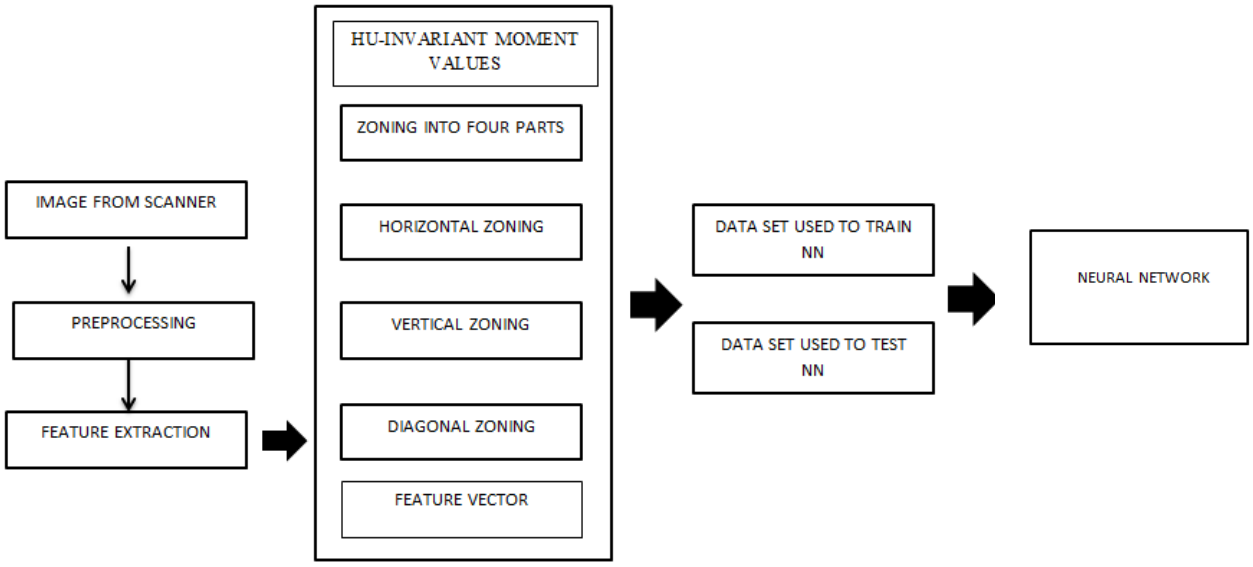

Fig 5:Block diagram of character Recognition schema

5. THE METHOD

The process includes Dataset colletion, Preprocessing, Feaure Extraction and Classification. The process diagram depicted in the figure 5.

5.1 Dataset Collection

A database containing different samples, size and variation in character is very important for evaluating a handwritten character recognition system. If a standard dataset is available, it could use for the purpose of training and testing, but no benchmarked data set is available. For the present study, we have considered only the consonant and vowels of the Malayalam language. About 250 samples of each Malayalam character have collected from the peoples of different age groups. 300 DPI is used for scanning the page. Each character written in a box, from which it segmented.

5.2 Preprocessing

Segmented characters are stored in the 128 x 128-image size. Some images with noise also collected. These noisy images are preprocessed using median filter to remove the salt and pepper noise. After applying the filter, binarization process in carrying out. Edge detection carried out to know the edges of the character. An even slight variation of the handwritten character could found using the edge detection. Here the method follows a moment-based approach of character recognition, so the area of interest only the region were character is written, all other information is to be removed. Using the principle of bounding box an algorithm has developed only to crop the character area. After completing these preprocessing steps, the image fed for feature extraction.

5.3 Feature Extraction

Application of Hu-invariant moments to the character without zoning will not give an effective feature vector. To improve accuracy of recognition the Hu-invariant momentaapplied after zoning the image. The following ways are adopted to divide and extract features from the image.

5.3.1 Zoning Into Four Parts (Feature Vector Set 1)

Prerocessed image divided into four parts through the centroid of the image. Hu-invarat moments calculated for each parts sepeartly.Thefeature vector formed by the four parts include 7 x 4 = 28. The figure 6 shows the method of division of preprocessed image.

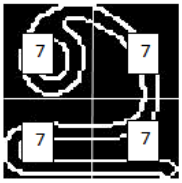

Fig 6:Method of zoning of image into four parts and its Hu-moment usage.

5.3.2 Horizontal And Vertical Zoning (Feature Vector Set 2)

Image is divided horizontally and the hu-moment applied to the two parts of the image results in 7 x 2 =14 feature vector values

Similarly, the image divided vertically and hu-moment is considered, which again constitutes 7 x 2 =14 feature vector values.Thus a total of 28 values formed in the feature vector set 2. The figure 7 shows the method of division of preprocessed image.

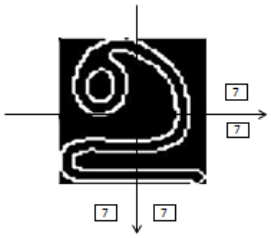

Fig 7:Method of division of horizonatal and vertical directions and its Hu-moment usage

5.3.3 Diagonal Zoning (Feature Vector Set 3)

The third feature set considered are extracted after dividing the image through the diagonals of character. This feature vector results more accurate values (7 x 2) x 2 =28. The figure 8 shows the method of division of preprocessed image.

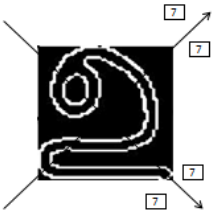

Fig 8:Method of division through diagonal and its Hu-moment usage

5.4 Neural Network Classifier

Classification accomplished by using feed forwadbackpropagationneural networks. Feature vector combined in a fashion which had feature vector set 1 and 2 (54 values) in one set and feature set 1,2 and 3 in another set (84 values).So the neural network with input layer 56 and 84 used and results compared. Neural Network with two hidden layers used, which hadTansig as activation function. The output layer depends on number of characters to be detected uses a Logsig function. From the handwritten Malayalam character feature vector datasets, values taken in a random order for testing and training purpose. They split in the ratio of 1:3. Number of elements in hidden layers chosen based on the lowest mean square error. At first the neural network trained with the backpropagation algorithm later on with feature vector values. The backpropagation algorithm is depiceted in the subsection

5.4.1 Back-propagation Algorithm

1. Randomly choose the initial weights
2. While error is too large
3. For each training pattern (presented in random order)
4. Apply the inputs to the network
5. Calculate the output for every neuron from the input layer, through the hidden layer(s), to the output layer
6. Calculate the error at the outputs
7. Use the output error to compute error signals for pre-output layers
8. Use the error signals to compute weight adjustments
9. Apply the weight adjustments

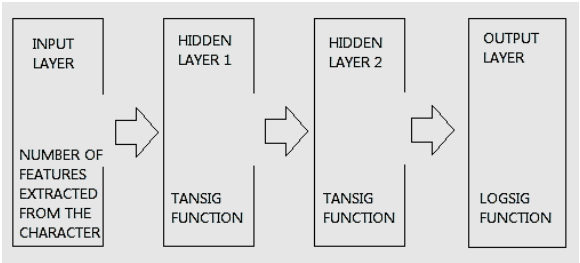

Fig 9:Architecture of Neural Network Used for Classification

6. RESULTS AND CONCLUSIONS

In this paper currently, 26 Malayalam characters trained on the system. Comparison made between the 56 values, without considering two diagonals and 84 values inclusive of diagonal characteristics. It found that features extracted inclusive of the diagonal came up with better performance and efficiency. Specific information about the character will be included when diagonal features also included. 4000 handwritten characters used to train and test the system out of which 100 characters from 26 classes were used for training the Neural Network and 30 characters used for testing the system. Characters from specific set also considered for training. Table 1 formulates the investigational results.

An accuracy of 86.7% obtained in the first experiment considering the feature set 1, 2. Better accuracy of 93.2% obtained for the classification having features set three along with the first experimental data set. Using more features and better classifiers improve efficiency of the system.

Classification of confusion characters that is, characters they look visually similar, also can be recognized in a better way.

TABLE 1: Performance of Neural Network on Malayalam Handwritten Character Recognition

Experiment Number	Feature sets	Neural Network Architecture	Recognition Rate (%)
1	1,2	56:75:37:26	86.7
2	1,2,3	84:93:58:26	93.2

7. REFERENCES

[1] G. Raju, "Recognition of Unconstrained Handwritten Malayalam Characters using Zero-crossing of Wavelet Coefficients", ADCOM 2006, Page(s): 217-221, 2006.

[2] Renju John, G.Raju , D. S. Guru, "1D Wavelet Transform of Projection Profiles for Isolated Handwritten Malayalam Character Recognition, International Conference on Computational Intelligence and Multimedia Applications 2007.

[3] Bindu Philip, R. D. Sudhaker Samuel , A Novel Algorithm for Segmentation and Classification of Malayalam Characters through Characterization using Dominant Singular Values, International Conference on Computing, Communication and Networking 2008.

[4] Lajish.V.L, Handwritten Character Recognition Using Gray-scale Based State-Space Parameters and Class Modular NN, International Conference on Signal processing, Communications and Networking,2008.

[5] Lajish.V.L, Handwritten Character Recognition using Perceptual Fuzzy-Zoning and Class Modular Neural Networks, IEEE 2008.

[6] Raju G., Wavelet Transform and Projection Profiles in Handwritten Character Recognition –A Performance Analysis, International Conference on Advanced Computing and Communications (ADCOM), 2008.

[7] Bindu Philip, A Malayalam OCR System using Column-Stochastic Image Matrix Approach, International Conference on Advances in Recent Technologies in Communication and Computing,2009.

[8] Bindu Philip, A Novel Bilingual OCR for Printed Malayalam-English Text based on Gabor Features and Dominant Singular Values, IEEE International Conference on Digital Image Processing, 2009.

[9] Bindu Philip, A Novel Bilingual OCR System based on Column- Stochastic Features and SVM Classifier for the Specially Enabled , Second International Conference on Emerging Trends in Engineering and Technology, 2009.

[10] Binu P Chacko, Discrete Curve Evolution Based Skeleton Pruning for Character Recognition, Seventh International Conference on Advances in Pattern Recognition, 2009.

[11] M Abdul Rahiman, Printed Malayalam Character Recognition Using Back-propagation Neural Networks, IEEE International Advance Computing Conference, 2009.

[12] S. V. Rajashekararadhya,Zone-Based Hybrid Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Indian Scripts, IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, TX, USA - October 2009.

[13] M Abdul Rahiman, AswathyShajan, Amala, Isolated Handwritten Malayalam Character Recognition using HLH Intensity Patterns, Second International Conference on Machine Learning and Computing, 2010.

[14] Abdul Rahiman M, An HCR System for Combinational Malayalam Handwritten Characters based on HLH Patterns International Journal of Computer Applications, Page(s) 19 – 23, Volume 8– No.11, October 2010.

[15] Binu P Chacko, Pre and Post Processing Approaches in Edge Detection for Character Recognition,12th International Conference on Frontiers in Handwriting Recognition, 2010.

[16] Abdul Rahiman M, Bilingual OCR System for Printed Documents in Malayalam and English, IEEE, 2011.

[17] Bindu S Moni, Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition, IJCA Special Issue on "Computational Science - New Dimensions & Perspectives" NCCSE, 2011.

[18] Bindu S Moni, Modified Quadratic Classifier for Handwritten Malayalam Character Recognition using Run length Count, International Conference on Emerging Trends in Electrical and Computer Technology, 2011.

[19] Jomy John, Pramod K. V, KannanBalakrishnan, Offline Handwritten Malayalam Character Recognition Based on Chain Code Histogram, International Conference onEmerging Trends in Electrical and Computer Technology, 2011.

[20] Abdul Rahiman M, Recognition of Handwritten Malayalam Characters using Vertical & Horizontal Line Positional Analyzer Algorithm, IEEE,2011.

[21] R. J. Ramteke, International Journal of Computer Applications (0975 - 8887), Volume 1 – No. 18, Page(s):1-5, 2010.

[22] Ming-KueiHu ,Visual pattern recognition by moment invariants,IRE Transactions on Information Theory, Volume: 8 , Issue: 2 Page(s): 179 – 187,February 1962.

[23] Zhihu Huang; JinsongLeng, Analysis of Hu's moment invariants on image scaling and rotation ,2nd International Conference on Computer Engineering and Technology (ICCET), Volume: 7, Page(s): V7-476 - V7-480,2010.

[24] R. J. Ramteke, S. C. Mehrotra. Feature Extraction Based on Invariant Moments for Handwriting Recognition, at Proc. of 2006 IEEE International Conference on Cybernetics and Intelligent System (CIS-2006), Bangkok, Thailand, ISBN: 1-4244-0023-6, DOI:10.110, 2006.

[25] Jan Flusser.On the independence of rotation moment invariants, Pattern Recognition 33 (2000) 1405-1410, 2000.

[26] Khalid M. Hosny.Efficient Computation of Legendre Moments for Gray Level Images, International Journal of Image and Graphics, Vol. 7, No. 4 (2007) 735–747, 2007.