

# A Wavelet Based Recognition System for Printed Malayalam Characters

M. Abdul Rahiman<sup>1</sup>, M. S. Rajasree<sup>2</sup>

<sup>1</sup> Asst Professor, Department of Computer Science & Engg  
LBS Institute of Technology for Women, Trivandrum, Kerala, India  
Research Scholar, Karpagam University, Coimbatore, India  
E-mail:rehman\_paika@yahoo.com

<sup>2</sup>Professor in Computer Science & Engg and Principal  
LBS Institute of Technology for Women, Trivandrum, Kerala, India

**Abstract**— This paper specifies an OCR system for printed Malayalam characters. Malayalam is the principal language of the South Indian state Kerala. It belongs to the family of Dravidian Language. The input to the system would be the scanned image of a page of text and the output is a machine editable file. Malayalam Character recognition is a complex task because of the presence of two scripts; old script and new script and a lot of combinational characters. Initially, the image is preprocessed to remove noise. Then skew correction methods are applied to the document. Lines, words and characters are segmented from the processed document image. The proposed method uses wavelet analysis for extracting features of the image and Back propagation neural network is used to accomplish the recognition tasks.

**Index Terms**— Optical Character Recognition, Neural Networks, Feature extraction, Segmentation, Wavelet, Malayalam Character.

## I. INTRODUCTION

The aim of Document Image Analysis is to process the image of a printed page containing characters and render the information into a suitable form for modification and manipulation on a system. Optical Character recognition (OCR) is the process of translating the image of typewritten or handwritten in to a machine editable text. Today efficient and inexpensive OCR packages are available to recognize printed texts in English, Chinese and Japanese etc. Relatively less work is reported for the recognition of Indian languages especially in Malayalam. In this paper we describe a system that can handle printed text documents in Malayalam, which is the official language of the South Indian State, Kerala. The presence of two different scripts, Old script and New script makes Malayalam OCR a difficult one. Moreover there exist a lot of combinational characters which are also a major constraint in developing the Malayalam OCR system. This system recognizes only the printed Malayalam characters. An attempt for developing manuscript recognition and other types such as palm leaf, stylized writing, and typed fonts may be dealt separately. The input to this system is an image of printed page of Malayalam text. This system produces an output which is

an editable computer file containing the information on that document. Due to the tremendous need for digitization of printed documents in recent days, OCR has received considerable attention. The important practical applications of OCRs are seen in converting documents into electronic

format, library catalogue, postal mail system, bank cheques and reading aid for the blind.

## II. MALAYALAM SCRIPT

The Dravidian Language Malayalam, which is one of the 22 official languages of India, is being used by around 36 million people, predominantly in Kerala, the Southern State of India. Malayalam is also widely used in the union territories of Lakshadweep Island, the west coast of India and Mahe. The term Malayalam comes from the words mala (Mountain) and alam (Place) and the word Malayali stands for the people of Kerala State which means Mountain people who lived beyond the Western Ghats and Malayalam is the language that was spoken there. Malayalam first appeared in writing in the vazhappalli inscription which dates from about 830 AD. In the early thirteenth century the Malayalam script began to develop from a script known as vattezhuthu or round writing, a descendant of the Brahmi script. In the early ninth century vattezhuthu is traceable through the Grantha script, to the pan-Indian Brahmi script, gave rise to the Malayalam writing system. It is syllabic in the sense that the sequence of graphic elements means that syllables have to be read as units, though in this system the elements representing individual vowels and consonants are for the most part readily identifiable. In the 411960s Malayalam dispensed with many special letters representing less frequent conjunct consonants and combinations of the vowel with different consonants.

### A. The Character Set of Malayalam

Malayalam language script consists of 51 letters including 15 vowels and 36 consonants. It also consists of 3 left vowel signs, 7 right vowel signs and 2 left and right vowel signs. These vowel signs are known as Dependant

vowels as they do not stand on their own. They got an existence only when combines with a consonant or conjunct. The complete Malayalam alphabet set, vowels, vowel signs and consonants are shown in figure 1.



(a)



(b)

Vowel Signs	Left or Right of the consonant/ conjunct
ഓ	Right
ഐ	Right
ഐ	Right
ഐ	Right
ഐ	Right
ഐ	Left
ഐ	Left
ഐ	Left
ഐ ഓ	Left & Right
ഐ ഓ	Left & Right
ഐ	Right

(c)

Figure 1: Malayalam alphabet set (a) vowels (b) consonants, (c) Vowel signs.

The earlier style of writing is now substituted with a new style from 1981. This new script reduces the different letters for typeset from 900 to fewer than 90. This was mainly done to include Malayalam in the keyboards of typewriters and computers. The important features of this script are listed below.

- Malayalam is a syllabic alphabet in which all consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel.
- When they appear the beginning of a syllable, vowels are written as independent letters.

- When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter.

### III. THE PROPOSED METHOD

The system involves four phases, scanning of the image, pre-processing, feature extraction [1] and recognition. In this method wavelet [2] analysis is used for extracting features of the image and Back propagation neural network is used to accomplish the recognition tasks.

#### A. Scanning of Image

Scanning of the text which converts the paper document into an image is the starting stage. The document is scanned by any standard scanners with a minimum of resolution 200 dpi. Lower resolution results in poor performance of the system and misidentifications. A scanned image will be any one of the format jpeg, bmp or tiff. This image is processed in many stages.

#### B. Preprocessing

Preprocessing steps are required for any kind of OCR system before the actual recognition. This is needed because of the existence of a lot of noise [2] with the scanned image. The different steps and type of preprocessing algorithms depend on many factors such as paper quality, resolution of the scanned image, age of the document, the amount of skew in the image and the layout of text etc. The scanned image contains noise, and these should be preprocessed to remove them from the image. Then skew correction [3] of the image is done. The characters should be individually extracted from the original image. The preprocessing stage in character recognition consists of Removal of noise if any, Binarization [4] of the Image, Separation of words, Identification of line and space in the passage and Resizing to a standard size.

#### C. Noise Removal

Even if we take extra care, the image obtained after scanning may contain noise. This is due to the poor qualities of scanner, printer, age of the document etc. Hence it is needed to filter out these noises from the image for better recognition results. Filtering is applied here to remove the noise. There can be Gaussian noise and Salt & Pepper [2] noise. A suitable filter is designed to remove the noise while retaining the entire signal in an image. Image filtering removes the Gaussian noise. Even though salt and pepper noise can be removed, their effect

will be there throughout the process. Filtering done here is by spatial domain (by using filter masks). Frequency domain filtering leads to the loss of data when reconstructing. In the filtering stage, a spatial domain filter is used for the removal of noise. The isolated and dilated points in the image are removed, where a clear image is obtained after filtering.

#### D. Binarization

Binarization [4] is used to convert the gray scale image into binary images. It separates the foreground and background information. This is an essential part as we have to identify the objects of interest from other part of the

image. The most common method employed in binarization is to select a threshold for the intensity of the image and then convert all the intensity values above the threshold to one intensity value and all intensity values below the threshold value to the other chosen intensity. Binarization is performed either locally or globally. Local or adaptive thresholding method uses different intensity values to different portions of the image and Global thresholding uses a common intensity value to the entire image. A lot of binarization techniques are available In a binary image, only two levels will be there which are 0 and 1. They are also called as logical images. A binary image with its pixel values is shown in figure 2. Here the value 0 corresponds to black pixel and 1 corresponds to a white pixel.

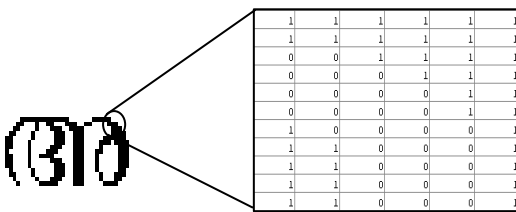


Figure 2: Binary Image.

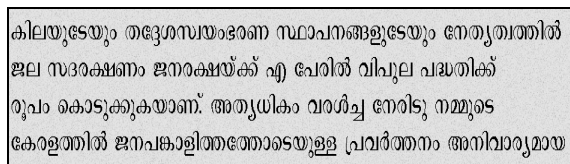
In the pre-processing stage the scanned image obtained is converted to binary image for easy analysis and to get the information about the lines and spaces between the characters. To convert a grayscale image to binary, a threshold level [2] of the grayscale image is found out. This threshold level corresponds to the gray level threshold of the grayscale image. A hard limiter is set with this gray level threshold.

$$B(x, y) = 1 ; \text{ if } I(x, y) > \text{threshold,} \quad (1)$$

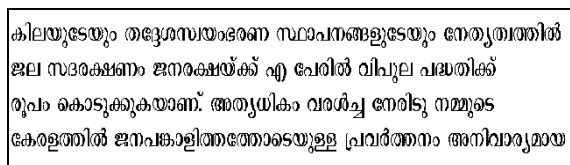
$$0 ; \text{ otherwise;}$$

where  $B(x, y)$  is the binarized image &  $I(x, y)$  is the input grayscale image.

Figure 3 illustrates scanned image of a Malayalam document. The second part of the picture is the binarized image in which the pixels are separated from the background. A lot of binarization techniques are available.



(a)

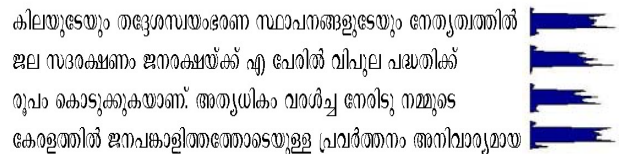


(b)

Figure 3: (a) Scanned Image, (b) Binarized Image.

IV. LINE, WORD AND CHARACTER SEGMENTATION

Characters can be separated by vertical and horizontal scanning [5] of the acquired image. The horizontal scan is done for the separation of lines from a document. The horizontal projection will have separated peaks and valleys if the lines are well separated. These peaks are easily detected and this is used to find out the boundaries between lines and so each line is extracted. The scanning starts from the top by vertical path and finds the gap present and then separates the pixels above the gap and vertical scanning is done. The line separation is illustrated in figure 4.



Word separation is the next step in the preprocessing. Vertical projection profile is applied to the Lines separated. The vertical scanning involves the scanning of image from horizontal scanning process and then determines the gap between the words and separates it. The peaks and valleys are observed and words are separated by looking at the minima in the vertical profile. In order to identify the words from individual character a threshold is set for the gap between two consecutive words. For example the space between the words is greater than 6 pixels. Figure 5 illustrates the segmentation of words. After the word separation individual characters are separated by the same method.



Figure 5: Word Segmentation.

A. Combinational Letters in Malayalam

Horizontal and vertical scanning only separates the characters by identifying the gap between them as explained earlier. But in a Malayalam character set there are different combinational letters that cannot be separated by this process. One such combination can be seen at the starting of figure 6 (first two symbols), which is pronounced as 'ki', in Malayalam, where there is no gap between the two letters. A separate algorithm for this should be developed to separate these types. One of the best methods for the separation is labeling the image. Labeling is done by the four and eight connected pixels of an image. The connected component is done only for a binary image with pixel varies form '0' and '1'. The labeling process will be done for the connected components for the value '1'. So the image is converted to complement image, where the back ground will be black and foreground will be white.

A four connected components in an image  $f(x, y)$  is four pixels in the adjacent boundaries of a pixel  $(i, j)$ . i.e.  $(i-1, j)$ ,  $(i, j+1)$ ,  $(i+1, j)$ ,  $(i, j-1)$ . Similarly for the pixel 8-connected component is defined as  $(i-1, j)$ ,  $(i-1, j+1)$ ,  $(i, j+1)$ ,  $(i+1, j+1)$ ,

(i+1,j), (i+1,j-1),(i,j-1),(i-1,j-1). Figure 6 shows the labeled image. Here the first letter is labeled as '1' and the second letter is labeled as '2'.

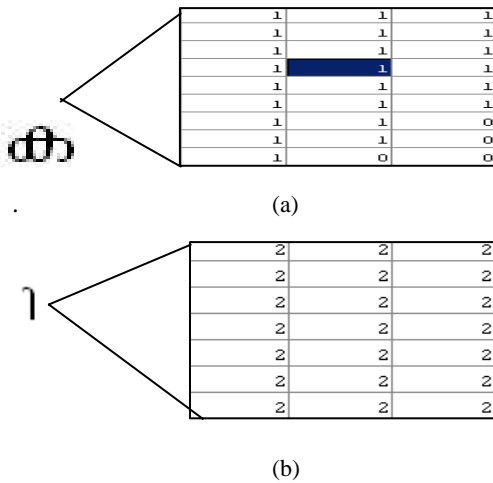


Figure 6: Labeled Images (a) labeled as 1 (b) labeled as 2.

The character should be in a standard form for the purpose of recognition. Therefore it should be resized [6] to a standard size. This resizing will lead to font invariant recognition. Bilinear Interpolation technique is used here. It maps each pixel in the new image back to a point in the old image where the four nearest pixels to this point are used to assign the value to the new pixel. These four pixels are used in a calculation that is linear in x and linear in y but not linear in both together, in other words, it is bilinear.

$$g(r,c) = (1-a)(1-b)f(m,n)+a(1-b)f(m+1,n)+(1-a) * bf(m,n+1)+abf(m+1,n+1) \quad (2)$$

In this formula g(r,c) is a pixel in the new image, m and n are row and column coordinates in the old image where m is the integer of the backward mapping of x (x=r\*(rows in old)/(rows in new)) and n is the integer of the backward mapping y (y=c\*(columns in old)/(columns in new)), a=x-m and b=y-n.

$$f(x,y) = A \cdot x + B \cdot y + C \cdot x \cdot y + D. \quad (3)$$

The characters are resized to a standard size of 32 x 64, by the above-mentioned resizing algorithm. Resizing the image will obtain font invariance during the recognition process.

### V. FEATURE EXTRACTION

For the extraction of the features [1] from the character the Daubechies (db4) wavelet [2] is used. The wavelet transform of a signal s is the family C(a,b), which depends on two indices a and b. The wavelet decomposition consists of calculating a "resemblance index" between the signal and the wavelet located at position b and of scale a. If the index is large, the resemblance is strong, otherwise it is slight. The indexes C(a,b) are called coefficients

$$C(a,b) = \int_{\mathbb{R}} s(t) \cdot (1/\sqrt{a}) \cdot \Psi((t-b)/dt). \quad (4)$$

$$\text{Where } a = 2^j, \quad b = k \cdot 2^j, \quad (j, k) \in \mathbb{Z}^2.$$

#### A. Details and Approximations

Let us fix j and sum on k. A detail Dj is nothing more than the function

$$D_j(t) = \sum_{k \in \mathbb{Z}} C(j, k) \Psi_{j,k}(t). \quad (5)$$

The details have just been defined. Take a reference level called J. There are two sorts of details. Those associated with indices j <= J correspond to the scales a = 2^j <= 2^J which are the fine details. The others, which correspond to j > J, are the coarser details. We group these details into A\_j = \sum\_{(j>J)} Dj, which defines what is called an approximation of the signal s. The equality s = A\_j + \sum\_{j>J} Dj, signifies that s is the sum of its approximation A\_j and of its fine details. This can be easily implemented using the filter [2] designed in the figure 7. The four filter coefficients are approximation coefficient (LL), vertical detail (LH), horizontal detail (HL) and diagonal detail (HH).

The LH channel represents information of a low horizontal and high vertical frequency. Since the HH channel contains most of the image noise, it is discarded at each decomposition level. The horizontal and vertical components at each frequency are combined to get the rotation invariants.

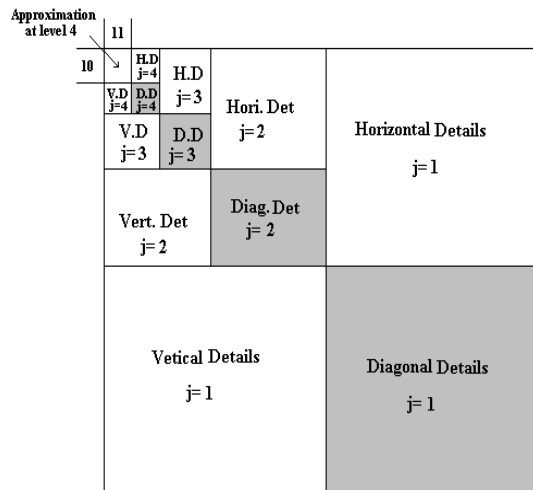


Figure 7: Conceptual diagram of Feature extraction.

For the 32 X 64 character image obtained from the preprocessing stage, wavelet transform is applied four times in order to get the 21 X 13 sub images. Figure 8 shows the wavelet analysis of Malayalam letter "Ka". Finally, a feature vector is organized by combining 100 features in the approximation detail sub image in level 2. The second level approximation coefficients are extracted. The figure 8 represents the feature extracted coefficients at the 2<sup>nd</sup> level of decomposition. These coefficients are further used for the identification process. Wavelet analysis of an image using the above method is illustrated in figure 8, which is the image of a Malayalam character "Ka".

Figure 9: Feed Forward Back Propagation Network

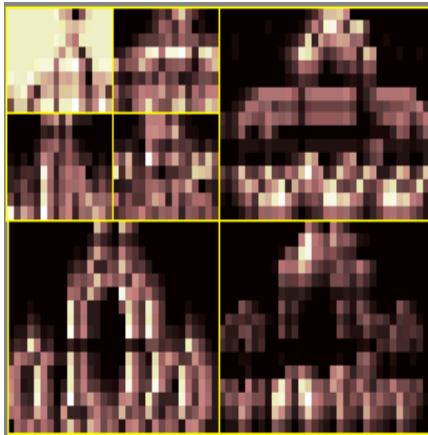
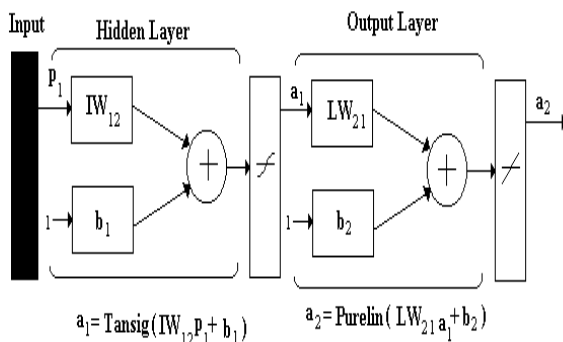


Figure 8: Wavelet analysis of Malayalam character “Ka”

### VI. RECOGNITION

Recognition is accomplished by using neural networks [7]. The most commonly used family of neural networks for pattern classification tasks is the feed-forward network, which includes multi layer perceptron and Radial-Basis Function (RBF) networks. These networks are organized into layers and have unidirectional connections between the layers. Another popular network is the Self-Organizing Map (SOM), or Kohonen-Network, which is mainly used for data clustering and feature mapping. The learning process involves updating network architecture and connection weights so that a network can efficiently perform a specific classification/clustering task. The increasing popularity of neural network models to solve pattern recognition problems has been primarily due to their seemingly low dependence on domain-specific knowledge (relative to model-based and rule-based approaches) and due to the availability of efficient learning algorithms for practitioners to use.

Having localized and coded acquired image that corresponds to the character, the final task is to decide if this character code matches a previously stored character code. Here we are using a feed forward back propagation neural network for that purpose. In general, competitive learning neural networks are used for fast learning mechanism. But their performance is easily affected by initial weight vectors. A feed forward back propagation network, shown in figure 9 overcomes this difficulty by automatically initializing the weight vectors.



The process of algorithm is shown below which is depicted in figure 10.

Step I : Setup the network by initializing the input from the feature extracted.

Step II : The weighted input vector is added with the bias and applied to the output layer through the *tansig* function.

Step III: The output layer input (weighted) is again added with the bias and passed through the *purelin* for obtaining the output.

Step IV: The output vector is compared with the target vector and the mean square error (mse) is calculated.

Step V : The weights are adjusted in order to reduce the mse.

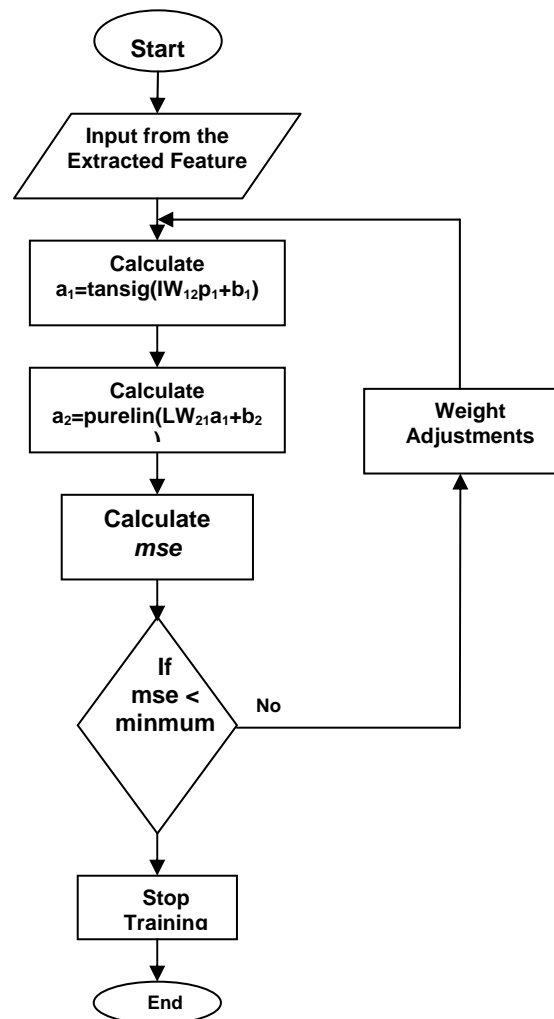


Figure 10: Flow chart of Pattern Training.

VII. TESTING & RESULT ANALYSIS

For the experimental purpose a total of 715 images were analyzed. Out of these the images with noise and without noise accounts to 315 and 400 respectively. The details are shown in Table 1.

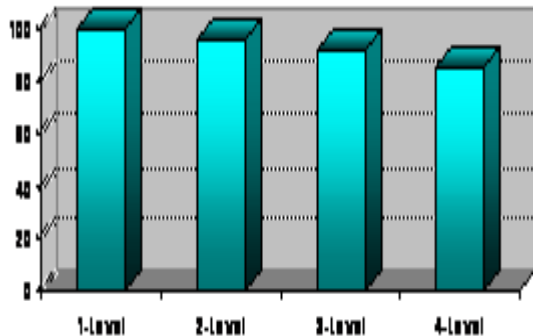


Figure 11: Identification without noise

TABLE 1

RESULT ANALYSIS OF THE PROCESS

Set No. wrt to size	No. of characters experimented	No. of noisy characters experimented	
		Gaussian Noise characters	Salt & pepper noise characters
1	90	40	30
2	70	55	66
3	85	33	47
4	65	42	57

Experimental result analyses with different levels of wavelet filters in different environments are tested. Figure 11 illustrates the identification at different levels without noise and that of the figure 12 shows the misidentification with salt & pepper noise.

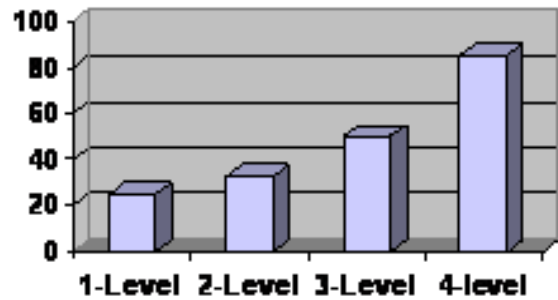


Figure 12: Misidentification with salt & pepper noise.

The detailed experiments are shown in the Table 2, which shows the results in noisy and noiseless environments. The different character in Malayalam character set is experimented. Vowels and consonants with Gaussian and salt & pepper noise are tested.

TABLE 2.

DETAILED RESULT ANALYSIS OF THE CHARACTERS WITH NOISY AND NOISELESS ENVIRONMENT

Levels	Noiseless Environment (%)		Noisy Environment (%)			
	Vowels (V)	Consonants (C)	Gaussian Noise		Salt & pepper noise	
			(V)	(C)	(V)	(C)
1st	92.8	95.8	90.5	88.5	82.5	77.2
2nd	85.6	92.3	82.5	80.5	77.5	68.9
3rd	80.8	85.4	75.1	71.4	66.4	60.2
4th	75.4	80.2	65.2	62.8	51.2	49.7

VIII. CONCLUSION

This paper describes an OCR system for recognition of printed text in Malayalam, the official language of Kerala, a south Indian State. The image of the page of text to be recognized as given as input to the system. The system separates it into lines, words and then characters. The features of the characters are extracted and classified using wavelets. For the experimental purpose a total of 715 images were analyzed. Out of these the images with noise and without noise accounts to 315 and 400 respectively. The system has been found to give an accuracy of 92 to 94%. The performance of this system can be further improved by using efficient preprocessing algorithms.

REFERENCES

[1] Anil K.Jain,O.D.Trier and T.Taxt,“Feature extraction methods for character recognition—a survey”, Pattern Recognition, Vol 29, pp 641-662.

[2] Anil K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, 1989.  
 [3] B.B. Chaudhuri and U. Pal, “Skew Angle Detection of Digitized Indian Script Document”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, February 1997.  
 [4] B. Anuradha and B. Koteswarra,“An efficient Binarization technique for old documents”, Proc.of International conference on Systemics,Cybernetics and Informatics,Hyderabad, 2006, pp 771-775.  
 [5] B.B Chaudhuri, U Pal and M Mitra, “Automatic recognition of printed Oriya script”,Sadhana, Vol. 27, Part 1, Feb 2002, pp. 23-34 .  
 [6] C.V. Lakshmi and C. Patvardhan, “Optical Character Recognition of Basic Symbols inTelugu”, IE(I)Journal-CP, 2003, Vol 84,pp.66-71.  
 [7] K. Pujari Arun, C Dhanunjaya Naidu and B C Jinaga, “An Adaptive character recognizer for Telugu scripts using Multiresolution Analysis and Associative memory”, ICVGIP, Ahmadabad, 2002.