ANALYSIS OF TELUGU PALM LEAF CHARACTERS USING MULTI-LEVEL RECOGNITION APPROACH

Panyam Narahari Sastry[1], T. R. Vijaya Lakshmi[2], N. V. Koteswara Rao[3] and Rama Krishnan Krishnan[4]
[1,3]Department of Electronics and Communication Engineering, CBIT, Hyderabad, India
[2]Department of Electronics and Communication Engineering, MGIT, Hyderabad, India
[4]ADRIN, Department of Space, ISRO, IIST, Trivandrum, India
E-mail: ananditahari@yahoo.com

ABSTRACT

Palm leaf character recognition is an area which is at the nascent stage of research. Although character recognition is a well-known application of pattern recognition, lot of work is still to be exploited in handwritten character recognition. The recognition accuracy as per the literature survey for handwritten English characters is very low and for Indian languages it is just started. Research has been started for Indian languages like Bangla, Hindi, Telugu, Tamil, Devangari, etc., but still at the starting stage. Palm leaf character recognition is an open area of research and is also very important since these palm leaves contain huge amount of information related to astronomy, astrology, architecture, law, medicine and music. In the present work, an additional feature called depth of indentation at important pixel points like the starting point, curves, joints, loops and end points is considered which is directly proportional to the pressure applied by the scriber on the palm leaf. This depth of indentation is considered in the Z-direction measuring in microns. In the proposed work, multistage recognition approach is used to improve the recognition accuracy up to 92.8%.

Keywords: Optical character recognition, palm leaf character recognition, 3D feature, 2-D FFT, 2-D DCT, multi-level recognition system, nearest neighbourhood classifier, pattern recognition.

1. INTRODUCTION

Machine recognition involves the ability of a computer to receive input from sources such as paper and other documents, photographs, touch screens and other devices, which is an ongoing research area. Optical Character Recognition (OCR) is a technology that uses an optical mechanism which allows a machine to recognize the characters automatically. The history of handwriting recognition systems is not complete without mentioning the (OCR) systems which preceded them.

The extensive applications of OCR today are banks, post offices, defence organizations, reading / visual aids for the blind, library automation, language processing and multi-media design. With the introduction of higher computer speeds, scanners are being developed which use the concept of OCR. In the future, OCR is expected to become more powerful and less expensive. Similarly, the need for improvement in handwritten character recognition will also increase.

Handwritten data is also characterized by the writing style of users [1]. The first style to be considered is the block or isolated style. In this style, symbols, letters and words are clearly separated by boxes, used as guides, or by leaving a distinguishable space between them. These characteristics correspond to the boxed and spaced styles respectively. In contrast, free style allows more freedom on writing. Symbols may overlap, share strokes, or the writer can even mix these styles.

The non existence of standard / benchmark databases is the biggest difficulty to do a research on handwritten character recognition of Indian scripts [2]. Small databases collected in laboratory environments have been reported in previous studies [2].

The progress of character recognition in Asian and particularly Indian scripts is in a relatively nascent stage as compared to English, which is in a mature stage of development. Telugu is one of the prominent scripts in India with more than 65 million worldwide speakers [3]. There are 18 vowels, 36 Consonants, and three dual symbols in this language. In Telugu, consonants take modified shapes when attached with the vowels. Additionally, vertical extent of the character varies depending on the modifying vowel or consonant. Such characters are even more difficult for a machine to recognize.

In almost all the Indian regional scripts there are compound characters apart from numerals, consonants and vowels [4]. Each compound character is the combination of two or more basic characters. There are several problems in recognizing the handwritten characters, because of the difference in writing style, size and shape of the characters which vary from person to person.

Palm leaf manuscripts contain religious texts and treaties on a host of subjects [5] such as art, medicine, astronomy, astrology, mathematics, law and music. The main causes of deterioration are climatic factors, light and insects [5]. In addition, constant handling and adverse storage conditions act as forces of deterioration. Hence there is a need for character recognition of palm leaf characters.

ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

The related work is described in section 2. The collection of characters from palm leaf is explained in section 3. The steps involved in the proposed model are described in section 4. The experimental results are discussed in section 5. Finally the paper is concluded in section 6.

2. EARLIER RELATED WORK

B. B Chaudhuri and U Bhattacharya [2] have published that a major obstacle to research on handwritten character recognition of Indian Scripts is the nonexistence of standard / benchmark databases. They further state that previous studies were reported on the basis of small databases collected in the laboratory environment.

Lakshmi Vasantha and others [6] have worked on isolated handwritten Devnagari numerals and obtained best accuracy of 94.25% using PCA. Dataset consisted of 9800 handwritten numerals out of which 9000 were in database whereas 800 were used for testing. It is reported that the recognition accuracy increased from 93.1% to 94.25 % using PCA and NNC.

Aradhya, Hemantha Kumar and Noushath [7] have worked on OCR systems for South Indian scripts (Kannada Telugu Tamil and Malayalam) and English documents based on Principal Component Analysis (PCA) and Fourier Transform (FT). It is reported that with 50 training samples / character class, the best recognition accuracy is 57% for PCA and 59% for Fourier PCA (F-PCA). It is further reported that when the number of training samples were increased to 175, the recognition accuracy for English is 95.1%. The accuracy obtained was 90.8 and 93.8 for PCA and F-PCA, respectively.

Lakshmi, Patvardhan and Prasad [3] gave a novel approach for improving recognition accuracy and resolved confusion pair of printed Telugu characters like Va, Pa, Na, Ya etc. the character is divided into two halves horizontally. To differentiate Pa and Sa the bottom half is considered. If there is a loop it is recognized as Pa, otherwise it is Sa. One more algorithm depending on zero crossings is proposed to remove the confusions between Ra and La. It is reported that improvement from 98 to 99 % in recognition for using neural networks is possible. KNN and ANN also gave approximately 1 % increase in recognition accuracy.

K. Pujari and others [8] have used wavelet multi resolution analysis and associative memory for Telugu scripts character recognition. This system learns the style and font from the document itself and recognizes remaining characters by itself. It has a Hopfield-based Dynamic Neural network (DNN) for the purpose of learning and recognition. Different fonts of the same text were tested. The recognition accuracy ranged from 85 % to 92.1 %.

Ashwin and Sastry [9] have proposed a font and size independent OCR system for printed Kannada documents using Support Vector Machines (SVM). The input to the system is a scanned image of a page text and the output is a machine editable file. In this system the word is extracted first and then it is further segmented into characters. The Zerike, structural and the modified structural features were used and the percentage accuracy for NNC and SVM were tabulated. It ranged from 88.18% to 94.9%.

Chaudhuri, Pal and Sinha [10] proposed separation of many languages using peak and valley positions. This helps in feeding the separated languages text to their respective OCR's. The feature selection was based on water reservoir principle, contour tracing, profile etc. Automatic recognition of text line of different Indian scripts was possible with an overall accuracy of about 97.52 %. This scheme does not depend on the size of the characters in the text line.

Negi and Chereddi [11] worked on printed Telugu character recognition using the candidate search and elimination approach. The initial candidates for recognition are found by applying the zoning method on input glyphs. Cavities present in Telugu script are used to prune the candidates found by zoning. Template matching is used for recognition. The images scanned from Telugu literature gave a recognition accuracy of 97 to 98 %.

Chakravarthy Bhagavati and others [12] have listed number of factors that are important in achieving high recognition accuracy in Telugu and other Indian scripts. It is reported that to obtain an accuracy of 85 – 93 % is relatively easy and to increase the accuracy further is a difficult task. In the proposed work nearly 97% accuracy is reported.

Sastry, Krishnan and Sankar Ram [13], [14] have worked on the three dimensional dataset for Telugu character recognition on Palm leaves using PCA and 2D correlation. The reported accuracy for these palm leaf Telugu characters using PCA along with minimum Euclidean distance concept is 36% in 'XY' plane of projection. It is further reported that using the depth information (Z) in the 'YZ' plane of projection and 2D correlation concept 84% of recognition rate was obtained. Although some of the characters like Va, Ma, Pa, Ya are very similar and create confusion in 'XY' plane, this proposed method successfully could remove this confusion to distinguish the characters.

3. METHODOLOGY

A. Data acquisition

The characters from the palm leaf are extracted using a digital measuroscope and a dial indicator plunger [13], [14], [15], [16] and [17]. The (X, Y) coordinates are measured using a digital measuroscope and the depth information, the 'Z' coordinate, is measured using a dial indicator plunger [13], [14], [15], [16] and [17]. These (X, Y, Z) coordinates are used to obtain the characters with the help of Microsoft Excel and are stored in the computer.

ARPN Journal of Engineering and Applied Sciences

Using the combination of two coordinates at a time the characters are obtained i.e. 'XY', 'YZ' and 'XZ' planes of projection [13]-[17]. Further the stored characters are processed for recognition. The number of characters taken into consideration in the proposed work is 28. The number of samples used for training are 1232 i.e. 44 per character in each plane of projection. For testing 308 samples are used in each plane of projection. The data collection process is described in Figure-1.

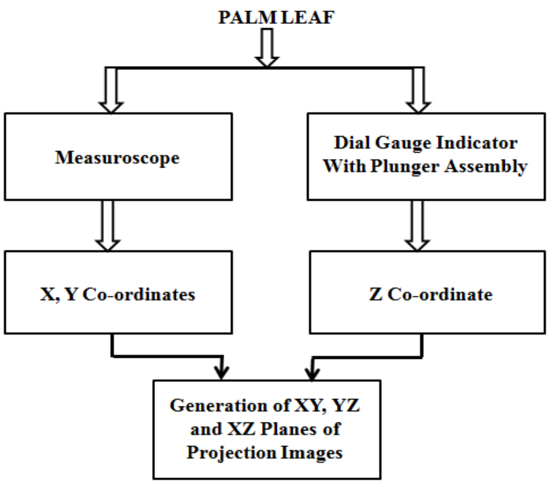

Figure-1. Data collection process.

The left most point of the character is fixed as Origin O (0, 0). The other points are selected at the junctions, turning points loops and end points as defined above. These selected points along the contour of the Palm Leaf character "Va" are illustrated in Figure-2.

The 'Z' dimension is the depth of indentation proportional to the pressure applied by the scriber using stylus at each pixel point of the Telugu character [13], [14], [15], [16] and [17]. A special needle is fabricated made up of Teflon which is attached to the dial indicator plunger [13], [14], [15], [16] and [17]. This needle is so designed that it does not damage the palm leaves during the measurements at various pixel points. The measurements were also taken with utmost care and interest. The needle attached to the dial indicator plunger is first positioned at the pixel point where the 'Z' Dimension is to be measured for any character. The distance of the bottom of the pixel point is measured and recorded. This distance is termed as D1.

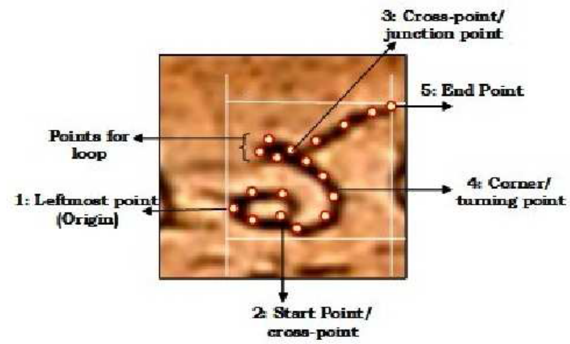

Figure-2. Selection of pixel points along the contour of a Palm Leaf Character "Va".

Then this needle is placed on the surface of the palm leaf nearer to the pixel point and again the distance is measured called D2. The depth of indentation D for the selected pixel point is found by subtracting D2 from D1. Hence the depth of indentation at any pixel co-ordinate is D = D1 - D2. This procedure is repeated at all the pixel points of the Palm Leaf Character where 'X' and 'Y' measurement were taken. Hence for every Palm leaf character at any pixel point there are 3 dimensions 'X', 'Y' and 'Z' [13], [14], [15], [16] and [17].

B. Multi-level Recognition system

The architecture of the proposed multi-level system is shown in Figure-3. The system is trained and tested for each plane of projection i.e., for 'XY', 'XZ' and 'YZ'. Each character is taken as an input image and the size of the images is normalized to 32X32. These images are preprocessed and binarized using a threshold value of 0.7. In the first level 2-D FFT is used to extract the features for both the training and testing images. The Euclidean distance is calculated between the test image and all the training character images. The database image whichever has the smallest Euclidean distance to the test image is considered to be matched with the test image. The proposed algorithm checks whether the test image is correctly identified. The wrongly identified test characters in the first level are given as input to the second level. In the second level 2-D DCT is used to extract the features for both the training and the wrongly identified (test) images. Again Euclidean distance is calculated between the test and training characters (new feature set using 2-D DCT). The minimum distance is found out and then the corresponding database image is set to be recognized as the test image. The overall recognition accuracies are reported for all the three planes of projection in the next section.

ARPN Journal of Engineering and Applied Sciences

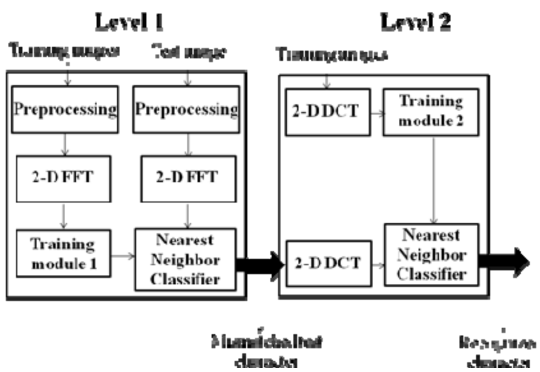

Figure-3. Architecture of multi-level feature recognition system 2-D FFT in the first level and 2-D DCT in the second level.

Let the image be denoted by $f(x,y)$ of size MXN. The definition of 2-D FFT for the image $f(x,y)$ is depicted in equation (1).

$$F(u,v) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\, e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \qquad (1)$$

$$for\ 0 \le u \le M-1 \qquad 0 \le v \le N-1$$

Here x, y are the pixel coordinates in the image and u, v are coordinates in the "transformed image". In the first level all the characters are pre-processed and classified by taking the 2-D FFT of the training and testing character images. The Euclidean distance between the training and the testing characters are calculated and the minimum the distance between them is considered as the matched image. The mismatched test characters from the first level are given as input to the second level to improve the recognition accuracy. The 2-D DCT is used in the second level of the algorithm which is depicted in equation (2).

$$C(p,q) = \alpha_p\alpha_q \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\cos\frac{\pi(2m+1)p}{2M}\cos\frac{\pi(2n+1)q}{2N} \qquad (2)$$

$$for\ 0 \le p \le M-1 \qquad 0 \le q \le N-1$$

Where

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{2/M}, & 1 \le p \le M-1 \end{cases} \qquad \alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{2/N}, & 1 \le q \le N-1 \end{cases}$$

After using the 2-D DCT features for the character images, they are classified by computing the Euclidean distance between the mismatched character and the training character images. The minimum the distance between them the more they are similar.

Further, in the next step the algorithm was modified, i.e., interchanging the feature vectors (2-D DCT in the first level and 2-D FFT in the second level). The architecture after interchanging the feature vectors is shown in Figure-4. The overall recognition accuracies are reported in the next section by interchanging the combinations in the two levels.

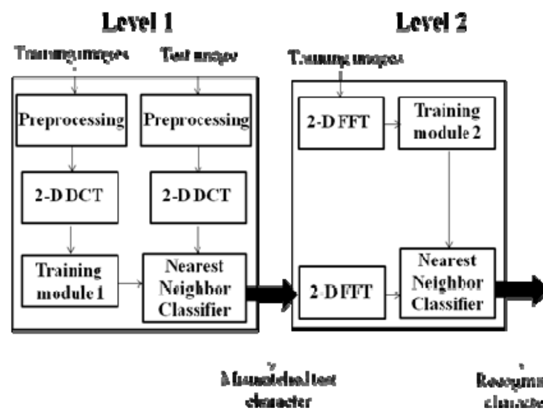

Figure-4. Architecture of multi-level feature recognition system 2-D DCT in the first level and 2-D FFT in the second level.

4. RESULTS AND DISCUSSIONS

There are many characters in Telugu which are highly similar and hence their patterns are always confusing for recognition. These similar characters can be grouped [13], [14], [15], [16] and [17] together and further certain special features can be extracted to distinguish between these similar characters. The highly similar characters Ae, Pa and Na are shown in Figure-5. The characters Ae, Pa and Na in Telugu are very similar and can be grouped in the same group which have similar pattern and this causes decrease in recognition accuracy for any recognition model. Hence the recognition accuracy obtained is very low for Telugu characters [13], [14], [15], [16] and [17].

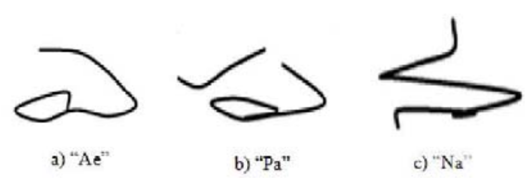

Figure-5. Some confusing telugu characters in XY plane of projection.

VOL. 10, NO 20, NOVEMBER, 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

However this problem is solved to a great extent in the proposed method even for Ae, Pa and Na characters if we consider the respective ‘YZ’ and ‘XZ’ plane of projection images. The patterns obtained for Telugu characters Ae, Pa and Na in ‘YZ’ and ‘XZ’ plane of projections are shown in Figure-6. It is very clear from these patterns that these patterns are completely different to each other; thereby recognition accuracy would naturally increase for the proposed algorithm.

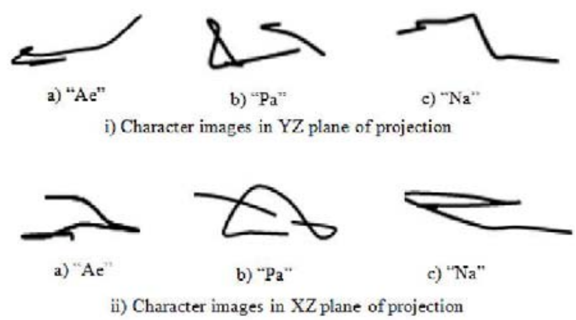

Figure-6. Same confusing characters shown in the above fig in YZ and XZ planes of projection.

The experimental results for all planes of projection viz., ‘XY’, ‘XZ’ and ‘YZ’ are shown in Table-1. Initially features are extracted at every pixel of the character image using 2-D FFT in level 1. The Euclidean distance is used to classify the characters as described in methodology section. The wrongly identified (test) characters from level 1 are further classified by using 2-D DCT features in level 2. It is observed from Table-1 that the best recognition accuracy obtained in level 1 is in ‘YZ’ plane of projection i.e., 71.4% and the overall recognition accuracy after level 2, obtained in ‘YZ’ plane of projection is 92.8%. For Telugu characters the maximum variation is found in ‘Y’ direction, which is the inherent characteristic of the Telugu script [15], [16], [17]. The measured depth at different pixel points of a Telugu character (a 3D feature) which is proportional to the pressure applied by the scriber is an important feature for palm leaf character recognition [15], [16], [17]. Hence the combination of these ‘Y’ and ‘Z’ features has yielded excellent recognition accuracy of 92.8%. If we consider ‘XZ’ projection plane, ‘Z’ component i.e., the 3D feature has contributed for getting 67.85% recognition accuracy. However, in the absence of this 3D feature ‘XY’ plane of projection yielded only 32.2% of recognition accuracy.

Table 1. Two-Level feature recognition system.

Plane of projection	**Recognition accuracy %**	
	Level 1 2-D FFT	**Level 2 2-D DCT**
XY	25	32.2
YZ	71.4	92.8
XZ	50	67.85

There is an adequate increase in recognition accuracies with a two-level system compared to a single-level system. Also on comparing the level 1 and level 2 results it is observed from Table-1 that in ‘YZ’ plane of projection there is a good improvement in accuracy compared to the other planes of projection.

The two-level system is also tested by extracting features at every pixel point using 2-D DCT in the first level and 2-D FFT in the second level. The experimental results after interchanging the feature sets are shown in Table-2. Even for this combination it is observed that ‘YZ’ plane of projection yielded better results. The overall recognition accuracies are same for both the architectures. From Table-2 it is observed that in ‘YZ’ plane of projection in level 1 the best recognition accuracy obtained is 85.7%. On comparing the level 1 accuracies from tables 1 and 2 it is observed that for all the plane of projections 2-D DCT has given better performance compared to 2-D FFT.

Table-2. Two-Level feature recognition system.

Plane of projection	**Recognition accuracy %**	
	Level 1 2-D DCT	**Level 2 2-D FFT**
XY	32.14	32.2
YZ	85.7	92.8
XZ	67.8	67.85

The recognition accuracies for the published and the proposed two-level feature recognition systems are shown in Table-3. Sastry *et al*., [13] have worked on the same database using PCA and reported the best recognition accuracy as 40%. Sastry *et al*., [14] also published that, by extracting the 2-D correlation features, the best recognition accuracy as 90%. In both published and proposed methods the best accuracies achieved are in ‘YZ’ plane of projection. The highest recognition accuracy achieved is 92.8% with the proposed two-level approach in ‘YZ’ plane of projection. This is in line with the earlier results. In both the published and proposed methods, ‘YZ’ plane of projection only gave the best results, as discussed earlier.

ARPN Journal of Engineering and Applied Sciences

Table-3. Comparison of proposed two-level method with existing methods.

Plane of projection	% Recognition accuracy		
	Published methods		Proposed two-level method
	PCA [13]	2D Correlation [14]	
XY	40	54	32.2
YZ	40	90	92.8
XZ	37	70	67.85

5. CONCLUSIONS AND FUTURE WORK

a) The recognition accuracy of handwritten character recognition (HWCR) is very low as reported in the literature.
b) Successfully developed palm leaf character set comprising 1232 images for training and 308 images for testing.
c) The recognition accuracy of palm leaf character recognition is very low in 'XY' and 'XZ' planes of projection. In the 'YZ' plane the recognition accuracy is high since there is lot of variation between any two Telugu characters in 'Y' direction which is the inherent characteristic of Telugu script. The additional feature 'Z' adds on along with the 'Y' characteristic to achieve high recognition accuracy in 'YZ' plane of projection.
d) There are some confusion characters like ba, bha, ja, etc., which are similar and confuse the computer. Hence the recognition accuracy obtained for the proposed algorithm is only 92.8%.
e) In future automation of palm leaf characters with matras can be carried out.

ACKNOWLEDGEMENT

This work is done as a part of AICTE project titled "Design and Development of Palm Leaf Character Recognition System" under RPS (Research Promotion Scheme). Hence the authors express their sincere thanks to the funding agency, AICTE, New Delhi, India for their support and encouragement.

REFERENCES

[1] Aburas, Abdurazzag Ali Gumah and Mohamed E. 2008. Arabic Handwriting Recognition: Challenges and solutions. Intl. Symposium On information Technology. pp. 1-6.

[2] Ujwal Bhattacharya and B.B.Chaudhuri. 2009. Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. IEEE transactions on pattern analysis and machine intelligence. 31(3): 444-457.

[3] C.Vasantha Lakshmi and C.Patvardhan. 2003. A high accuracy OCR System for Printed Telugu Text. IEEE Conference on Convergent Technologies for Asia-Pacific region, TENCON. 2: 725-729.

[4] K. Sharma Rajiv and S. Dhiman Amardeep. 2010. Challenges in segmentation of text in handwriting Gurumukhi Script. communications in Computer Information Science. 70: 388-392.

[5] Zhixinshi, Srirangaraj Setlur and Venu Govindaraju. 2005. Digital enhancement of palm leaf manuscript images using Normalization techniques. Center of Excellence for Document Analysis and Recognition (CEDAR). pp. 1-27.

[6] C.Vasantha Lakshmi, Ritu Jain and C.Patvardhan. 2007. Handwritten Devnagari Numerals Recognition with higher accuracy. International Conference on Computational Intelligence and Multimedia Applications, IEEE. 3: 255-259.

[7] V.N. Manjunath Aradhya, G. Hemanth Kumar and S. Noushat. 2008. Multilingual OCR system for South Indian Scripts and English documents: An approach based on Fourier transform and PCA. Elsevier, engineering applications of artificial intelligence. pp. 658-668.

[8] Arun K.Pujari, C.Dhanunjaya Naidu, M.Sreenivasa Rao and B.C. Jinaga. 2004. An intelligent character recognizer for Telugu scripts using multi resolution analysis and associative memory. Image and Vision Computing. Vol. No. 22, pp. 1221-1227.

[9] T V Ashwin and P S Sastry. 2002. A Font and size-independent OCR system for printed Kannada documents using support vector machines. Sadhana. 27(part 1): 35-58.

[10] U.Pal, S.Sinha and B.B.Chaudhuri. 2003. Multi-Script Line identification from Indian Documents. Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), IEEE. pp. 880-884.

ARPN Journal of Engineering and Applied Sciences

[11] Atul Negi and Chandra Kanth Chereddi. 2003. Candidate Search and Elimination Approach for Telugu OCR. Conference on convergent technologies for Asia-pacific region, TENCON. 2: 745- 748.

[12] Chakravarthy Bhagvati, Tanuku Ravi, S. Mahesh Kumar and Atul Negi. 2002. On Developing High Accuracy OCR Systems for Telugu and Other Indian Scripts. Language Engineering Conference (LEC'02), IEEE. pp. 18-23.

[13] P. N. Sastry, R. Krishnan, and B. V. S. Ram. 2010. Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach. J. Applied Engn. Sci. 5(3): 22-32.

[14] P. N. Sastry, R. Krishnan, and B. V. S. Ram. 2008. Telugu character recognition on palm leaves- a three dimensional approach. Technology Spectrum. 2(3): 19-26.

[15] P. N. Sastry and R. Krishnan. 2012. Isolated Telugu palm leaf character recognition using radon transform, a novel approach. In World Congress on Information and Communication Technologies (WICT). pp. 795-802.

[16] P. N. Sastry and R. Krishnan. 2012. A data acquisition and analysis system for palm leaf documents in Telugu. In: Proceeding of the Workshop on Document Analysis and Recognition, ser. DAR '12. New York, NY, USA: ACM. pp. 139-146.

[17] Lakshmi T.R.V., Sastry P.N., Krishnan R., Rao N.V.K., Rajinikanth T.V. 2015. Analysis of Telugu Palm Leaf Character Recognition Using 3D Feature. International Conference on Computational Intelligence and Networks (CINE-2015). pp. 36-41.