Original Research Paper

VOLUME-6 | ISSUE-4 | APRIL - 2017 • ISSN No 2277 - 8179 | IF : 4.176 | IC Value : 78.46

Handwritten Recognition of Tamil Scripts Using HMM With Fuzzy Logic

Engineering

KEYWORDS: (Fuzzy Logic, Character Recognition, *Fuzzy Set Theory, Membership FunctioN)*

L.G.X.Agnel Livingston — Asst. Professor, CSE Dept., St.Xavier's Catholic College of Engineering, Nagercoil , India

L.M.Merlin Livingston — Associate Professor, ECE Dept., Jeppiaar Institute Of Technology,Chennai , India

ABSTRACT

Ancestors recorded and used the information in various ways of documents. The handwritten text decayed over a period of the time. It is very difficult to preserve them in the same form. It is a need to preserve these olden valuable data's and they should be converted to computerized Unicode. In this paper the Hidden Markov Model is combined with the fuzzy approach in recognition of Handwritten Tamil characters. Identified unknown characters are classified as one among the prototype characters using features distance from the frame and a suitable membership function. The unknown and prototype characters are pre-processed and considered for recognition. The algorithm is tested for about 750 samples of Tamil characters and the success rate obtained varies from 89% to 93%.

I. INTRODUCTION

HANDWRITING processing is a domain in great expansion. The interest devoted to this field is not explained only by the exciting challenges involved, but also the huge benefits that a system, designed in the context of a commercial application, could bring .Two classes of recognition systems are usually distinguished: online systems for which handwriting data are captured during the writing process, which makes available the information on the ordering of the strokes, and offline systems for which recognition takes place on a static image captured once the writing process is over. The field of personal computing has begun to make a transition from the desktop to handheld devices, thereby requiring input paradigms that are more suited for single hand entry than a keyboard. Online handwriting recognition allows for such input modalities. On-line handwritten scripts are usually dealt with pen tip traces from pen-down to pen-up positions. There is extensive work in the field of handwriting recognition, and a number of reviews exist. General methodologies in pattern recognition and image analysis are presented in. Character recognition is reviewed in for off-line recognition, and in for on-line recognition. Most of the researchers have chosen numeric characters for their experiment. So, some maturity can be observed for isolated digit recognition. However, when we talk about the recognition of Tamil characters, results a complicated process in character recognition

Handwritten character recognition refers to the process of conversion of handwritten character into Unicode. Among different branches of character recognition it is easier to recognize English alphabets and numerals than Tamil characters. In this Tamil character recognition process, the most difficult problem is the great variation among handprints. Hence only a few selected Tamil characters are considered in this work. Here we use fuzzy concept as a tool for recognition of handwritten Tamil Characters. The feature vector consists of distances of the pattern from the frame in sixteen different directions. Actually, two sets are considered for recognition, one with segments and the other with membership values.

II. HANDWRITTEN TAMIL CHARACTER RECOGNITION

Character recognition gets complicated by differences such as multiple patterns to represent a single character, cursive representation of letters, and the number of disconnected and multi–stroke characters. Few researches have addressed this complicated subject. In fact, it can be said that character recognition still an open problem. Neural Nets (NN) and Hidden Markov Models (HMM) are the popular, amongst the techniques which have been investigated for handwriting recognition. It has been observed that NNs in general obtained best results than HMMs, when a similar feature set is applied. The most widely studied and used neural network is the Multi-Layer Perceptron (MLP). Such an architecture trained with back-propagation is among the most popular and versatile forms of neural network classifiers and is also among the most frequently used traditional classifiers for handwriting recognition. Other architectures include Convolutional Network (CN), Self-Organized Maps (SOM), Radial Basis Function (RBF), Space Displacement Neural Network (SDNN), Time Delay Neural Network (TDNN), Quantum Neural Network (QNN), and Hopfield Neural Network (HNN). Few attempts have been found in the literature in which counter-propagation (CPN) architecture has been used for the recognition of handwritten characters. The main objective of this work is the implementation of Hidden Markov Model (HMM) with Fuzzy approach.

Using Fuzzy theory, the badness in handwritten characters can be interpreted directly as a fuzzy membership function representing the degree to which the actual pattern is a member of a fuzzy set Line, or Arc. The scanned image is segmented into paragraphs using spatial space detection technique, paragraphs into lines using vertical histogram, lines into words using horizontal histogram. Each image glyph is subjected to feature extraction procedure, which extracts the statistical and syntactical features such as character height, character width, number of horizontal lines, number of vertical lines, the horizontally oriented curves, the vertically oriented curves, number of slope lines, image centroid and special dots of the image glyph. Then these classes are mapped onto Unicode for recognition and the text is reconstructed using Unicode fonts. The concept of Fuzzy set was introduced by L.A. Zadeh [3,9]

Recognition of handwritten letters is a very complex problem. The letters could be written in different size, orientation, thickness, format and dimension. These will give infinity variations. The capability of neural network to generalize and be insensitive to the missing data would be very beneficial in recognizing handwritten letters. The proposed Tamil handwritten character recognition system uses a neural network based approach to recognize the characters represented by scale and shift invariant features. Feed forward Multi Layered Perceptron (MLP) network with one hidden layer trained using back-propagation algorithm has been used to recognize handwritten Tamil characters.

III. SYSTEM ARCHITECTURE

In this paper, Fuzzy technique to identify between Tamil scripts was proposed. The proposed framework can be divided into three stages according to the framework model, which consists of the preprocessing stage, Extraction using Hidden Markov Model and the retrieval stage which uses Fuzzy Logic. The design of our script identification system is illustrated in figure 1. Details of each stage are discussed in the following subsections.

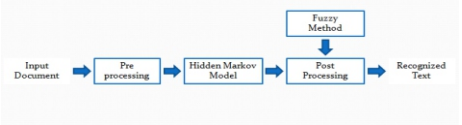

Fig:1 Block Diagram Represents Hidden Markov Model with Fuzzy Logic approach in Post Processing

A. Preprocessing

The most popular class of nonlinear vector operators in handwritten recognition was based on the order statistics, where the output is equal to the vector associated with the smallest accumulated distance to all other vectors in the sliding window or which is most similar to all neighboring pixels. These classical filters are efficient in removing outliners but also blur image details due to their pure order statistics approaches. 1) weighted vector filtering techniques which utilize the local spatial relationship of the samples inside the supporting window 2) switching schemes so that the filter is only applied to pixels corrupted with impulse noise and 3) fuzzy-based-approaches in order to distinguish between noise and image characteristics.

Preprocessing includes the steps that are necessary to bring the input data into an acceptable form for feature extraction. The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing stages. Preprocessing stage involves noise reduction, slant correction, size normalization and thinning. Among these size normalization and thinning are very important. Normalization is required as the size of the numeral varies from person to person and even with the same person from time to time. The input numeral image is normalized to size 50x50 after finding the bounding box of each handwritten numeral image. Thinning provides a tremendous reduction in data size, thinning extracts the shape information of the characters. It can be considered as conversion of off-line handwriting to almost on-line data. Thinning is the process of reducing thickness of each line of pattern to just a single pixel. In this research work, we have used morphology based thinning algorithm for better symbol representation. Figure 2 shows the steps involved in the method as par preprocessing is considered.

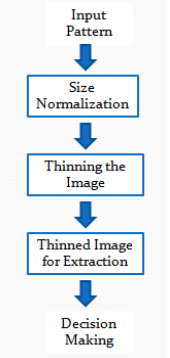

Fig:2 Block Diagram Represents Preprocessing

B. Hidden Markov Model

Hidden Markov Models (HMM) are suitable for handwriting recognition for a number of reasons. The importance of HMMs in the area of speech recognition has been observed several years ago. In the meantime, HMMs have also been successfully applied to image pattern recognition problems such as shape classification and face recognition. HMMs qualify as suitable tool for cursive script recognition for a number a reasons. A Hidden Markov Model is a doubly stochastic model. The underlying stochastic process corresponds to state transitions that are hidden, but the state changes are observed through another set of stochastic processes that produce the output characters. The output character is said to be discrete if it is from a finite alphabet, and it is continuous if it has real-valued attributes. Accordingly, the model is called discrete or continuous HMM. In this experiment, continuous HMMs were used to model the

Tamil characters since the features are real-valued. The most commonly used HMM topology for both speech and handwriting is the left-to-right model, also known as the Bakis model. An HMM state is said to generate feature vectors following a probabilistic distribution, usually a mixture of Gaussians. The number of Gaussians in the mixture and the number of states in the HMM were determined empirically. HMM training was done using the well known Baum-Welch re-estimation procedure. In this experiment a different handwriting samples of HMMs corresponding to different character classes were trained. Given a test character, the probability associated with each one of the character- HMMs was computed and the character that has the maximum probability is declared was the recognition result. The probability associated with each character was computed using the HMM forward algorithm.

Hidden Markov Models are increasingly being used to model sub strokes of characters. However, the published work employing HMMs is limited, and generally focused on isolated character recognition. In this best result obtained from lexicon-driven approach where character hidden Markov models are concatenated to obtain word hidden Markov model that is matched with input word image the features extracted from the symbols were used to train a continuous density HMM for each symbol. For modeling a symbol using HMM, a simple left-to-right topology with no state skipping was adopted and the training was carried out using the Baum-Welch re-estimation procedure.

The number of states per model was determined based on the shape complexity of the symbol and this has been shown to model the symbols better than having a fixed number of states for each symbol. The number of states was computed as a fraction of average length of the training observation sequences of the symbol. The fraction was empirically determined as 0.2, and similarly the number of Gaussians per state were set to two.

C. Fuzzy Method

Fuzzy concept on handwritten Tamilcharacters classifies them as one among the prototype characters using a feature called distance from the frame and a suitable membership function. The prototype characters are categorized into two classes: one is considered as line characters/patterns and the other as arc patterns. The unknown input character is classified into one of these two classes first and then recognized to be one of the characters in that class.

IV. EXPERIMENTAL RESULTS

A. Database

The training data consists of 750 samples of Tamil character, out of which 180 patterns are machine printed and the remaining are handwritten elements including both cursive and block handwriting besides signatures, dates and address locations. Our test data consists of 250 Tamil document images, scanned at 300 dpi and stored in 1-bit depth monochrome format. These documents contain handwritten elements, signatures, logos along with free flowing text paragraphs. The testing also contains documents which have only signatures as part of their handwritten components. It is observed that higher accuracy was obtained on such type of documents.

B. Evaluation

Upon closer examination of the results, one letter has been confused with the three scripts. These three scripts have the same writing direction and they have similar stroke length and density, which could explain the misclassifications. Comparisons with the work on the identification of handwritten Arabic and Roman scripts by Ben-Jlaiel et al. show that our proposed methodology outperforms them. They attained an accuracy of 93% handwritten scripts. Another comparison that is closer to our work on online handwritten script identification.

CONCLUSION

From the results it can be concluded that Fuzzy is a good promise in terms of recognition capability which has not been put on trial in the field of handwriting recognition. More over Fuzzy is more economical than convergence of other NN architectures e.g. back-propagation where the training time can take long time. The

experiments provided the authors an opportunity to explore this pattern recognition methodology; the exercise provided a theoretical base for further investigations and impetus for development work in this discipline. The obtained results motivate the continuity of the system development considering a preprocessing mechanism including normalization and slant removal. Other future work might involve some new feature extraction approaches.

REFERENCES

[1] A.K.Dutta,"An experimental procedure for Handwritten character recognition",IEEE Tr. On Computer vol c- 23,no5, pp 536-545, May 1974.
[2] E.T. Lee, and L.A. Zadeh, "Note on Fuzzy Languages", Info. Sc. 1, pp. 421-434, 1969.
[3] L.A. Zadeh, "Fuzzy Languages and their relation to Human and Machine Intelligence", Man and Computer, Proce. Int. Conf. Bordeaux 1970, pp. 130-165.
[4] L.A. Zadeh,"Quantitative Fuzzy Semantics", Info. Sc. 3, pp.159-176, 1971.
[5] M.Shimura,"Application of fuzzy sets Theory to Pattern Recognition", J. JAACE, 43,243,1975
[6] M.M. Gupta,G.N Sardis and S.R Gaines,"Fuzzy Automata and Decision Process",North Holland, New York, 1977.
[7] P. Siy and C.S. Chen,"Minimization of Fuzzy Functions", IEEE Tr. on Comp. c-21,100,1972.
[8] Siromoney,and et al,"Computer Recognition of printed Tamil Character",Patt. Rec. vol 10, 243-247, 1978.
[9] L.A. Zadeh, " Fuzzy Sets", Info. Cont. 8,pp. 338-353, 1965.
[10] S.K. Pal et al, " Fuzzy Set in Handwritten Character Recognition", Recent Develop. in Pattern recognition and Digital Techniques, ISI, Cal. pp. 63-71, 1977.
[11] E.T. Lee, " Fuzzy Tree Automata and Syntactic Pattern recognition", IEEE PAMI 4,4,458-462, 1982.
[12] M.G. Thomsan," Finite Fuzzy Automata, Regular Fuzzy Languages, and Pattern Recognition", PR,5, pp. 383-390,1973.
[13] P. Siy and C.S. Chen,"Fuzzy Logic for handwritten Numerical Character recognition", IEEE SMC 1,1,pp. 61- 66,1971
[14] W.J.M. Kickert, and H.Kopelaar,"Application of Fuzzy Set theory to Syntactic pattern recognition", IEEE SMC 6,pp. 148-151, 1976.
[15] R.M.Suresh&S.Arumugam,"Fuzzy Context-free Grammar to Handwritten Numerical Recognition CSI`98¸ Proceeding of Annual Convention of CSI, Sept16-20, 411-419,1998
[16] RM. Suresh,"Fuzzy Context-free Grammar, Fuzzy Tree Automata, and Syntactic Pattern Recognition CSI` 96 Proceedings of Annual Convention on CSI Oct. 30 th Nov. 3,499-502,1996.
[17] N.V.Subba Reddy and P.Nagabhusan,"A Three-Dimensional Neural Network Model for Unconstrained Handwritten Numeral Recognition: A New Approach", Pattern Recognition 31(5), 511-516, 1998. [18] I.K.Sethi and B.Chatterjee, "Machine Recognition of constrained handprinted Devanagari",Pattern Recognition 9, 69 ,1977.
[18] B.N.Chatterjee, "Fuzzy Set Theory to recognize Handwritten Characters", Proceeding edited by D.Dutta Majumder of Pattern Recognition and Digital Techniques conference held at ISI, Calcutta, pp 166-172, 1982.
[19] V.K.Govindan and A.P.Shivaprasad,"Character Recognition – A Review", Pattern Recognition 23, No. 7, pp671- 683,1990..
[20] P.Chinnuswamy and S.G.Krishnamoorthy, "Recognition of Handprinted Tamil Characters", Pattern Recognition Vol. 12, pp141-152, 1980