Recognition of Tamil Handwritten Characters using Daubechies Wavelet Transforms and Feed-Forward Backpropagation Network

Tiji M Jose
PG Scholar
Department of IT
Bannari Amman Institute of Technology
Sathyamangalam-638 401

Amitabh Wahi
Professor
Department of IT
Bannari Amman Institute of Technology
Sathyamangalam-638 401

ABSTRACT

This article suggests wavelet transform based feature extraction technique for extracting robust features from Tamil handwritten characters. The algorithm uses feed-forward back propagation neural network as the classifier. It presents the relevant features of Tamil script and describes various techniques used for character recognition. In common, all the pattern recognition tasks focus on extracting more differentiating features. This is the most important and complicated job. The proposed system concentrates on two dimensional discrete wavelet transformations for extraction of features. For multiresolution analysis of images, Wavelet Transform is used. This capability can be used to study the character image in different frequency bands. Localized basis functions of WT are used for extracting localized features of a character image. This enables us to obtain more distinct traits as features for each character. Feed forward back propagation neural network is one of the general neural network architectures and this architecture can be applied to many different tasks and is very popular.

Keywords

Back Propagation Neural Networks, BPNN, offline Tamil character recognition, Wavelet Transform, WT, localized basis functions, MatLab, haar, approximation coefficients

1. INTRODUCTION

Optical character recognition (OCR) is a vital research area in pattern recognition. OCR system is recognizing alphabetic letters, numbers, or other characters, which are in the form of digital images, without any human interference [1]. This is achieved by investigating a match among the features extracted from the given character's image and the library of image models. The features should be different for every character images so that the computer can mine the right replica from the library without any misunderstanding. Two different classifications are integrated in the common term of character recognition [2]:

• On-line character recognition

• Off-line character recognition

On-line character recognition uses a data stream from a transducer as input while the user is writing. The data is collecting through a electromagnetic or pressure sensitive digitizing tablet. When the user writes on the tablet, the consecutive movements of the pen are translated to a sequence of electronic signal which is stored and inspected by the computer [3].Off-line character recognition is performed only after the writing is finished. The major difference among on-line and off-line character recognition is in time-series background knowledge of on-line character recognition.This differentiation creates a considerable deviation in processing architectures and systems. The off-line character recognition is again divided into [4]:

• Magnetic character recognition (MCR)

• Optical character recognition (OCR)

In MCR, magnetic ink is using for the characters to print. The reading device can recognize the characters in accordance with the unique magnetic field of each character. For checking authentication, MCR is normally used in banks. Recognition of characters gained by optical means, normally through a scanner or a camera, deals with OCR [5]. As early as 1929, Tausheck got a patent named "Reading Machine" in Germany [6]. This patent replicates the basic concept of today's OCR. The OCR can be categorized into handwritten character recognition and printed character recognition [7]. Handwritten character recognition is more complicated to employ than printed character recognition due to the complexity in human handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial, Courier, etc

The idea behind an OCR is to identify and analyse a document image by dividing the page into line elements, further sub-dividing into words, and then into characters. These characters are compared with image patterns to predict the probable characters. Recognition of characters can be done either from printed documents or from handwritten documents. Handwritten document recognition can be done offline or online. Offline character recognition is more complicated than online. In particular, Tamil handwritten OCR is more complicated than other related works. This is because Tamil letters have more angles and modifiers. Additionally, Tamil script contains large number of character sets. A total of 247 characters; consisting of 216 compound characters, 18 consonants, 12 vowels and one special character. The main challenges in OCR research are due to the curves in the characters, number of strokes and holes, sliding characters, differing writing styles so on.Pre-processing, segmentation, feature extraction and classification are major steps involved in character recognition.

Translating scanned images to machine readable text focuses paperless environment which leads to the concept of optical character recognition. It increases the demand in many merging applications like postal system, banks , institutions, word processing, library system etc where all the processing are automated. It is one of the field of research in artificial intelligence which is branch of computer science and aims at to create intelligence in machines. Recognition of hand

printing, handwriting and printed text are the main focusing research area because still no 100% recognition is possible even though the available scanned image is accurate. Text can be in different scripts, numerals, images.

The paper is organized as ten Sections. A short review of Previous OCR research in Tamil scripts is provided in Section II. Section III elaborates salient features of Tamil Script. Implementation model of the proposed system is given in Section IV. Results and Discussion is explained in Section V. The paper is concluding in Section VI and Future work is also covered in that section. Finally, references are given in section VII.

2. PREVIOUS OCR RESEARCH IN TAMIL

Siromoney et al. [8] illustrated a scheme for recognition of machine printed Tamil characters using an encoded character string dictionary.Chinnuswamy et al. [9] put forwarded an idea for hand-printed Tamil character recognition. In this, the characters are consisting of line-like elements, called primitives.Topological matching procedure to calculate the correlation coefficients and then maximizes the correlation coefficient.Suresh et al. [10] elucidates the fuzzy concept on handwritten Tamil characters using a feature called distance from the frame and a membership function. The system is tested for 250 samples for numerals and seven Tamil characters and the recogntion rate achieved ranging from 76% to 94%.Hewavitharana, S, and H.C. Fernando [11] presents an idea to recognize handwritten Tamil characters with two-stage classification approach, which is an amalgam of structural and statistical techniques. In [12] illustrated offline unconstrained handwritten Tamil character recognition system based on support vector machine (SVM).Shivsubramani et al. [13] proposes a well-organized technique for recognizing printed Tamil characters by using the interclass relationship between them, which should be achieved using Multiclass Hierarchical Support Vector Machines.

3. TAMIL SCRIPT

Tamil , the South Indian language is one of the oldest languages in the world.It is the official language of not only the Indian state of Tamil Nadu, but Also countries like Singapore, Malaysia and Sri Lanka. There are millions of speakers of this language all around the world. Tamil alphabets are very old and are categorized in a methodical way. The alphabets are divides into vowels, consonants, composite letters, and special letter.Tamil script is having 12 vowels, 18 consonants, 216 composite letters, one special character (AK) and 14 other characters. Composite letters are are obtained by the combinations of consonants and vowels [9]. In general,there are 67 Tamil characters as the basic characters (Vowels, Consonants, and composite letters).By identifying the 67 basic characters, all the 247 characters can be recognized.

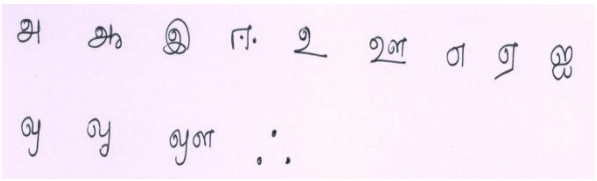


Fig 1: Handwritten vowels

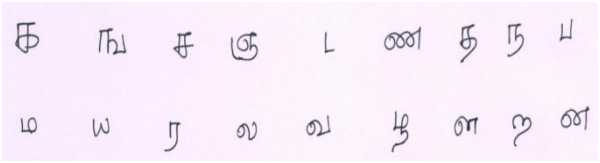


Fig 2: Handwritten consonants

The Universal Character encoding scheme used for writing characters and text in Tamil script is Unicode Standard . The Tamil Unicode range is U+0B80 to U+0BFF. The Unicode characters are comprised of 2 bytes in nature. For example, the Unicode for the character is 0B85; the Unicode for the character is 0BAE+0BC0. The Unicode is designed for various other Tamil characters. Like other indian scripts tamil script also evolved from brahmi script also called tamil brahmi.Tamil brahmi almost similar to writing system present in Tholkoppiyam which is one of the tamil grammer followed in ancient days.

4. SYTEM IMPLEMENTATION MODEL

Proposed System is having modules like Preprocessing, Feature extraction, Training and Testing. Preprocessing consists of grayscale conversion, binarization and thinning. Wavelet decomposition is used for Feature extraction. For classification BPNN is used.The proposed system has obtained a maximum recognition rate of 89%.

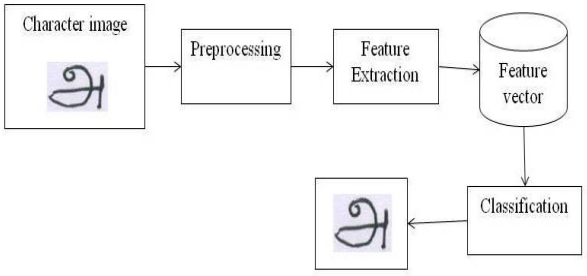


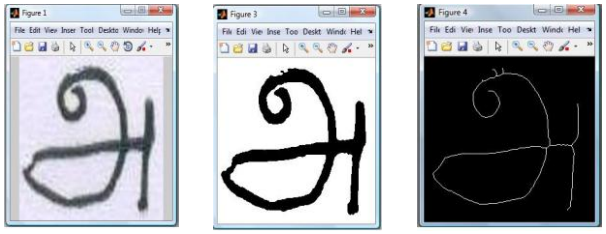
Fig 3: Proposed Model of OCR system

Table 1. Wavelet decomposition

Wavelet decomposition level	Wavelet
Level 1	db1
Level 2	db1
Level3	db1
Level3	db2
Level4	db2

4.1 Preprocessing Techniques

As an initial step, data set is collected from 100 different persons with pen and ink variations. The data samples are scanned at 300 DPI and stored at JPG format. Each character is segmented using morphological structuring element and bounding boxes. Suitable filters are used to reduce various noises in the segmented characters. Input image is converted into gray scale. Grayscale Image is converted into Binarized Image. For Binarization, Otsu's method of global thresholding is used. Binarized Image is converted into Thinned Image for applying wavelet transform.



a **b** **c**

Fig 4: a) Gray Image b)Binarized Image c)Thinned Image

4.2 Feature Extraction Scheme

Feature extraction is done by using wavelet decomposition. Wavelet decomposition at various levels is done by using different wavelet families like 'db1' and 'db2'. Wavelet decomposition at several levels are varying from Level 1 to Level 4. Features are extracted for various wavelet families and those features are normalized between [-1 1].

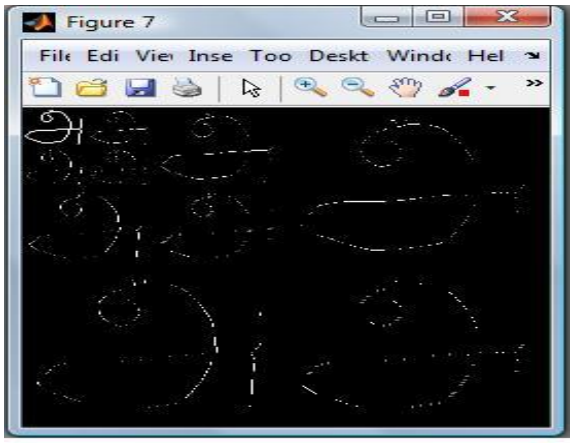


Fig 5: Level 3 wavelet decomposition using 'db1'

4.3 Classification and Recognition

For neural network training and testing, normalized features of feature vector is used.A three layer feed forward back propagation neural network with sigmoid activation function is used for classification.The neural network model is given in Fig.The neural network is trained with gradient descent backpropagation algorithm.Mean Squared error is used as the performance measure.

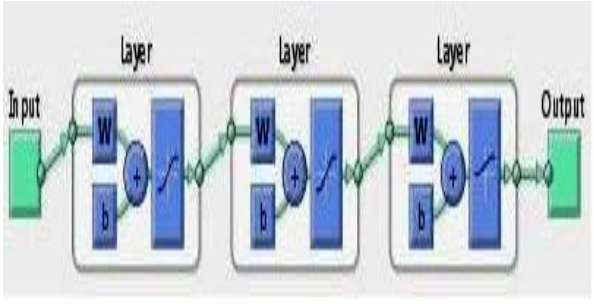


Fig 6: Neural network model

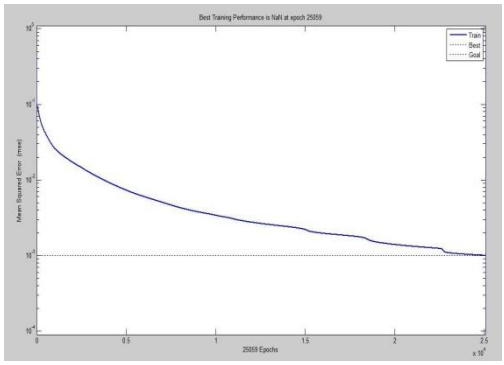


Fig 7: Performance plot -Level 3 wavelet decomposition using 'db1' features

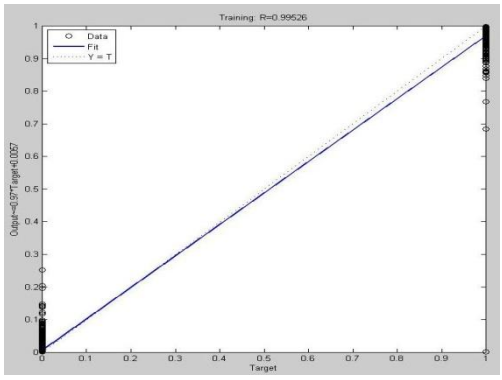


Fig 8: Regression plot -Level 3 wavelet decomposition using 'db1' features

5. RESULTS AND DISCUSSIONS

75% of data samples is used for training.Each class is having 100 samples collected from 100 different people.From the total samples of each class, 75 random samples were used for training.25% of data is used for testing.

Table 2. Results comparision

Method	Recognition Rate
Level 2 db1	84%
Level 3 db1	86%
Level 3 db2	87%
Level 4 db2	89%

6. CONCLUSION AND FUTURE WORK

A system for offline Tamil handwritten character recognition was developed using wavelet transform. By increasing the level of wavelet decomposition, more efficient features for each character were extracted and the system was trained using back propagation neural networks (BPNN) and obtained a maximum accuracy rate of 89% for 31 class problem.

In the future, the system can be modified using better feature extraction methods and classifiers like KNN classifier, Support Vector Machines (SVM) and Extreme Learning Machines (ELM). So the character recognition system will get more accurate results.

7. REFERENCES

[1] T. Reiss, "Recognizing Planar Objects Using Invariant Image Features", Springer-Verlag, New York, Inc. Secaucus, NJ, USA, 1993. .

[2] S. Impedove, L. Ottaviano, and S. Occhinegro, "Optical character recognition – a survey", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 5 (1-2), pages.1-24, 1991.

[3] C. C. Tappert, C. Y. Suen, and T. Wakahara, "State of the art in online hand-writing recognition" ,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, No.8, August, 1990.Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] "Code: a system in which arbitrary values are given to letters, words, numbers or symbols to ensure secrecy or brevity,"http://homepages.cwi.nl/~dik/english/codes/charrec.html.

[5] T. Pavdilis and Z. Jiangying, "Page segmentation and classification," CVGIP:Graphical models and image processing, Vol.54, No.6, 1992.Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[6] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey," IEEE Transactions on Pattern Analysis and Machine Vision, Vol.22, No.1, January 2002.

[7] G. Tauschek, "Reading machine," U.S. Patent 2026329, December 1935.

[8] Siromoney et al., 1978. Computer Recognition of Printed Tamil Character, Pattern Recognition 10: 243-247.

[9] Chinnuswamy, P., and S.G. Krishnamoorthy, 1980. Recognition of Hand printed Tamil Characters, Pattern Recognition, 12: 141-152.

[10] Suresh et al., 1999. Recognition of Hand printed Tamil Characters Using Classification Approach, ICAPRDT' 99, pp: 63-84.

[11] Hewavitharana, S, and H.C. Fernando, 2002. "A Two-Stage Classification Approach to Tamil Handwriting Recognition", pp: 118-124, Tamil Internet 2002, California, USA.

[12] N. Shanthi and K. Duraiswamy, "Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM", Department of Information Technology, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, India.

[13] Shivsubramani K, Loganathan R, Srinivasan CJ, Ajay V, Soman KP, "Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters", Centre for Excellence in Computational Engineering, Amrita Vishwa Vidyapeetham, Tamilnadu, India.