Telugu Handwritten Isolated Characters Recognition using Two Dimensional Fast Fourier Transform and Support Vector Machine

Raju Dara
Research Scholar,
Department of Computer Science and Engineering
Jawaharlal Nehru Technological University,
Kakinada, Andhra Pradesh, India

Urmila Panduga
Senior Software Engineer
Cubic Transportation Systems India Pvt. Ltd

ABSTRACT
Research in character recognition is an old application in the area of pattern recognition and has attracted many researchers during the last few decades. Handwritten character recognition (HCR) is of two types namely, Online and Offline. The recognition accuracy for HCR is less than 60% as per the literature survey. Also the non existence of standard database for Indian languages is another reason for motivation of this work. This work describes Offline HCR by extracting features using 2D FFT and using the support vector machines for Telugu documents. The best percentage recognition accuracy for Telugu handwritten characters is 71%.

Keywords
Handwritten character recognition, 2D FFT, support vector machine Classifier, Pattern Recognition.

1. INTRODUCTION
Optical character recognition (OCR) is predicated on optical system, which allows a device to identify the characters automatically. There are several applications of OCR few of them are as follows like automatic mail sorting, bank cheques, library automation, reading aid for the blind, language processing and defense applications that produce lots of interest for researchers. The popularity accuracy is higher in OCR systems for printed characters compared to hand written character recognition (HCR) systems [1, 2, 3 and 4]. Basically, the HCR is of 2 types: Online and Offline. The information is captured throughout the calligraphy process with the assistance of a special pen on an electronic surface in Online HCR, whereas documents are scanned pictures of prewritten text in Offline HCR [5]. The accuracy is a smaller for Offline HCR as reported within the literature [1, 2 and 3], it is true because of the issues associated with skew angle, legibility, overwriting, and authentication of document and noise issues [2]. Also, there is no standard or benchmark information available and it could be a major obstacle for analyzing on HCR of Indian scripts [1, 2, 3, 4 and 6]. Several customary databases like MNIST, NIST, CEDAR and CENPARMI are accessible for Latin numerals [6], however there does not exist any standardized database for Indian scripts. The previous studies are supported tiny databases, which are collected in the laboratory [6]. The expressive style of users is one in every of the characteristics for written knowledge. It's a tedious task to tell apart from the written characters after they overlap.

India being a bilingual country has sixteen major languages and over a hundred regional languages [4] that clearly show the necessity of bilingual and multi-script recognition systems. Additionally majority of the documents in Asian countries consists of text information in addition to the script or language forms. There are approximately hundred million Telugu speaking folks [7] within the world that indicates the necessity of HCR for Telugu. Several Telugu characters have high rate of similarity and each of these characters is classified into six completely different categories [3]. The method of character recognition in Asian and significantly in Indian scripts is in a very initial stage. A number of the elucidations [1, 2, 3, 4] are as follows:

1. Indian languages have a lot of range of basic and composite characters compared to English.
2. Telugu or the other Indian languages have tiny range of users as compared to English.
3. There are several junctions, union of characters or modifiers in Telugu and different Indian languages.

There are 18 vowels and 36 consonants in Telugu language. Some of the Telugu characters and their pronunciations are shown as follows:

అ	→	**Aa**	మ	→	**Ma**
ఆ	→	**Aaa**	వ	→	**Va**
ఎ	→	**E**	స	→	**Sa**
ఏ	→	**Ee**	న	→	**Na**
ఒ	→	**O**	ద	→	**Da**
ఓ	→	**Oo**	ధ	→	**Dha**
బ	→	**Ba**	చ	→	**Ca**
భ	→	**Bha**	ఛ	→	**Cha**

This language is mostly used in the states of Telangana and Andhra Pradesh. Telugu being the local language and known to most of the village community, and there are many documents having English and Telugu characters together. These types of documents such as Railway reservation form, Birth report, Death report, Aadhaar card acknowledgement etc. Some of the documents are shown in figures 1, 2 and 3.

Figure-1 shows news regarding two magistrates delivered a verdict in Telugu. Lord Hanuman Prayer is shown in Telugu in figure-2.

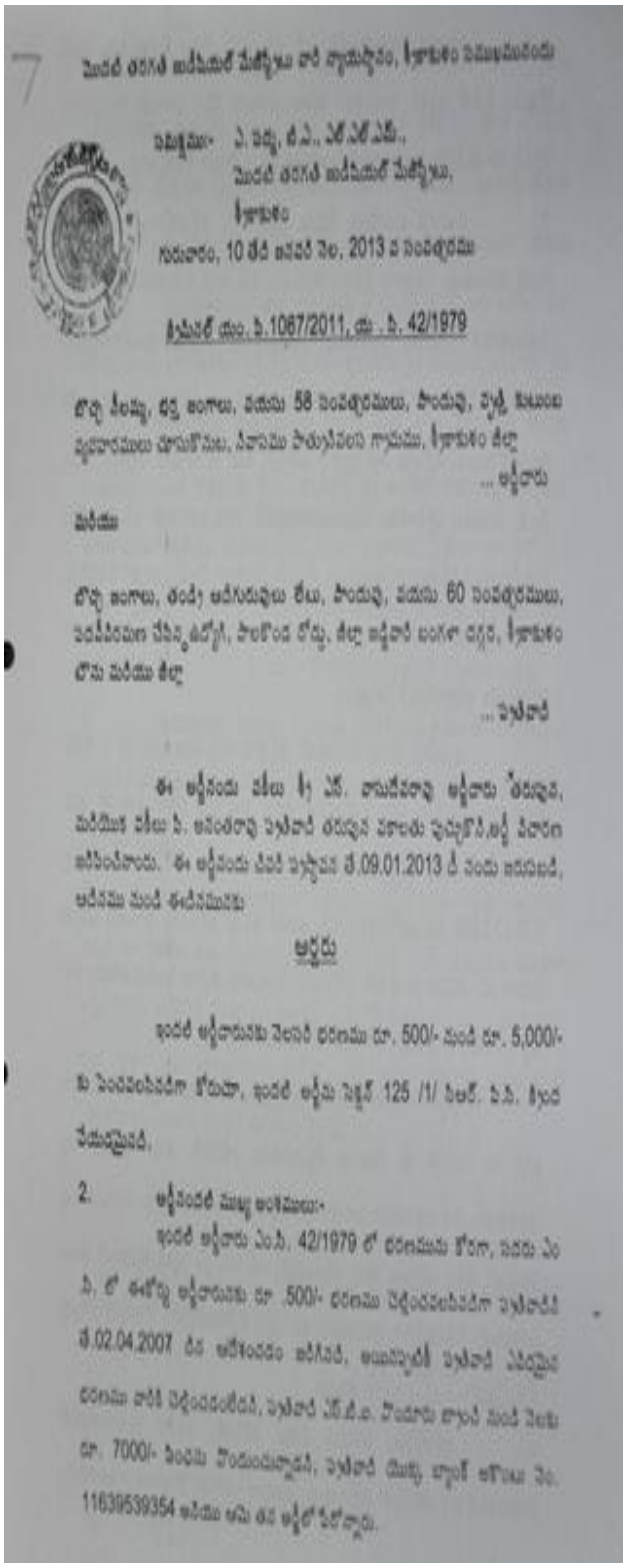

Fig 1: Two magistrates deliver verdict in Telugu

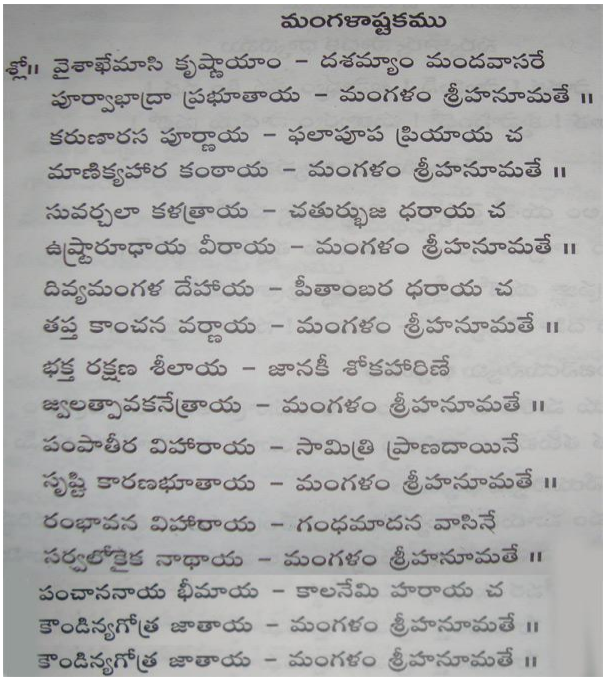

Fig 2: Lord Hanuman Prayer in Telugu

This paper describes the process of recognizing the isolated basic characters in a document. Section II describes existing methods for Handwritten Character Recognition systems. The step by step algorithm is described in section III. In section IV the experimental results are shown. Section V reports the final conclusions.

2. BACK GROUND WORK FOR HANDWRITTEN CHARACTER RECOGNITION SYSTEMS

Manjunath Aradhya and Hemanth Kumar [4] worked for English handwritten documents based on Fourier transform followed by PCA for 75 training samples or character classes, and reported the recognition accuracy as 79.6% for Fourier PCA (F-PCA) whereas the recognition accuracy for PCA is reported as 76.6%. When the number of training samples was increased to 175 classes, the recognition accuracy was reported as 93.8% and 90.8% for F-PCA and PCA respectively. The bilingual recognizer for printed documents consists of Hindi and Telugu is presented by Jawahar et al [8], which is based on principal component analysis (PCA) followed by support vector classification (SVC), and determined Hindi characters provide good recognition even at a resolution of 15x15, whereas Telugu characters need more details with 40x40, finally reported the overall accuracy of 96.7%. Using cross correlation coefficient concept, the offline HCR for the Gujarati script is presented by Prasad et al [9], in this work, the shape of the character is analyzed and the features are compared to identify character and the overall efficiency of the system is found to be 71.66%. Sutha and Ramaraj [10] worked on Tamil HCR, Fourier descriptors are used as feature vectors for identifying a character, implemented multilayer perceptron with a hidden layer in networks and reported recognition accuracy of 97%. Slant correction is improved using octal graph and independent of the writing style the basic form of a character is represented using the graph [11] and the recognition accuracy is found to be 82%. Ramakrishnan et al [12] worked on a specific font in English based on Zernike moments and Delaunay triangulation, Support Vector Machine Classifier is used for

classification, and the accuracy is found to be 85.85% for words. When the written text had a signature, then the reported accuracy is 92.22%. Pujari A.K. et al [13] worked for Telugu HCR using wavelet multi resolution analysis and associative memory, Hope field-based Dynamic neural network is used for learning the style and font from the document, and tested the same text for different fonts and the efficiency ranged from 85% to 92.1%. Bunke et al [14] worked on offline cursive HCR system using Hidden Markov Model; the words are recognized by extracting the edges in the order of skeleton graph of the word, in this work 9000 words for training and 3000 words for testing are used and reported the accuracy to be 98%. Hanmandlu et al [15] proposed a method for language independent HCR system using neural networks and fuzzy logic, the vector distance between each point and a fixed point is used as a feature vector and reported a recognition accuracy of 97% to 98% for neural network and fuzzy logic respectively.

3. METHODOLOGY

There is no any common obtainable place for databases for Indian languages and hence it becomes terribly troublesome for the event of hand written character, thus the databases are developed by the researchers within the laboratories for hand written character recognition [6]. There are eighteen vowels and thirty six consonants in Telugu.

3.1 Dataset Details

In this article, the amount of categories or characters information developed is 100. Every character is written on a paper in an exceedingly rectangular manner in completely different sizes and designs by 50 individual writers. The written documents were then inheritable by scanning by a flatbed scanner. These pictures are preprocessed to the minimum boundary parallelogram conception; social control is performed to a size of 100x100. Thus a complete variety of 1500,750 samples were developed for coaching whereas 750 samples were developed for testing purpose. To increase the amount of databases, every image is resolved by +30, -30, +50 and -50 and keeps the feature in coaching sample set. By this the amount of the information is raised to 5 times.

3.2 Algorithmic Rule for Implementing 2-D FFT

The following is the step by step algorithmic rule for 2-D FFT, enforced on basic and isolated written characters.

1. Load the pictures of size of 100x100 pixels.
2. Scan all the pictures.
3. Binarize the pictures employing a threshold of 0.85
4. Rotate the pictures by -3, +3, -5 and +5 degrees to get artificial information.
5. Notice the 2-D FFT for all the first and synthetically generated pictures.
6. Reshape the matrix into column matrix to make the feature vector for every image.
7. Repeat the higher procedure to take a look at pictures additionally.
8. Notice the space between the column matrices and take a look at image and every training image.
9. Notice the minimum distance.
10. Increments the match counts if category matches or else increment the couple count.
11. Show the take a look at image and also the matched info image.

3.3 Mathematical model of 2-D FFT

Two dimensional Fourier transforms involves a number of one dimensional Fourier transforms. From the definition of 2D FFT equation (1) shows the 2D FFT of the image f (p, q) and equation (2) represents inverse 2D FFT. The 2D FFT F (s, t) for the image f (p, q) can be found using formula

$$F(s,t) = \frac{1}{MN} \sum_{p=0}^{M} \sum_{q=0}^{N} f(p,q) e^{-j2\pi(\frac{sp}{M}+\frac{tq}{N})} \quad (1)$$

Whereas, s = 0, 1, 2 ….M, & t = 0, 1, 2 ….N.

Here *p, q* are the pixel coordinates in the image and *s, t* are coordinates in the "transformed image".

The inverse 2D FFT (from FFT back to the original image, possibly after filtering) can be obtained using below given formula

$$f(p,q) = \sum_{s=0}^{M} \sum_{t=0}^{N} F(s,t) e^{j2\pi(\frac{sp}{M}+\frac{tq}{N})} \quad (2)$$

Where p = 0, 1, 2 ….M, & q = 0, 1, 2 ….N.

These formulas assume calculations using complex numbers ($j = \sqrt{(-1)}$).

4. RESULTS AND DISCUSSIONS

The image of a document is scanned and saved in the computer which is used as an input image for the recognizer. The scanner used is 600 dpi using WIA Cano scan LiDE 100 flatbed scanner. The images are preprocessed first using Adobe Photoshop. Further they are preprocessed using MATLAB tools and the synthetic data is generated, then the 2-D FFT is applied to all the images. Finally, the Euclidean distance is measured between each test image and the database images. It is found that the database image has the lowest Euclidean distance to the test image. The test image is displayed on the left side, whereas the matched image of the database is shown on the right side.

The total number of isolated character set of Telugu database consists of 4,250 samples. Out of 4,250 samples 3,750 samples are used for training and 500 samples are used for testing. Further the size of the database is increased by rotation as described earlier, thus the available size of the database is 18,750 samples (375samples or classes). The system is trained by varying the training sample number by 75, 150, 225, 300 and 375. Table I shows the recognition accuracy by varying the number of training samples or classes.

All the Telugu characters can be divided into 6 groups, each character within the group has high similarity measure with any other character in the same group [3]. A few matched and mismatched samples of Telugu database are shown in the figure- 4 and figure-5 respectively. In every set, the test image is displaced first and then its matched database image. From figure-5 it is very clear that due to more number of similar characters in Telugu script there is a lot of confusion between two characters of the same group [3]. Hence the recognition accuracy drastically decreases.

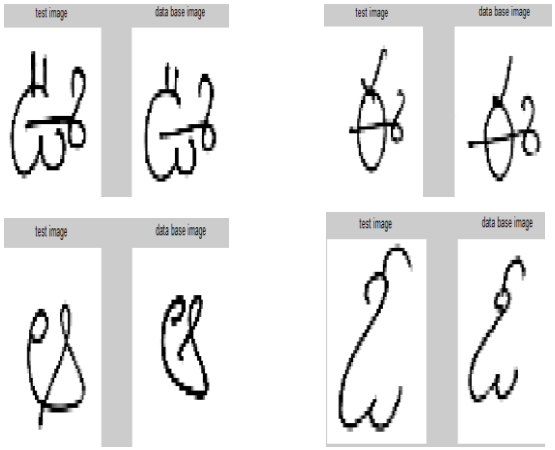

Fig 3: Characters matched correctly

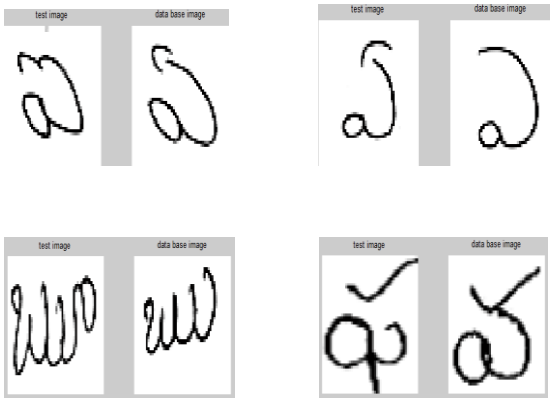

Fig 4: Mismatched characters

Table 1. Table captions should be placed above the table

Number of Training Sample Classes	Recognition Accuracy (%)
75	32
150	56
225	59
300	61
375	63

In table 2, the 2D FFT results are compared with the published method. Aradhya and Hemanth Kumar [4] experimented on English characters by extracting the features using Fourier transform followed by PCA and Support Vector Machine for classification. The recognition accuracy was reported as 85% for 100 training samples or classes. In this paper, the features are extracted using 2-D FFT and Support Vector Machine, and the recognition accuracy is 71% even though the proposed method uses a single stage for feature extraction.

Table 2. Comparison of 2D FFT results

Description	Published Method [4]	Proposed Method
Language	English	Telugu
Feature extraction	Fourier transform followed by PCA	Two Dimensional Fast Fourier Transform
Number of training samples or classes	75	375 (75 original data + 300 synthetic data)
Medium	Written on paper	Written on paper
Recognition Accuracy	85%	71%

5. CONCLUSION

Handwritten character recognition for Telugu characters is explored during this work. A complete variety of 1500, 750 samples used for training and 750 samples for testing area is developed and additionally normalized to 100x100 pixels. The popularity accuracy obtained by exploitation of 2D FFT is 71%.

6. REFERENCES

[1] P. N. Sastry and R. Krishnan. "Isolated Telugu palm leaf Character recognition using radon transforms–a novel approach". In *IEEE-WICT*, 2012.

[2] P. N. Sastry, R. Krishnan and T.V. Rajinikanth. Palm leaf Telugu character recognition using Hough transform. In *Proceedings of International Conference on Advanced Computing Methodologies (ICACM-2011)*, pages 21–28, Dec 2011 Elsevier.

[3] P. N. Sastry, R. Krishnan and B. V. Sanker Ram. Telugu character recognition on palm leaves-a three dimensional approach technology. *Spectrum (JNTU Hyderabad)*, 2(3):19–26, Nov. 2008.

[4] S.V.N.Manjunath Aradhya and G.Hemanth Kumar. Multilingual OCR system for south Indian scripts and English documents. In *5th IFIP International Conference on Intelligent Information Processing*, pages 658–668, 2008 Elsevier.

[5] Munish Kumar, R.K.Sharma and M.K.Jindal. Offline Handwritten Gurumukhi Character Recognition: Study of Different Feature-Classifier Combinations. *In the proceedings of the Workshop on Document Analysis and Recognition, Dec 2012.*

[6] U. Bhattacharya and B.B. Chaudhuri. Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals. *IEEE transactions on pattern analysis and machine intelligence*, 31(3):444–457, Mar. 2009

[7] P. N. Sastry, R. Krishnan and B. V. Sanker Ram. Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach. *ARPN Journal of Engineering and Applied*

Sciences, 5(3), Mar. 2010.

[8] Jawahar, C.V., Pavan Kumar, M.N.S.S.K., Ravi Kiran, S.S., 2003. A bilingual OCR for Hindi-Telugu documents and its applications. *In proceedings of ICDAR, 3-6 August, Edinburgh, pp 656-660.*

[9] Prasad J.R., Kulkarni U.V., Prasad R.S., "Offline handwritten character recognition of Gujarati script using pattern matching". *ASID (Anti-counterfeiting, Security and Identification in Communication) 2009 pp.611-615.*

[10] Sutha, J. and Ramaraj N., "Neural Network Based Offline Tamil handwritten character recognition system". *International conference on computational Intelligence and Multimedia Applications", 2007, vol.2, pp.446-550.*

[11] R. Jagadeesh Kannan and R. Prabhakar, "An improved Handwritten Tamil Character Recognition System using Octal Graph", *Journal of computer science, vol.4, No.7, 2008, pp.509-516.*

[12] Kandan Ramakrishnan, Arvind K.R. and A.G.Ramakrishnan, "Localization of handwritten text in documents using moment invariants and Delaunay Triangulation", *International Conference on computational Intelligence and Multimedia Applications, vol.3 2007, pp.408-414.*

[13] Arun K. Pujari, C. Dhanunjaya Naidu, M.Sreenivasa Rao and B.C.Jinaga, "An intelligent character recognizer for Telugu scripts using multiresolution analysis and associative memory", *Image and vision computing, vol.22. 2004, pp.1221-1227.*

[14] H.Bunke, M.Roth and E.G.Schukat-Talamazzini, "Offline Cursive Handwriting Recognition using Hidden Markov Models", *Pattern Recognition, Vol.28, No.9, 1995, pp. 1339-1413.*

7. AUTHORS' PROFILE

Raju Dara is a Research Scholar of Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Kakinada. He has 12 years of teaching experiences for Graduate and Post Graduate engineering courses. His current research interests are Data Warehousing, Image Processing. He published 5 research papers in international journals and 3 research papers in international conferences.

Urmila Panduga is working as a senior software engineer for the past 8 years with various software industries and worked as an Assistant Professor for 2 years in an Engineering college, her areas of interest are Image Processing, Databases, and System Programming. She published 2 research papers in international journals and 1 research paper in international conference