A novel handwritten character recognition system using gradient based features and run length count

G RAJU¹, BINDU S MONI^{2,*} and MADHU S NAIR³

¹Department of Information Technology, Kannur University, Kannur 670 567, India

²School of Computer Sciences, Mahatma Gandhi University, Kottayam 686 560, India

³Department of Computer Science, University of Kerala, Kariavattom, Thiruvananthapuram 695 581, India

e-mail: kurupgraju@rediffmail.com; bindusmony@gmail.com; madhu_s_nair2001@yahoo.com

MS received 11 February 2014; revised 29 May 2014; accepted 6 June 2014

Abstract. In this paper, we propose a novel hand written character recognition system using a combination of gradient-based features and run length count (GBF–RLC). The performance of the proposed method has been tested on Malayalam script, a South Indian language. The gradient of image is the intensity at each point, giving the direction of the largest possible increase from light to dark and the rate of change in that direction. RLC is the count of contiguous group of 1's encountered in a left to right/top to bottom scan of a character image or block of an image. Classification was carried out with a Simplified Quadratic Classifier (SQDF) and Multi Layer Perceptron (MLP). A database containing 19,800 isolated handwritten characters pertaining to 44 classes was used for the study. The feature vector is augmented by including aspect ratio, position of centroid and ratio of pixels on the vertical halves of a character image. The recognition accuracy of 99.78% was achieved with minimum computational and storage requirement.

Keywords. Gradient; RLC; handwritten character recognition; quadratic classifiers; MLP.

1. Introduction

Recognition of handwritten characters is a challenging area in pattern recognition and image processing. Offline HCR systems are matured only in few languages like English and Chinese (Nagy 2000; Plamondan & Srihari 2000; Srihari *et al* 2007; Trier *et al* 1996). Imparting human perception capabilities into the systems is the major goal of pattern recognition research. The issues faced by the handwritten character recognition systems are the similarity

*For correspondence

between characters, infinite variety of character shapes, and deformed and illegible characters. Intelligent system which can identify handwritten scripts is still an open problem for the researchers.

In offline character recognition, the character images are obtained by applying appropriate segmentation algorithm to handwritten documents images. These character images are then subjected to preprocessing and an appropriate feature extraction method is applied to form feature vectors. Feature extraction is the step to extract the discriminating features from the given patterns. The feature extraction methods for handwritten character recognition are based on two types of features: statistical and structural. The choice of features decides the efficiency of the recognition system. Some of the effective features are the different variants of direction features. These include chain code, gradient, and curvature features. Gabor transforms and statistical/structural features have also been successfully used in character recognition research (Bindu *et al* 2012). When a few promising feature extraction methods have been identified, they need to be evaluated experimentally to find the best method for the given application. Some among the feature set may mislead the classifier to a wrong classification. So the most discriminative features must be identified, which in turn does a feature reduction. The popular classifiers include statistical classifier such as Modified Quadratic Discriminant Function (MQDF), Simplified Quadratic Discriminant Function (SQDF) (Chergui *et al* 2011), Neural Classifiers such as Multi Layer Perceptron (MLP), Radial Basis Function (RBF) classifier, Polynomial classifier (PC), Learning Vector Quantization (LVQ) and Support Vector Machine (SVM).

The gradient features were used in a work (Bindu & Raju 2011b), which gave 94.96% recognition rate for 44 characters with an MQDF classifier. The experiment with RLC (Bindu & Raju 2011c) and Normalized Vector Distance, based on block centroid (NVD-BC) (Bindu & Raju 2010d) using MQDF reports a better recognition rate of 94.18% and 92%, respectively. RLC is the count of contiguous group of 1's encountered in a left to right/top to bottom scan of a character image or block of an image. The recognition rate achieved with RLC (Bindu & Raju 2011c, 2011d) and gradient features (Bindu & Raju 2011b) are found to be promising. The work (Bindu & Raju 2010c), using RLC as the feature, reported a recognition rate of 89.64% for a feature size of 18, with an MLP network. The accuracy is further enhanced to 91.16%, by incorporating two simple features — aspect ratio and a feature based on the ratio of pixels on the vertical halves of a character image. Combining these two features, and centroid of the image, with RLC and NVD-BC (Bindu & Raju 2009a) based feature, reports a recognition rate of 94.64% and 92.52%, respectively (Bindu & Raju 2011e). This work focuses on offline recognition of unconstrained, isolated handwritten Malayalam characters. RLC, gradient features, aspect ratio, centroid of the image and a feature based on the ratio of pixels on the vertical halves of a character image, together forms the feature vector. Classification was done by implementing a QDF and SQDF classifier. The results obtained with the combinations of features are compared using MLP classifier.

The rest of the paper is organized as follows. The review of previous work is given in section 2. In section 3, feature extraction method is described. The classifier is presented in section 4. The experiment and results are included in section 5. Finally, section 6 concludes the paper.

2. Review of previous work

Effective hand written character recognition systems have been developed by researchers during the past decades. A brief report of, some of the recent works done in various languages are as follows.

A state-of-the-art survey about the techniques available in the area of offline handwriting recognition (OHR) in Indian regional scripts was made in the article by Pal *et al* (2012) to discuss the advancements reported in this regard during the last few decades. Various feature extraction and classification techniques associated with the offline handwriting recognition of the regional scripts were discussed in the survey. As it is important to identify the script before the recognition step, various handwritten script identification techniques were also well-discussed in the survey.

A novel approach for recognition of unconstrained handwritten Marathi compound characters was proposed by Sushama & Shaila (2011). The recognition was carried out using multistage feature extraction and classification scheme. The average recognition rate was found to be 96.14% and 94.22%, respectively for training and testing samples with wavelet approximation features and 98.68% and 96.23%, respectively for training and testing samples with modified wavelet features. Chergui *et al* (2011) presented an off-line Multiple Classifier System (MCS) for Arabic handwriting recognition which combines two individual recognition systems based on Fuzzy ART network and RBF. The best combination ensemble has a recognition rate of 90.1%, which is significantly higher than the 84.31% achieved by the best individual classifier. Georgios *et al* (2010) proposed a methodology which relies on a feature extraction technique based on recursive sub divisions of the character image. Feature extraction is followed by a two-stage classification scheme based on the level of granularity of the feature extraction method. The authors report that the recognition rate achieved is the highest for the well-known CEDAR Character Database (94.73%) and among the best for the MNIST Database (99.03%). Nibaran *et al* (2010), presents a novel approach for recognition of handwritten compound Bangla characters, along with the Basic characters of Bangla alphabet. The approach makes an attempt to identify compound character classes from most frequently to less frequently occurred ones. On experimentation, the technique produces an average recognition rate of 79.25% using MLP and 80.510% using SVM after three fold cross validation.

The statistical approaches keep in practical applications due to their robust characteristic and simple training schemes. Omachi *et al* (2000), proposed SQDF, a new approximation method of the Quadratic Discriminant Function (QDF). Character recognition was performed using character images included in the NIST Special Database. As the feature vector, the improved directional element feature, a 196-dimensional vector was used. The error rate of SQDF with the dominant principal axes (k) as 30 was much smaller than QDF, in which, k was 196. So, the dimensionality is fixed to $k = 30$. SQDF reduces the computational time and improves classification accuracy. Junichi *et al* (2007), proposed the higher accuracy discriminant method for similar hand-written Japanese character recognition. Experiment was done on the hand-written character database ETL9B, and it was shown that the Difference Subspace-Compound Mahalanobis Function (DS-CMF) can achieve a higher recognition rate as high as 94.01%, a notable increase over that of the conventional methods.

2.1 Malayalam HCR

In Malayalam language, characters of the alphabet are rich in shape and they are subjected to many variations in terms of handwriting styles. Malayalam characters are isolated in nature; there is no cursive writing and no upper and lower case difference. Some among the works in Malayalam HCR are Raju (2006); Raju & Bindu (2009a); Raju & Bindu (2009b) and Renju *et al* (2007). Bindu & Raju (2009a) found that, fixed meshing together with measurement of NVD from block centroid yielded comparatively better performance (83.3%). An experiment with four meshing techniques for Offline recognition of handwritten isolated Malayalam characters was also conducted by Bindu & Raju (2010b). Fixed meshing, Fuzzy based meshing,

Global Elastic meshing and Local Elastic meshing were the four different meshing techniques experimented. The feature values are formed by finding NVD of blocks marked by the different meshing techniques. The above works used a general MLP classifier. From the results obtained, it is found that Fixed Meshing with Local Centroid as reference point gives the best recognition rate of 88.22%. Raju (2008), reported that, wavelet transform of projection profiles together with the aspect ratio of character images could give an overall recognition rate of 81.3%. But this work used only 33 characters and a general MLP classifier. A work based on skeleton pruning by contour partitioning with discrete curve evolution and discrete features (Binu & Babu 2009) reported a recognition accuracy of 90.18%. The features extracted are loops, end points, semi-circles in each direction, arcs in each quadrant, and horizontal and vertical crossings. During the training stage, classifier selects a unique combination of feature values to identify each character. An empirical study on the classification of thirty handwritten Malayalam characters using a multiple MLP classifiers was conducted by Bindu & Raju (2009b). Complexity of MLP network increases with number of classes. In the multiple MLP classifiers, instead of using a single network, a set of networks (MLP modules) of same structure is used. Each module is trained to recognize a subset of the character set used. The features employed for classification is based on Zoning and NVD. A database of 15000 handwritten samples for 30 classes was used for the study. It is found that the multiple MLP classifier with six modules, each trained for five characters, give best performance compared to other combinations. It gives an average recognition accuracy of 93.3% in comparison to 78.35% given by same features and a single MLP network.

A study with quadratic classifier instead of MLP, reported a recognition accuracy of 89.99% (Bindu & Raju 2010a). The features employed for classification is based on zoning and NVD. It is found that the quadratic classifier gives better performance compared to MLP. The work (Bindu & Raju 2010c), using RLC as the feature, reported a recognition rate of 91.16% for a feature size of 20, with an MLP network. Compared to other reported works in Malayalam HCR which are based on extraction of structural features as well as features based on normalized vector distance, RLC gives better recognition rate and is computationally very simple. Bindu & Raju (2010d), (2011a), (2011b), (2011c) have implemented MQDF, which is found to improve the recognition rate of Malayalam HCR. The experiment with RLC (Bindu & Raju 2011c) and NVD-BC (Bindu & Raju 2010d) using MQDF reported a better recognition rate of 94.18% and 92%, respectively. Compared to QDF, MQDF gives better recognition rate with lesser feature size. The effectiveness of gradient features was studied in another work (Bindu & Raju 2011b), which gave 94.96% recognition rate for 44 characters. The horizontal and vertical gradient along with the gradient strength are extracted to form the feature vector. From the 147 features extracted, the dimensionality is fixed as k (dominant principal axes) = 35. In another work (Bindu & Raju 2011a), the twelve directional codes depending on the gradient direction are used. QDF and MQDF classifier is used for classification, and it is found that the MQDF classifier gives better result of 95.42% with k as 118 from a total of 432 features. Compared to QDF, MQDF improves the classification performance by more than 10%, reduces the computation cost and also provides dimensionality reduction to a larger extent. A database of 19,800 handwritten Malayalam character samples was used for the experiment. Better recognition rates are obtained by implementing SQDF by Bindu & Raju (2011d). Using RLC as feature, SQDF decreases the dimensionality n to a value slightly greater than $n/4$. SQDF outperforms QDF for a small feature size ($k = 45$) from 162 features and had achieved 95.13% recognition rate. In the work by Bindu & Raju (2011e), experiment with different combinations of features is done. Classification was carried out with MQDF. Combination with RLC based feature is found to be better than with NVD-BC based feature. NVD is a distance metric and is not giving better results

compared to RLC, which extracts the edge information. Recognition rate of 94.64% and 92.52% was obtained with the two approaches, respectively.

The work by Binu & Babu (2010) presents edge detection in the preprocessing stage. Canny edge detector is used to produce thinned edges of the character. But for high resolution images, it gives spurious branches corresponding to noise and texture of the image. So a technique is applied prior to this process to focus the edge regions using nonlinear anisotropic diffusion via partial differential equations (PDE). Even though this method helped to get accurate edge points, the final result shows some broken edges. These broken parts are then linked using ant colony optimization (ACO) method. Classification was done by MLP, and a recognition accuracy of 95.16% was obtained. Recently, Binu *et al* (2012) used ELM (extreme learning machine), to overcome the slow speed of leaning algorithms which was a major hurdle and it produces only the smallest training error. The feature vector used for this pattern classification problem is based on wavelet energy. Wavelet energy feature (WEF) is defined by employing wavelets, which is a powerful tool for multi-resolution analysis. It can reflect the wavelet energy distribution of characters in several directions at different wavelet decomposition levels. WEF using daubechies wavelet at the sixth decomposition level achieved the highest recognition rate of 95.59%.

3. Gradient-based feature–run length count (GBF–RLC)

The feature extraction methods for handwritten character recognition are based on two types of features: (i) statistical and (ii) structural. The statistical features are derived from statistical distributions of pixels, such as zoning, moments, projection histograms or direction histograms. Structural features are based on topological and geometrical properties of the character, like strokes and their directions, end-points or intersection of segments and loops. To make the features more representative/discriminative, it is better to localize the extraction. Integration of structural and statistical information is a better solution to accommodate the large variations observed on the handwritten character images and highlight different character properties. Since these two types of features can be considered complementary in nature, the features representing the structural properties of different regions of the character image extracted with the aid of partial decomposition (blocking or zoning) should work efficiently for the Malayalam scripts.

3.1 Character feature

The discriminating properties of character images are not represented in a single feature. Single type of feature may highlight only specific characteristics of the image. The inherent properties of the character images are extracted by different features. To model a better recognition system, features representing different characteristics are combined in our proposed system.

3.1a Meshing technique: Features can be made more representative, by meshing the images into different zones. The vital and minute information's which discriminates the images is captured by meshing technique.

Fixed meshing. Fixed meshes are constructed by dividing an $m \times n$ image into N equal sized blocks. For example, figure 1 shows fixed meshing of an $m \times n$ image into 9 blocks of size $p \times q$.

Figure 1. Fixed meshing.

3.1b Run length count (RLC): This is a technique to represent each row/column of an image by a count of successive runs of white pixels. In a binary image, transitions from 0 to 1 and from 1 to 0 are important from the point of edge identification. Hence they significantly contribute to the object recognition process. The meshing ensures that uniqueness is found in at least one block. We represent the transition using run length count (RLC). In Horizontal Run Length Count, the image/block is scanned from left to right and the contiguous blocks of 1s in all the rows are counted. In vertical RLC, the image/block is scanned from top to bottom and the contiguous blocks of 1s in all the columns are counted. The vertical and horizontal RLCs of the blocks in an image can be taken as a feature vector. Figures 2 a and b show the horizontal and vertical RLC count for a block. RLC captures the distribution of the pixels such that the shape information inherent in the image is modelled. The variations in writing style can be addressed by RLC as exactness in count is not required.

3.1c Directional features: The gradient of image is the intensity at each point, giving the direction of the largest possible increase from light to dark and the rate of change in that direction. Gradient direction feature is suitable for application using machine print/handwriting, binary/gray scale and low-resolution images. It is the directional change in the intensity or colour

Figure 2. Run length count. (a) Horizontal RLC. (b) Vertical RLC.

in an image. The rate of change of image/pixel intensity is modelled in gradient direction feature. A gradient vector of arbitrary direction is decomposed into two components coincident with the two neighbouring standard directions (figure 3). The components are assigned to the corresponding direction planes. On decomposing all the gradient vectors, a number of feature values are extracted from each direction plane. Conventionally, the gradient is computed on each pixel of the normalized image. The gradient feature represents local characteristic of a character image properly, but it is sensitive to the deformation of handwritten character (Bindu & Raju 2011a).

In this work, the gradient feature is extracted on grayscale image. The grayscale level is standardized by mapping the minimum and maximum value to 0 and 255, respectively. To extract gradient feature, 3×3 sobel operators are used to get the horizontal and vertical gradient at each image pixel.

The Sobel operator uses two templates to compute the gradient components in horizontal and vertical directions, respectively. The templates are shown in figure 4 and two gradient components at location (i, j) are calculated by:

$$g_v(i,j) = f(i-1, j+1) + 2f(i, j+1) + f(i+1, j+1) - f(i-1, j-1) - 2f(i, j-1) - f(i+1, j-1), \quad (1)$$

$$g_h(i,j) = f(i-1, j-1) + 2f(i-1, j) + f(i-1, j+1) - f(i+1, j-1) - 2f(i+1, j) - f(i+1, j+1). \quad (2)$$

Figure 5 shows an image and its corresponding gradient images.

3.2 Additional features (AF)

Characteristics of images are represented in different types of features. Some simple features show better discriminating capabilities when combined with other effective features. Such features can be utilized for better recognition systems. The following features are used in the experiment. These features represent the local characteristics of the characters.

Figure 3. Direction decomposition of gradient vector.

-1	0	1
-2	0	2
-1	0	1

Vertical
Template

1	2	1
0	0	0
-1	-2	-1

Horizontal
Template

Figure 4. Sobel operator templates.

- (i) *Aspect ratio*: The aspect ratio of character image, cropped into minimum rectangle, is one of the simple features. It is the ratio of width to the height of an image. Coupled with others, it is found to improve the recognition accuracy. Most of the isolated Malayalam characters are greater in width. The aspect ratio of an $m \times n$ image is given by,

$$\text{Aspect ratio} = n/m. \quad (3)$$

- (ii) *Centroid*: The position of the centroid of the character image is another feature used. The centre of mass (unit value) of all the on pixels will be acting at the centroid. It is calculated as the weighted (unit value) average of all the on pixels locations. Figure 6 represents the image centroid. The coordinates (x_c, y_c) of centroid are calculated by equations (4) and (5).

$$x_c = \frac{\sum_{i=1}^n x_i}{n}, \quad (4)$$

$$y_c = \frac{\sum_{i=1}^n y_i}{n}. \quad (5)$$

- (iii) *Character code*: It is based on the ratio of number of pixels on the vertical halves of character images. For this, the cropped character image without resizing is used. Each image is divided vertically into two. The count of pixels on either halves is found. Let N_l and N_r be

Figure 5. Image and the gradient. (a) Image; (b) Horizontal gradient; (c) Vertical gradient.

Figure 6. Image centroid.

the count of pixels on the left and right side, respectively. Calculate $D = N_l - N_r$, $N = N_l + N_r$. According to the number of pixels on either side, a code is assigned based on the following:

$$\left. \begin{array}{l} \text{If } (\text{ABS}(D) \leq (0.05 * N)) \text{ then code} = 3; \\ \text{else if } (D > 0) \text{ then} \\ \quad \{ \text{if } \text{ABS}(D) \leq (0.15 * N) \text{ then code} = 2; \\ \quad \quad \text{else code} = 1; \} \\ \text{else} \\ \quad \{ \text{if } \text{ABS}(D) \leq (0.15 * N) \text{ then code} = 4; \\ \quad \quad \text{else code} = 5; \} \end{array} \right\} \quad (6)$$

The value of code obtained for a character image is used as a feature value. Figure 7 shows the assignment of code for a character image. Code 3 is assigned to the character.

The left/right concentrated characters can be identified by this approach. Equally concentrated images with a small percentage of tolerance are given the central code. Left concentrated characters are given codes less than the central code, based on the frequency. Similarly, right concentrated characters are given codes greater than the central code. We are categorizing the characters into 5 groups with five arbitrary codes. This grouping/assignment of code increases the recognition rate. The feature extraction algorithm is given in figure 8. The block diagram of the proposed method GBF-RLC is given in figure 9.

Figure 7. Extraction of character code.

1. Find the horizontal gradient g_h , and vertical gradient g_v , at each pixel.
2. Apply Fixed Meshing: Divide the $m \times n$ image into N ($p \times q$) equal sized block.
For each block, find
 - a. Horizontal run length count.
 - b. Vertical run length count.
 - c. Sum of g_h
 - d. Sum of g_v
3. Repeat Step 2 for different values of N ($p \times q$).
4. For each character image, find out the aspect ratio using eq. (3).
5. For each image in the database, find the centroid of the image using eq. (4) and (5).
6. Find out the character code using eq. (6).
7. Form the feature vector by combining the features.

Figure 8. Feature extraction algorithm.

Figure 9. Block diagram of the proposed method (GBF-RLC).

4. The classifier

The statistical classifiers typically construct generative probabilistic model or learn prototypes for each character class. Here, the parameters of one class are estimated from the samples of its own class only. The training time is linear with the number of classes and is easy to add a new class to an existing classifier. For many statistical pattern recognition methods, distributions of sample vectors are assumed to be normal, and the quadratic discriminant function (QDF) derived from the probability density function of multivariate normal distribution is used for classification. The multivariate normal distribution is usually used as the density function because it is easy to handle and in many cases the distribution of sample vectors can be regarded as normal if there are enough samples. If there are not enough training samples, accurate estimation of the covariance matrix is not possible. Also, for the higher dimensions, the estimation error increases in eigen value expansion. Since the estimation error of the covariance matrix is critical when the training samples are limited, (Kimura *et al* 1987) proposed the modified quadratic discriminant function (MQDF), which has been successful in character recognition (Liu & Ding 2005).

To achieve high recognition accuracy in statistical pattern recognition, the distribution of patterns must be precisely estimated. For the precise estimation, a great number of sample patterns are required, especially when the feature vector is high dimensional. Simplified Quadratic Discriminant Function (SQDF) can be estimated with small number of samples and it approximates the QDF (Omachi *et al* 2000).

4.1 Quadratic discriminant function

For an n - dimensional feature vector x , the QDF can be represented as follows, (Omachi *et al* 2000)

$$g(x) = (x - \mu)^t \rho^{-1} (x - \mu) + \log |\rho| \quad (7)$$

$$= \sum_{i=1}^n \frac{((x - \mu)^t \phi_i)^2}{\lambda_i} + \sum_{i=1}^n \log \lambda_i, \quad (8)$$

where μ is the mean vector and ρ is the covariance matrix of the given character class. λ_i is the i th eigen value of ρ sorted by descending order and ϕ_i is the eigenvector that corresponds to λ_i . This will be the minimum-error-rate classifier if (i) the distributions are normal, (ii) prior probabilities of all categories are equal and (iii) the parameters μ and ρ are known. Since the parameters are unknown, the sample mean vector $\hat{\mu}$ and sample covariance matrix $\hat{\rho}$ are used.

$$\hat{g}(x) = (x - \hat{\mu})^t \hat{\rho}^{-1} (x - \hat{\mu}) + \log |\hat{\rho}|, \quad (9)$$

$$= \sum_{i=1}^n \frac{((x - \hat{\mu})^t \hat{\phi}_i)^2}{\hat{\lambda}_i} + \sum_{i=1}^n \log \hat{\lambda}_i. \quad (10)$$

Here, $\hat{\lambda}_i$ is the i th eigen value of $\hat{\rho}$ and $\hat{\phi}_i$ is the eigenvector. Small eigen values in equation 10 usually contain many errors that cause the reduction of recognition accuracy. Also when n is large, it requires enormous computational cost (Omachi *et al* 2000).

4.2 Simplified quadratic discriminant function

To avoid the bad influence caused by small eigen values and thereby improve the recognition rate, the small eigen values are substituted by a constant. The approximated function of equation 10 is

$$g_s(x) = \sum_{i=1}^k \frac{((x - \hat{\mu})^t \hat{\phi}_i)^2}{\hat{\lambda}_i} + \sum_{i=k+1}^n \frac{((x - \hat{\mu})^t \hat{\phi}_i)^2}{\lambda} + \sum_{i=1}^k \log \hat{\lambda}_i + \sum_{i=k+1}^n \log \lambda. \quad (11)$$

Here, λ is the mean value of $\hat{\lambda}_i$ ($i = k+1, \dots, n$) and $k \leq n$. k denotes the number of dominant principal axes. In recognition systems which consider large number of categories with high dimensional vectors, small value of k should be chosen to limit the computational cost. It can be chosen arbitrarily or experimentally and it is fixed as $n/2$ (Omachi et al 2000). The execution time of the system can also be reduced by choosing a smaller k value.

4.3 Multi layer perceptron

MLP is the class of networks consisting of multiple layers of computational units, interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. Back propagation is used as the learning technique. The experiment was conducted using the WEKA tool. The features (36, 39, 64, 67, 144 and 147) extracted for each method are given as the inputs. The MLP function in WEKA selects the number of neurons in the hidden layer as the average of input size and number of classes. And, the sigmoid function is used as the activation function in the output layer.

5. Experimental analysis

The experiment consists of preprocessing, feature extraction, training and testing. We have carried out the feature extraction techniques described in section 3.1 for fixed meshing with different block sizes. A database consisting of 19800 character images was used for the study.

5.1 Dataset

A database containing real samples of handwritten characters with different size, shape variation, thickness, etc. is important to evaluate the performance of any HCR system. In the present study, we considered only isolated characters of vowels and consonants. Four hundred and fifty handwritten pages containing the selected 44 characters are collected from different persons belonging to different age groups, qualification, and profession. Each page is scanned at 300 DPI. The characters are segmented using the projection histogram method, cropped to fit the minimum sized window and stored as BMP images. The resultant database contains 450 samples of each of the 44 selected character classes. It includes broken and distorted characters. Samples of the 44 characters used in this study are given in figure 10.

5.2 Preprocessing

A series of preprocessing steps are applied on the raw images to obtain images suitable for feature extraction. Binarization, normalization and thinning are the preprocessing steps carried out.

Figure 10. Character samples.

Binarization. This is the first step which converts the given input grey scale or colour document into a bi-level representation. This representation is particularly convenient because most of the documents that occur in practice have one colour for text (e.g., black), and a different colour (e.g., white) for background. These colours are typically chosen to be high-contrast so that the text is easily readable by humans.

Normalization. The segmented characters can be of different size because of the variation in handwriting. For further processing we can apply size normalization on all images. For this, the height and width of all characters is fixed to a value.

Thinning. Objects in document images have line structures. To represent an object by its boundary is line representation which can simplify subsequent operations. Line representation can be obtained by a thinning process. The basic idea of thinning is to repeatedly delete object boundary pixels so as to reduce the line width to one pixel. It must be done without locally disconnecting the object or deleting line end points. Thinning is conditional deletion of boundary pixels.

The segmented character images are first binarized. Salt and pepper noise are eliminated with a median filter. Each character image is then thinned using a **Matlab** function. In few cases the thinning function failed to give satisfactory output (figure 11).

Figure 12 shows the sample image and image after the preprocessing steps.

The preprocessing steps will reflect in feature extraction and thereby in the accuracy of the system. The feature, extracted after the preprocessing steps may uniquely classify the characters.

5.3 Feature extraction

For each image in the database, the feature extraction methods are applied. For gradient feature extraction, grayscale image is standardized by mapping the minimum and maximum value to 0 and 255, respectively and is resized (72×72). To extract RLC, the segmented character images

Figure 11. Samples showing poor thinning.

Figure 12. Preprocessed images.

are first binarized. Each character image is then resized (72×72) and thinned. The aspect ratio of the image is found without applying any preprocessing steps. The binarized and thinned images are used for finding the character code, as explained in section 3.2. The centroid of the image is calculated using the binarized, resized (72×72) and thinned images. Figure 13 shows the feature extraction steps.

Figure 13. Feature extraction.

5.4 Training and testing

The training and testing was carried out with 450 samples and is divided into four data sets. Among the 450 samples, 70% (315) of are selected randomly for training and the remaining 30% (135) for testing. For the creation of next data set, another combination of the same character samples are considered, and the selected 70% training data and 30% testing data will be different from the first. The formation of the four different data sets ensures the distribution of all variations of characters across training and testing phases.

Classification was done with QDF, SQDF and MLP classifier. For classification with QDF, all the extracted features are used. To reduce the estimation error caused by the small eigen values, we replaced the small eigen values by lambda for SQDF classifier as described in section 4.2. The number of small eigen values to be replaced is found experimentally. Substitution was done for different values of k . For each feature set, we have experimented with different k values to find the best recognition rate.

5.5 Results and discussion

Classification was done with the features individually and in combination. Fixed meshing was applied to divide the image into N equal sized blocks. 72×72 images were divided into 9 (3×3), 16 (4×4), 36 (6×6), 64 (8×8) and 81 (9×9) blocks.

The value of k for each method is fixed experimentally, by examining the rate of recognition for different values of k for all the methods. For a small feature size (N), the number of small eigen values which can be replaced is also small, which results in a k value closer to N . It is seen that, when the number of blocks is increasing, the recognition rate is also increasing with a smaller value of k . So, better recognition rate is obtained with bigger values of N . But when we increase the number of blocks to a higher extent, the size of the block is reduced and hence the information content in the block will be reduced, which is resulting in lesser recognition rate. So an optimal value for N should be experimentally found.

The results with different number of blocks using QDF and SQDF with (RLC, AF) pair are given in tables 1A and B. The best recognition rate obtained with QDF- (RLC, AF) pair is 94.29% for 75 features. For higher number of blocks ($N > 36$), the accuracy is found to be decreasing. For all the methods, accuracy is increased by using SQDF, except with 9 (3×3) blocks. This is because; SQDF improves the recognition rate for higher dimensions only. From the table 1B, it is clear that 95.34% accuracy is achieved with SQDF for 165 ($k = 46$) features whereas QDF gives only 90.5%. This value of k is much less than $N/3$ and closer to $N/4$.

Table 1A. Classification performance—R LC and AF with SQDF.

Data Set	Classification accuracy (%)									
	18 F (3×3)		21 F (3×3)		32 F (4×4)		35 F (4×4)		72 F (6×6)	
	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF
	N = 18	K = 16	N = 21	K = 19	N = 32	K = 20	N = 35	K = 33	N = 72	K = 36
1	89.16	88.89	91.09	90.93	92.98	93.4	94.06	94.19	93.32	94.43
2	89.82	89.98	91.4	91.33	93.22	93.2	94.19	94.19	94.02	95.25
3	89.53	89.71	91.65	92.05	93.32	93.75	94.16	93.99	94.02	95.3
4	89.8	89.78	91.79	91.84	93.06	93.13	93.99	94.28	93.33	94.29
AVG	89.58	89.59	91.48	91.54	93.15	93.37	94.1	94.16	93.67	94.82

Table 1B. Classification performance—R LC and AF with SQDF.

Data Set	Classification accuracy (%)									
	75 F (6×6)		128 F (8×8)		131 F (8×8)		162 F (9×9)		165 F (9×9)	
	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF
	N = 75	K = 36	N = 128	K = 49	N = 131	K = 43	N = 162	K = 45	N = 165	K = 46
1	94.11	94.56	91.77	94.75	92.24	94.98	90.29	94.93	90.4	95.03
2	94.41	95.02	92.27	95.07	92.48	94.95	90.4	94.8	90.4	95.19
3	94.52	95.54	92.53	95.42	92.95	95.72	90.76	95.98	91.03	96.18
4	94.13	94.63	91.85	94.92	92.36	95	89.76	94.8	90.15	94.95
AVG	94.29	94.94	92.11	95.04	92.51	95.16	90.3	95.13	90.5	95.34

Table 2 shows the results of gradient based features, g_h , g_v and AF with SQDF. The best recognition rate obtained with QDF is 87.39% for 75 features. 92.66% is achieved with SQDF for 75($k = 13$) features. The value of k is less than even $N/5$ and more closer to $N/6$. Here also, there is considerable improvement in accuracy with SQDF, compared to QDF.

Combination of the two features, g_h and g_v with RLC is giving better results. Combining three more simple features (AF) given in section 3.2 is found to be promising in the all cases. The results are given in table 3A and 3B. 95.94% is achieved with even 9 (3×3) blocks. Here, the result with QDF is also increasing with higher number of blocks. This is because, when the features are combined, more information is included even with small block sizes. Better recognition rate is obtained by combining RLC, 3F and g_h and g_v . The best recognition rate 99.66 % is obtained with SQDF for 147 ($k = 58$) features (36 number of blocks). So among the 147 eigen values, 89 small eigen values were replaced, resulting in k to be 58. This k value is only slightly greater than $N/3$, which is much better than $N/2$ (Omachi *et al* 2000). Figure 14 shows the recognition rates of various dimensionality k of SQDF for 147 features. From the results obtained, it is found that combination of different types of features shows remarkably better performance.

In the work by Bindu & Raju (2011d), the experiment with RLC–SQDF pair, reports 95.13% accuracy for 162 features ($k = 45$). The recognition rate achieved with gradient features (Bindu & Raju 2011b) is also much promising. Combining the three simple features (section 3.2) with other features – RLC (Bindu & Raju 2010c) and NVD (Bindu & Raju 2009a), (Bindu & Raju 2011e) is also increasing the recognition rate. So, we have combined the gradient-based features and RLC along with the three simple features in this paper.

Table 2. Classification performance—Gradient and AF with SQDF.

Data set	Classification accuracy (%)											
	18 F (3×3)		21 F (3×3)		32 F (4×4)		35 F (4×4)		72 F (6×6)		75F (6×6)	
	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF	QDF	SQDF
	N = 18	K = 16	N = 21	K = 21	N = 32	K = 9	N = 35	K = 9	N = 72	K = 6	N = 75	K = 13
1	80.93	81.13	84.80	84.80	84.60	89.33	86.38	89.21	86.63	92.46	87.44	92.31
2	80.51	80.59	83.30	83.30	85.00	89.48	85.99	89.19	85.89	92.66	87.48	92.53
3	81.94	82.04	84.06	84.06	84.93	89.01	86.47	88.82	86.35	92.80	87.68	93.17
4	82.42	82.69	83.79	83.79	84.78	89.58	85.61	89.46	86.53	92.69	86.94	92.63
AVG	81.45	81.61	83.99	83.99	84.83	89.35	86.11	89.17	86.35	92.65	87.39	92.66

Table 3B. Classification performance–GBF–RLC with SQDF.

Data set	Classification accuracy (%)					
	144 F (6×6)			147 (6×6)		
	RLC -72 + Gx, Gy-72			RLC -72 + Gx,Gy-72 + 3F		
	QDF	SQDF(K)		QDF	SQDF(K)	
	N = 144	N/2 = 72	63	N = 147	N/2 = 73	58
1	98.55	99.38	99.61	98.67	99.38	99.66
2	98.89	99.71	99.76	99.01	99.7	99.73
3	98.75	99.6	99.63	98.77	99.58	99.65
4	98.92	99.5	99.56	98.77	99.51	99.6
Avg	98.78	99.55	99.64	98.81	99.54	99.66

The RLC technique does not include any complex time-consuming computations for the individual black pixels. It carries out simple addition for the individual blocks which results in lesser execution time and higher computational efficiency.

Compared to neural networks, quadratic classifiers take much time for testing. But training phase takes only very less execution time. The training time of neural networks is high, and it gets increased with higher number of features. The classification was also done using MLP classifier. The features were given individually and in combination. Four different data sets created (section 5.4) are used as training and testing files. The results with RLC and AF using MLP are given in table 4. The highest recognition rate obtained is 95.81% for 81 (9x9) blocks. Increase in accuracy is seen with increase in number of blocks.

Table 5 shows the results with gradient features and MLP classifier. Addition of three simple features are improving each method. The highest recognition rate obtained is 95.15 % for 36 (6x6) number of blocks.

Figure 14. The recognition rates of various dimensionality K for 147 features with GBF–RLC+SQDF.

Table 4. Classification performance—RLC and AF with MLP classifier.

Data set	Classification accuracy (%)									
	18 F (3×3) RLC-18	21 F (3×3) RLC-18+3F	32 F (4×4) RLC-32	35 F (4×4) RLC-32+3F	72 F (6×6) RLC-72	75 F (6×6) RLC-72+3F	128 F (8×8) RLC-128	131 F (8×8) RLC-128+3F	162 F (9×9) RLC-162	165 F (9×9) RLC-162+3F
Set 1	83.6	86.13	89.39	90.79	94.09	94.65	95.1	95.39	95.46	95.89
Set 2	84.14	85.67	88.97	90.30	94.26	94.85	94.75	95.62	94.73	95.62
Set 3	84.71	86.77	90.3	91.60	94.18	95.15	95.15	94.76	95.29	95.81
Set 4	84.24	86.58	89.81	90.83	93.57	94.85	95.37	95.88	95.05	95.99
AVG	84.17	86.29	89.64	90.88	94.03	94.88	95.09	95.41	95.13	95.81

Table 5. Classification performance—gradient and AF with MLP classifier.

Data set	Classification accuracy (%)					
	18 (3×3) Gx,Gy	21 (3×3) Gx,Gy+3F	32 (4×4) 4x4-Gx,Gy	35 (4×4) Gx,Gy+3F	72 (6×6) Gx,Gy	75 (6×6) Gx,Gy+3F
Set 1	82.96	83.06	81.90	84.5	92.31	95.14
Set 2	83.22	83.12	84.43	85.98	94.14	94.71
Set 3	82.34	84.24	83.69	86.04	93.62	95.40
Set 4	82.48	84.23	82.64	84.21	93.2	95.34
AVG	82.75	83.66	83.17	85.18	93.32	95.15

The recognition rate with smaller number of blocks is seen to be less with all the methods using single features (RLC/Gradient). But combining these simple features, it is possible to get higher recognition rate, even with smaller number of blocks. Combination of gradient features, RLC and AF with MLP classifier is given in table 6. Recognition rate of 95.08% is achieved with 9 (3x3) number of blocks, with MLP. The best recognition rate obtained is 99.78 for 147 features (36 blocks). Figure 15 shows the performance of different methods. The training and

Table 6. Classification performance—GBF-RLC with MLP classifier.

Data set	Classification accuracy (%)					
	36 F (3×3) RLC -18, Gx,Gy-18	39 F (3×3) RLC -18 + Gx,Gy-18 + 3F	64 F (4×4) RLC-32 + Gx,Gy-32	67 F (4×4) RLC-32 + Gx,Gy-32 + 3F	144 F (6×6) RLC-72 + Gx,Gy-72	147 F (6×6) RLC-72 + Gx,Gy-72 + 3F
Set 1	95.76	97.32	97.41	99.44	99.7	99.73
Set 2	94.07	97.39	98.01	98.72	99.73	99.76
Set 3	95.39	97.48	98.25	99.31	99.9	99.9
Set 4	95.08	97.71	98.03	99.14	99.7	99.73
AVG	95.08	97.48	97.93	99.15	99.76	99.78

Figure 15. Performance of different methods.

Table 7. Time (seconds) plot for the methods.

Methods	36 Features	39 Features	64 Features	67 Features	144 Features	147 Features
QDF	115	126	205	219	477	487
SQDF($k = n/2$)	112	122	198	208	462	474
SQDF(k)	114	124	196	209	456	466
MLP	1103	1560	2457	2271	5468	5660

Table 8. Classification result of GBF-RLC USING MLP classifier for data set 3.

Letter	NCCS	%	Letter	NCCS	%	Letter	NCCS	%	Letter	NCCS	%
1	135	100	12	135	100	23	135	100	34	134	99.26
2	135	100	13	135	100	24	135	100	35	135	100
3	134	99.26	14	135	100	25	135	100	36	134	99.26
4	135	100	15	135	100	26	135	100	37	135	100
5	135	100	16	135	100	27	135	100	38	135	100
6	135	100	17	135	100	28	135	100	39	135	100
7	135	100	18	135	100	29	135	100	40	135	100
8	135	100	19	134	99.26	30	135	100	41	135	100
9	134	99.26	20	135	100	31	135	100	42	135	100
10	135	100	21	135	100	32	135	100	43	135	100
11	135	100	22	134	99.26	33	135	100	44	135	100

Total number of samples: 135

NCCS – Number of correctly classified samples

testing time is recorded for all the methods. Table 7 shows the time taken for different feature sizes with the classifiers.

High recognition rate achieved by our system shows that all of the testing samples were correctly classified. The classification result of the best method with MLP classifier is shown in table 8. Among the 5940 total number of instances, 5934 instances are correctly classified. Only 6 instances are incorrectly classified. 135 samples are used for the testing of 44 classes.

6. Conclusion

A novel hand written character recognition system using a combination of gradient based features and run length count (GBF–RLC) is proposed. Gradient based features, run length count (RLC) and three more simple features—aspect ratio, centroid of the image and a character code are used in this work. SQDF and MLP are used for classification. From the results obtained, it is found that combination of different types of features is a better method to achieve higher recognition rate. The best recognition rate of 99.66% is obtained with SQDF for 147 ($k = 58$) features. Our experimental results using MLP show that, the performance of the HCR system is reaching 99.78% and is remarkable. The combination method proposed in this paper gives the best results among the reported works.

References

- Bindu S Moni and Raju 2009a Meshing and normalized vector distance from centroid for handwritten Malayalam character recognition, *Proceedings of the 2nd International Conference on Signal and Image Processing (ICSIP – 2009)*, pp 398–403, ISBN: 978-93-80043-26-5
- Bindu S Moni and Raju G 2009b Multiple MLP classifiers for handwritten Malayalam character recognition, *Proceedings of the International Conference on Mathematical and Computational Models: Recent Trends- ICMCM-09*, pp 349–354, ISBN: 978-81-8487-042-8
- Bindu S Moni and Raju G 2010a Quadratic classifier for handwritten Malayalam character recognition, *Proceedings of the U G C Sponsored National Conference on Soft Computing*, pp. 59–68, ISBN: 978-81-908520-1-2
- Bindu S Moni and Raju G 2010b Study on different meshing techniques and normalized vector distances for handwritten Malayalam character recognition. *Int. J. Eng. Res. Ind. Appl. (IJERIA)* 3(1): 181–195 ISSN 0974–1518
- Bindu S Moni and Raju G 2010c Runlength counting for handwritten Malayalam character recognition, *Proceedings of the AICTE Sponsored International Conference, ICMCM 2010*
- Bindu S Moni and Raju G 2010d Modified quadratic classifier and normalised vector distance for handwritten Malayalam character recognition, *Proceedings of the International Conference on Emerging Trends in Mathematics and Computer Applications – ICETMCA 2010*, pp 356–360, ISBN: 978-81-8424-649-0
- Bindu S Moni and Raju G 2011a Modified quadratic classifier and directional features for handwritten Malayalam character recognition. *Int. J. Comput. Appl. Spec. Issue Comput. Sci. New Dimens. Perspect.* (1): 30–34
- Bindu S Moni and Raju G 2011b Modified quadratic classifier for handwritten Malayalam character recognition using gradient features, *Proceedings of the National Conference on Indian Language Computing*
- Bindu S Moni and Raju G 2011c Modified quadratic classifier for handwritten Malayalam character recognition using run length count, *IEEE Explore*, doi:[10.1109/ICETECT.2011.5760188](https://doi.org/10.1109/ICETECT.2011.5760188)
- Bindu S Moni and Raju G 2011d Simplified quadratic classifier for handwritten Malayalam character recognition using run length count, *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, IEEE Catalog No: CFP1120J-PRT, pp 1103–1107, ISBN: 978-1-61284-766-5
- Bindu S Moni and Raju G 2011e Handwritten Malayalam character recognition using feature combinations and modified quadratic classifier, *Proceedings of the International Conference on Mathematical and Computational Models- ICMCM-11*, Narosa Publishing House, pp 424–431. ISBN: 978-81-8487-164-7
- Binu P Chacko and Babu Anto P 2009 Discrete curve evolution based skeleton pruning for character recognition, *Proceedings of the IEEE International Conference in Pattern Recognition* pp 402–405
- Binu P Chacko and Babu Anto P 2010 Pre and post processing approaches in edge detection for character recognition, *ICFHR '10 Proceedings of the 2010 International Conference on Frontiers in Handwriting Recognition*, pp 676–681
- Binu P Chacko, Vimal Krishnan V R, Raju G and Babu Anto P 2012 Handwritten character recognition using wavelet energy and extreme learning machine. *Int. J. Mach. Learn. Cybern.* 3(2): 149–161
- Chergui Leila, Kef Maâmar and Chikhi Salim 2011 Combining neural networks for Arabic handwriting recognition, *Programming and Systems (ISPS), 2011 10th International Symposium on*, pp.74–79
- Georgios Vamvakas, Basilis Gatos, Stavros J and Perantonis 2010 Handwritten character recognition through two-stage fore ground sub-sampling. *Pattern Recog.* 43(2010): 2807–2816
- Junichi Hirayama, Hidehisa Nakayama and Nei Kato 2007 A classifier of similar characters using compound Mahalanobis function based on difference subspace, *Ninth International Conference on Document Analysis and Recognition (ICDAR:2007)*, Vol. 1, pp. 432–436
- Kimura F, Takashina K, Tsuruoka S and Miyake Y 1987 Modified quadratic discriminant functions and its application to Chinese character recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(1): 149-153

- Liu Hailong and Xiaoqing Ding 2005 Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes, *Proceedings of the 2005 Eight IEEE International Conference on Document Analysis and Recognition (ICDAR'05)*, Vol. 1, pp 19–23
- Nagy G 2000 Twenty years of document analysis in PAMI. *IEEE Trans. PAMI* 22(1): 38–61
- Nibaran Das, Brindaban Das, Ram Sarkar, Subhadip Basu, Mahantapas Kundu and Mita Nasipuri 2010 Handwritten Bangla basic and compound character recognition using MLP and SVM classifier. *J. Comput.* 2(2): 109–115 ISSN 2151–9617
- Omachi Shin'ichiro, Fang Sun and Hiroto Aso 2000 A new approximation method of the quadratic discriminant function, *Advances in Pattern Recognition, Joint IAPR International Workshops SSPR/SPR 2000*, 601–610 Proceedings, 1876 Springer 2000
- Pal Umapada, Ramachandran Jayadevan and Nabin Sharma 2012 Handwriting recognition in Indian regional scripts: A survey of offline techniques. *ACM Trans. Asian Lang Inf. Proc.* 11(1): 1–35
- Plamondon R and Srihari S N 2000 Online and offline handwriting recognition: A comprehensive survey. *IEEE Trans. PAMI* 22(1): 63–84
- Raju G 2006 Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients, *Proceedings of the 14th International Conference on Advanced Computing and Communications*, pp 217–221
- Raju G 2008 Wavelet transform and projection profiles in handwritten character recognition – A performance analysis, *Proceedings of the 16th International Conference on Advanced Computing and Communication (ADCOM)**, pp 309–313
- Raju G and Bindu S Moni 2009a Global elastic meshing for handwritten Malayalam character recognition, *Proceedings of the National Conference on Computational Science and Engineering - NCCSE*, pp 10–14, ISBN: 978-81-907044-0-3
- Raju G and Bindu S Moni 2009b Global and local elastic meshing for handwritten Malayalam character recognition. *Int. J. Comput. Inf. Technol. Eng.* 3(1): 149–153 ISSN: 0973-743X
- Renju John, Raju G and Guru D S 2007 1D wavelet transform of projection profiles for isolated handwritten character recognition, *Proc. ICCIMA07* 2: 481–485
- Srihari S N, Yang X and Ball G R 2007 Offline Chinese handwriting recognition: An assessment of current technology. *Front. Comput. Sci. China* 1(2): 137–155
- Sushama Shelke and Shaila Apte 2011 A multistage handwritten Marathi compound character recognition scheme using neural networks and wavelet features. *Int. J. Signal Proc. Image Proc. Pattern Recog.* 4(1)
- Trier D, Jain A K and Taxt T 1996 Feature extraction methods for character recognition – A survey. *Pattern Recog.* 29(4): 641–662