Off-Line Cursive Handwritten Tamil Character Recognition

R.JAGADEESH KANNAN and R.PRABHAKAR
RMK Engineering College
Chennai
INDIA
rjagadeeshkannan@yahoo.com

Abstract: - In spite of several advancements in technologies pertaining to Optical character recognition, handwriting continues to persist as means of documenting information for day-to-day life. The process of segmentation and recognition pose quiets a lot of challenges especially in recognizing cursive hand-written scripts of different languages. The concept proposed is a solution crafted to perform character recognition of hand-written scripts in Tamil, a language having official status in India, Sri Lanka, and Singapore. The approach utilizes discrete Hidden Markov Models (HMMs) for recognizing off-line cursive handwritten Tamil characters. The tolerance of the system is evident as it can overwhelm the complexities arise out of font variations and proves to be flexible and robust. Higher degree of accuracy in results has been obtained with the implementation of this approach on a comprehensive database and the precision of the results demonstrates its application on commercial usage. The methodology promises to present a simple and fast scaffold to construct a full OCR system extended with suitable pre-processing.

Key-Words: - Optical Character Recognition (OCR), Cursive Script Recognition, Handwritten Script Recognition, Segmentation, Offline Recognition, Hidden Markov Model (HMM)

1 Introduction

Character recognition can solve more complex problems and help ease the drudgery involved in maintaining obscure image files. Basically converting scanned images in to text document can enable manipulation through word processing applications. Optical Character Recognition has gained a momentum since the need for digitizing or converting scanned images of machine printed or hand written text (numerals, letters, and symbols), in to a format recognized by computers (such as ASCII). OCR has been extensively used as the basic application of different learning methods in machine learning literature [1]. Handwriting recognition is the task of transforming a language re-presented in its own spatial form of graphical marks into a symbolic representation [2]. Handwriting recognition inherited a number of technologies from optical character recognition (OCR). The main difference between handwritten and typewritten characters is in the variations that come with handwriting. It is also worth noticing that OCR deals with off-line recognition while handwriting recognition may be required for both on-line and off-line signals.

Handwriting recognition is one of the very challenging problems. Traditionally the field of handwriting recognition is divided into off-line and on-line recognition [2]. In off-line recognition, only the image of the handwriting is available for the computer, while in the on-line case temporal information such as pentip coordinates as a function of time is also available. Typical data acquisition devices for off-line and on-line recognition are scanners and digitizing tablets, respectively. Due to the lack of temporal information, off-line handwriting recognition is considered more difficult than on-line. Furthermore, it is also clear that the off-line case is the one that corresponds to the conventional reading task performed by humans [3].

The need for OCR arises in the context of digitizing Tamil documents from the ancient and old era to the latest, which helps in sharing the data through the Internet [5]. Tamil, the native language of a southern state in India has several million speakers across the world and is an official language in countries such as Sri Lanka, Malaysia and Singapore. The penetration of Information Technology (IT) becomes harder in a country such as India where the majority people read and write in their native language. Therefore, enabling interaction with computers in the native language and in a natural way such as handwriting is absolutely necessary [4].

In the literature, many papers have been published with research detailing new techniques for the classification of handwritten characters and

words. Some researchers have obtained very promising results for isolated/segmented characters using neural network based techniques [6]. However, the results for the segmentation and recognition of touching handwritten characters have not been very good and still there is a need for improvement so that they can be used in real world applications. In this paper we propose a complete off-line OCR system for cursive handwritten Tamil characters. The scanned document image is preprocessed to ensure that the characters are in a suitable form. Then the line, word and characters are segmented and features are extracted from the segmented characters. Finally Hidden Markov Models are used for the training of extracted features and the recognition of characters.

The remainder of the paper is organized as follows: Section 2 briefly reviews the prior work on offline recognition of Handwritten-characters. Section 3 briefly describes about the Tamil Script. The proposed system architecture is given in Section 4. Section 5 describes the preprocessing and feature extraction stages of the system proposed. Training and Recognition using HMMs is explained in Sections 6. The results of our experiment are given in Section 7 and conclusions are mentioned in Section 8.

2 Related Work

Yann Lecun et al. [1] discussed various methods applied to handwritten character recognition and compare them on a standard handwritten digit recognition task and described two systems for online handwriting recognition that experiments demonstrate the advantage of global training, and the flexibility of graph transformer networks.

Rejean Plamondon and Sargur N.Srihari [2] describes the nature of handwritten language, how it is transduced into electronic data and the basic concepts behind written language recognition algorithms and they indicate that algorithms for preprocessing, character and word recognition, and performance with practical systems.

Tam´as Varga [3] investigate the generation and use of synthetic training data in terms of the problem of off-line cursive handwritten text line recognition and they examine whether the recognition performance can be improved by expanding the natural training set using synthetically generated text lines, because the automatic generation of training data is much faster and cheaper than collecting additional human written samples.

Bharath A and Sriganesh Madhvanath [4] have proposed a data-driven HMM-based online handwritten word recognition system for Tamil, an Indic script and they discussed a writer-independent online handwritten Tamil word recognition system that employs HMMs for word modeling

Seethalakshmi R. et al. [5] discusses the various strategies and techniques involved in the recognition of Tamil text and they refers Optical Character Recognition (OCR) for the process of converting printed Tamil text documents into software translated Unicode Tamil Text and their extracted features are passed to a Support Vector Machine (SVM) where the characters are classified by Supervised Learning Algorithm.

Sung-Bae Cho [6] present three sophisticated neural network classifiers to solve complex pattern recognition problems: multiple multilayer perceptron (MLP) classifier, hidden Markov model (HMM)/MLP hybrid classifier, and structure adaptive self-organizing map (SOM) classifier and their three methods have produced 97.35%, 96.55%, and 96.05% of the recognition rates, respectively, which are better than those of several previous methods reported in the literature on the same database.

K.H.Aparna et al. [7] gives a complete OCR system for Tamil newsprint that includes the full suite of processes from skew correction, binarization, segmentation, text and non-text block classification, line, word and character segmentation and character recognition to final reconstruction and their approach gave reasonable segmentation results with the class of document images chosen in the present work.

Khalaf khatatneh et al. [8] proposes a new technique assists in developing a recognition system for handling the Arabic Hand Written text named Arabic Hand Written Optical Character Recognition (AHOCR) that concerned with recognition of hand written Alphanumeric Arabic characters and their final results indicate and clarify that the proposed AHOCR technique achieves an excellent test accuracy of recognition rated up to 97% for isolated Arabic characters and 96% for Arabic text.

H.Bunke et al. [10] have developed a method for the off-line recognition of cursive handwriting based on Hidden Markov Models and their letter models are presently insensitive to their surrounding contexts, and no attempts have been no made to build individual models for significantly different realizations of the same letter.

Lawrence R.Rabiner [11] review the theoretical aspects of Markov Source or hidden Markov modeling type of statistical modeling and show how

they have been applied to selected problems in machine recognition of speech.

Ninad Thakoor and Jean Gao [12] conversed a novel two step shape classification approach consisting of a description and a discrimination phase that the description phase, curvature features are extracted from the shape and are utilized to build a Hidden Markov Model (HMM) and they present classification results achieved for fighter planes in terms of classification accuracy and discriminant functions.

F.Samaria and F.Fallside [13] have proposed a new approach to the automatic face identification involving in the use of Hidden Markov Models and they illustrate how these models allow the automatic extraction of facial features and the classification of face images.

S. Hewavitharana et al. [16] developed a method to recognize off-line handwritten Sinhala characters, the language used by the majority of Sri Lanka based on discrete hidden Markov models and their feature extraction method reduced the two dimensional spatial information of character images into a single dimension array of values, thereby throwing away some information.

S. Hewavitharana and H. C. Fernando [17] described a system to recognize handwritten Tamil characters using a two-stage classification approach, for a subset of the Tamil alphabet and their first stage, an unknown character is pre-classified into one of the three groups: core, ascending and descending characters, then, the second stage, members of the pre-classified group are further analyzed using a statistical classifier for final recognition.

Dhanya D and A G Ramakrishnan[23] suggested Recognition via script identification and Bilingual approach methods for character recognition from bilingual text and that first one uses the knowledge of the script for identification depends on the efficiency of the individual OCRs ad also in the second approach they has been suggested a hierarchical classification scheme.

K.H.Aparna and V.S.Chakravarthy [24] presents a complete OCR system for Tamil newsprint that includes the full suite of processes from skew correction, binarization, segmentation, text and non-text block classification, line, word and character segmentation and character recognition to final reconstruction and their approach gave reasonable segmentation results with the class of document images chosen in the present work.

Shivsubramani K et al. [25] gives an efficient method for recognizing printed Tamil characters exploring the interclass relationship between them

and they accomplished using Multiclass Hierarchical Support Vector Machines, a new variant of Multi Class Support Vector Machine which constructs a hyperplane that separates each class of data from other classes.

B.M. Sagar, Dr. Shobha G, Dr. Ramakanth Kumar P [26] describes an Optical Character Recognition (OCR) system for printed text documents in Kannada, a South Indian language. The proposed OCR system for the recognition of printed Kannada text, which can handle all types of Kannada characters.

Hani Khasawneh [27] present a new and novel Arabic optical character recognition (AOCR) system. The system accepts a scanned-page image containing a set of text lines, and affected by typical noise level. The system preprocesses the image before the separate lines are handled to the character extraction phase. Our approach depends on segmenting the line into tokens and extracting characters features in order to recognize the characters and to assemble the output line of text. A neural network is used to process the features and to classify them into one of the characters.

Nadia Ben Amor And Najoua Essoukri Ben Amara [28] Pattern Recognition Is A Well-Established Field Of Study And Optical Character Recognition (OCR) has long been seen as one of its important contributions. In this paper we describe the performances of a hybrid classification approach which combines both neural networks and hidden Markov models. This classification technique is dealing with features extracted through the wavelet transform method.

3 Tamil Script

Tamil is the root of all Dravidian languages like Malayalam, Telugu, Kanadam and etc, essentially the mother tongue of the Tamil people and spoken predominantly by the people of Tamilnadu. The language has gained official status in India, Srilanka, and Singapore. Considerable percentage of minorities in Malaysia, Mauritius, and Reunion, as well as emigrant communities around the world also speaks Tamil [20]. Tamil is the first language to be conferred as a classical language by the government of India in 2004, followed by Sanskrit [21], [22]. The language is also the administrative language of the Indian state of Tamilnadu. Consisting symbols for vowels and consonants, Tamil scripts belong to the family of syllabic alphabets [4]. Using special diacritical marks known as mantras, the vowels implicitly present in the consonants can be modified to other vowels. The implicit vowel sound can be

eliminated in a consonant by changing to its half form using the vowel-muting diacritic. Referred here as a syllabic unit, a composite character, is obtained by combining a vowel and a consonant. Even though the Tamil Alphabets don't have much of cursive nature, the cursive nature seems to exist in the speed writing, writing style of individuals, old literatures in Tamil written in OlaiChuvadi, Stone scriptures etc.

Including 12 pure vowels and 23 pure consonants the language consist 156 unique symbols/characters. The combined sum of pure vowels and pure consonants, the 35 characters, is the basic character units of the script and the remaining categories of characters are vowel-consonant combinations. The basic character and the modifier symbol corresponding to the basic character are the two parts that primarily form the vowel-consonant combination. Figure 1 is the representation of basic character set in Tamil.
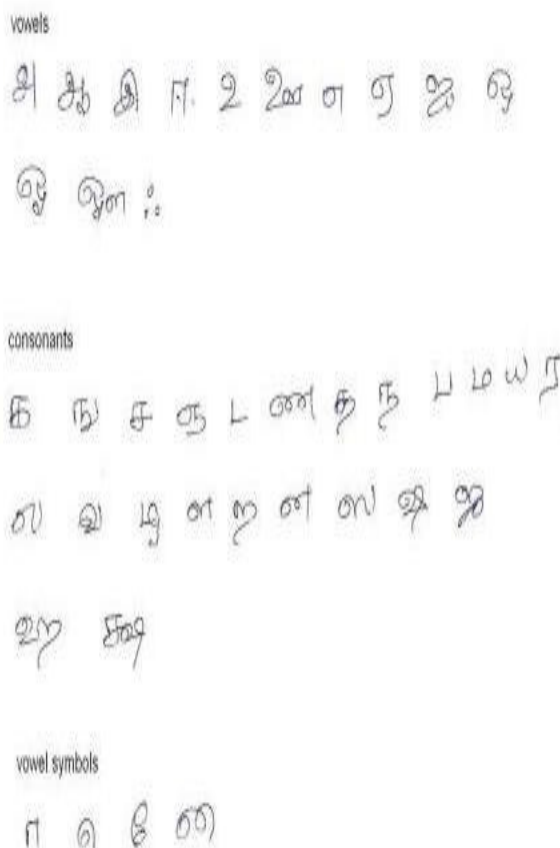

Figure 1: Basic Tamil Character Set

4. The Proposed System Architecture

Broadly, off-line handwritten character recognition system includes three stages: image preprocessing, feature extractor, and classifier. The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown character into one of the known classes. Preprocessing is primarily used to reduce variations of handwritten characters. A feature extractor is essential for efficient data representation and extracting meaningful features for later processing. A classifier assigns the characters to one of the several classes. The architecture of the proposed system is shown in Figure 2
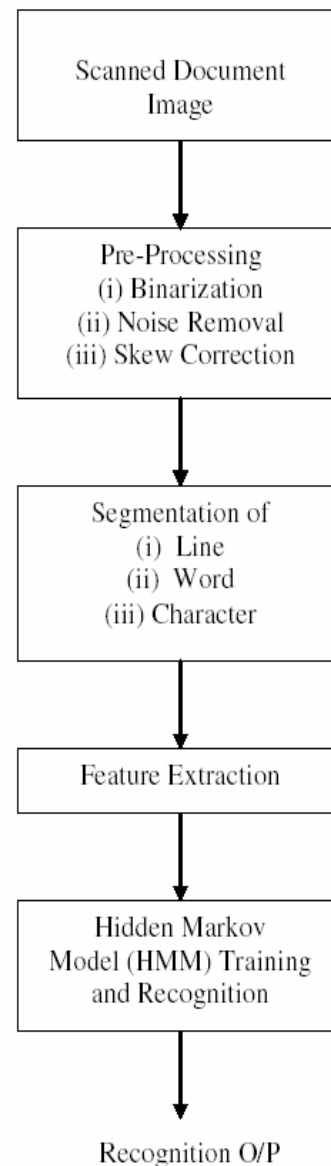

Figure 2: Architecture of the Proposed System

5. Recognition Methodology

The proposed research methodology for off-line cursive handwritten Tamil characters is described in this section.

5.1 Preprocessing

There exist a whole lot of tasks to complete before the actual character recognition operation is commenced. These preceding tasks make certain the scanned document is in a suitable form so as to ensure the input for the subsequent recognition operation is intact. The process of refining the scanned input image includes several steps that include: Binarization, for transforming gray-scale images in to black & white images, scraping noises, Skew Correction- performed to align the input (usually paper document) with the coordinate system of the scanner and etc., The preprocessing stage comprise three steps:

    (1) Binarization
    (2} Noise Removal
    (3) Skew Correction

5.1.1 Binarization

Extraction of foreground (ink) from the background (paper) is called as thresholding. Typically two peaks comprise the histogram gray-scale values of a document image: a high peak analogous to the white background and a smaller peak corresponding to the foreground. Fixing the threshold value is determining the one optimal value between the peaks of gray-scale values [2]. Each value of the threshold is tried and the one that maximizes the criterion is chosen from the two classes regarded as the foreground and back ground points. The scanned document image and its binarized output are shown in Figure 3 and 4 respectively.
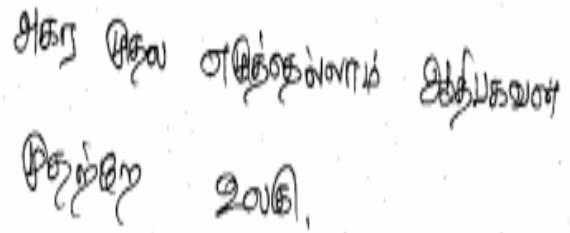

Figure 3: Scanned Document Image

Figure 4: Binarized Image

5.1.2 Noise Removal

The presence of noise can cost the efficiency of the character recognition system; this topic has been dealt extensively in document analysis for typed or machine-printed documents. Noise may be due the poor quality of the document or that accumulated whilst scanning, but whatever is the cause of its presence it should be removed before further processing. We have used median filtering and wiener filtering for the removal of the noise from the image [8]. The noise removed image is shown in Figure 5.
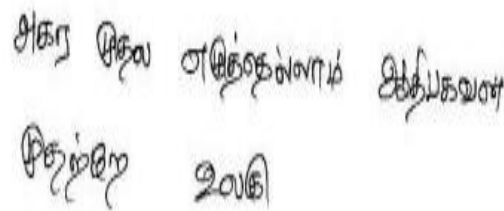

Figure 5: Noise Removed Image

5.1.3 Skew Correction

Aligning the paper document with the co-ordinate system of the scanner is essential and called as skew correction. There exist a myriad of approaches for skew correction covering correlation, projection, profiles, Hough transform and etc. For skew angle detection Cumulative Scalar Products (CSP) of windows of text blocks with the Gabor filters at different orientations are calculated [7]. Orientation with maximum CSP gives the skew angle. Alignment of the text line is used as an important feature in estimating the skew angle. We calculate CSP for all possible 50X50 windows on the scanned document image and the median of all the angles obtained gives the skew angle [7].

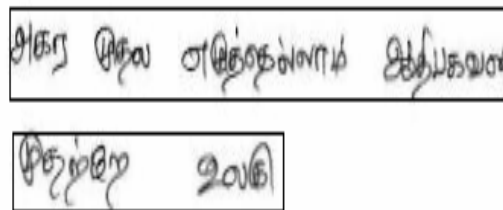Figure 6: Image before Skew Correction

Figure 8: Line Segmented Image

5.2.2 Word And Character Segmentation

The process of word segmentation succeeds the line separation task. Most of the word segmentation issues usually concentrate on discerning the gaps between the characters to distinguish the words from one another other. This process of discriminating words emerged from the notion that the spaces between words are usually larger than the spaces between the characters. There are not many approaches to word segmentation issues dealt in the literature. In spite of all these perceived conceptions, exemptions are quiet common due to flourishes in writing styles with leading and trailing ligatures. Alternative methods not depending on the one-dimensional distance between components, incorporates cues that humans use. Meticulous examination of the variation of spacing between the adjacent characters as a function of the corresponding characters themselves helps reveal the writing style of the author, in terms of spacing. The segmentation scheme comprises the notion of expecting greater spaces between characters with leading and trailing ligatures [3].

Recognizing the words themselves in textual lines can itself help lead to isolation of words. Segmentation of words in to its constituent characters is touted by most recognition methods. Features like ligatures and concavity are used for determining the segmentation points. The algorithm exploits the caps between character segments and heights of character segments too.

Figure 7: Image after Skew Correction

5.2 Segmentation

Segmentation is a process of distinguishing lines, words, and even characters of a hand written or machine-printed document, a crucial step as it extracts the meaningful regions for analysis. There exist many sophisticated approaches for segmenting the region of interest. Straight-forward, may be the task of segmenting the lines of text in to words and characters for a machine printed documents in contrast to that of handwritten document, which is quiet difficult. Examining the horizontal histogram profile at a smaller range of skew angles can accomplish it. The details of line, word and character segmentation are discussed as follows.

5.2.1 Line Segmentation

Obviously the ascenders and descenders frequently intersect up and down of the adjacent lines, while the lines of text might itself flutter up and down. Each word of the line resides on the imaginary line that people use to assume while writing and a method has been formulated based on this notion. The local minima points are calibrated from each component to approximate this imaginary baseline. To guesstimate and categorize the minima of all components and to recognize different handwritten lines clustering techniques are deployed [1].

Figure 9: Word Segmented Image

5.3 Feature Extraction

The size inevitably limited in practice, it becomes essential to exploit optimal usage of the information stored in the available database for feature extraction. Thanks to the sequence of straight lines, instead of a set of pixels, it is attractive to represent character images in handwritten character recognition. Whilst holding discriminant information to feed the classifier, considerable reduction on the amount of data is achieved through vector representation that stores only two pairs of ordinates replacing information of several pixels.

Vectorization process is performed only on basis of bi-dimensional image of a character in off-line character recognition, as the dynamic level of writing is not available. Reducing the thickness of drawing to a single pixel requires thinning of character images first.

Figure 10: Character before and after Thinning

After streamlining the character to its skeleton, entrusting on an oriented search process of pixels and on a criterion of quality of representation goes on the vectorization process. The oriented search process principally works by searching for new pixels, initially in the same direction and on the current line segment subsequently. The search direction will deviate progressively from the present one when no pixels are traced. The dynamic level of writing is retrieved of course with moderate level of accuracy, and that is object of oriented search.

Starting the scanning process from top to bottom and from left to right, the starting point of the first line segment, the first pixel is identified. According to the oriented search principle, specified is the next pixel that is likely to be incorporated in the segment. Horizontal is the default direction of the segment considered for oriented search. Either if the distortion of representation exceeds a critical threshold or if the given number of pixels has been associated with the segment, the conclusion of line segment occurs. Computing the average distance between the line segment and the pixels associated with it will yield the distortion of representation.

The sequence of straight lines being represented through ordinates of its two extremities character image representation is streamlined finally. All the ordinates are regularized in accordance to the initial width and height of character image to resolve scale variance.

6 Hidden Markov Model

Hidden Markov Models are suitable for handwriting recognition for a number of reasons [10]. The importance of HMMs in the area of speech recognition has been observed several ago [11]. In the meantime, HMMs have also been successfully applied to image pattern recognition problems such as shape classification [12] and face recognition [13]. HMMs qualify as suitable tool for cursive script recognition for a number a reasons. First, they are stochastic models that can cope with noise and pattern variations occurring is human handwriting. Next, the number of tokens representing an unknown input word may be of variable length.

We have used discrete Hidden Markov Models proposed by Hewavitharana et al. [16] in our approach. A discrete HMM is characterized by the following [11]:

1. N, the number of states in the model. The set of states in the model is $S = \{1, 2, ..., N\}$, the state at time t is denotes as q_t

2. M, the discrete alphabet size. We denote the individual symbols as $V = \{v_1, v_2, ..., v_M\}$.

3. $A = \{a_{ij}\}$, the states transition probability distribution where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i], 1 \le t, j \le N$.

4. $B = \{b_j(k)\}$, the observation symbol probability distribution in state j where $b_j(k) = P[v_k \text{ at t} | q_t = S_j], 1 \le j \le N \ 1 \le k \le M$

5. $\pi = \{\pi_i\}$, the initial state distribution, where $\pi_i = P[q_1 = S_i], 1 \le i \le N$

A compact notation for the above HMM would be, $\lambda = \{A, B, \pi\}$

Maximum likelihood parameter estimation for HMM is obtained by the iterative procedure, the Baum-Welch algorithm [8], with multiple observation sequences. Then, for a given observation sequence $O = O_1, O_2, ..., O_T\}$, the HMM is used to compute $P[O | \lambda]$, where T is the length of the sequence.

Two HMMs are created for every character, one for modeling the horizontal information and the other for modeling the vertical information. The discrete hidden Markov character models are trained using standard procedures [8][9]. The numbers of states for all the character HMMs is fixed and no skip states are allowed. Only the pre-classified Candidate characters are passed on for HMM recognition.

Two log probabilities for each candidate are calculated using the horizontal direction HMM. Then, the log probabilities are added together to obtain a final 3-best character recognition.

7 Experimental Results

We have tested the proposed system with many handwritten documents of different individuals, Olaichuvadi samples, Machine printed, Scanned documents. Picture 11, 12 are few examples of the scanned documents we have tested.

Figure 11: Scanned Document Image

Figure 12: Recognition Output

The experimental results are listed in Table 1.

Count of Words	Efficiency (%)
25	96.4
50	95.6
100	94.8
150	93.7
200	93.0

Table 1. Recognition Results of Proposed Approach

8. Conclusion

Focusing primarily on ability to deal with variability feature of vector we presented method for character recognition offline cursive hand-written scripts in Tamil. The methodology is a streamlined approach incorporating discrete hidden Markov model based classifier for recognition. Classification of off-line vectorized handwritten characters has been proven to be efficient on application of Hidden Markov models. Significant increase in accuracy levels has been found on comparison of our method with the others for character recognition. Furthermore, this recognition model poses to be more compatible for other Indian scripts too. With the addition of sufficient pre processing the approach offers a simple and fast structure for fostering a full OCR system.

References:
[1] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient based learning applied to document recognition", Proceedings of the IEEE, vol. 86, no.11, IEEE, pp. 2278- 2324, USA , 1998.
[2] R. Plamondon and S. Srihari. On-line and Off-line Handwriting Recognition: A Comprehensive Survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(1):63–84, 2000.
[3] T. Varga. Off-line Cursive Handwriting Recognition Using Synthetic Training Data. PhD thesis, University of Bern, Switzerland, 2006.
[4] Hidden Markov Models for Online Handwritten Tamil Word Recognition, Bharath A, Sriganesh Madhvanath, 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, Sept 23-26, 2007.
[5] SEETHALAKSHMI R., SREERANJANI T.R., BALACHANDAR T., Abnikant Singh, Markandey Singh, Ritwaj Ratan, Sarvesh Kumar, "Optical Character Recognition for printed Tamil text using Unicode", Journal of Zhejiang University SCIENCE, Vol. 6A No. 11, 2005.
[6] S-B. Cho, "Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals", IEEE Trans. on Neural Networks, Vol. 8, 1997, pp. 43-53.
[7] K.H.Aparna, Sumanth Jaganathan, P.Krishnan, V.S.Chakravarthy, "An optical Character Recognition System for Tamil Newsprint," International Conference on Universal Knowledge and Language - 2002, 25th-29th November 2002, Goa, India, 2002. [8] K. Khatatneh, "Probabilistic Artificial Neural Network for Recognizing the Arabic. Hand Written Characters", Journal of Computer Science 3 (12), 881-886, 2006.
[9] B. Gosselin, \A comparison of different feature extractors for handwritten character recognition", Proc. of PRORISC IEEE Benelux Workshop on Circuits, System and Signal Processing 94, pp. 27- 32, Pappendaele, Pays-Bas, November 1994.
[10] H. Bunke, M. Roth, and E. G. Schukat-Talamazzini. Offline Cursive Handwriting Recognition using Hidden Markov Models. Pattern Recognition, 28(9):1399–1413, 1995.
[11] L.R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech

Recognition, Proceedings of the IEEE, vol. 77 no. 2, 1989, pp. 257-286.

[12] Y.He, A.Kundu: 2-D shape classification Using Hidden Markov Model, IEEE Trans. On PAMI, vol.13, 1991, pp.1172-1184.

[13] F.Samaria, F.Fallside: Face Identification and Feature Extraction Using Hidden Markov Models, in G.Vernazza, A.N.Venetsanopoulos, C.Braccini (editors): Image Processing: Theory and Applications, Elsevier Science publishers B.V., 1993, pp.292-302.

[14] F. Jelinek, Statistical Methods for Speech Recognition, MIT-Press, 1998.

[15] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Processing, Prentice-Hall, 1993.

[16] S. Hewavitharana, H. C. Fernando and N.D. Kodikara, "Off-line Sinhala Handwriting Recognition using Hidden Markov Models", Proc. of Indian Conference on Computer Vision , Graphics & Image Processing (ICVGIP) 2002.

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

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

[19] R.M. Suresh, S. Arumugam and K.P. Aravanan, "Recognition of handwritten Tamil characters using fuzzy classificatory approach", Proc. The Tamil Internet 2000 Conference, Singapore, July 2000.

[20] Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International.

[21] BBC. India sets up classical languages. August 17,2004. http://news.bbc.co.uk/2/hi/south asia/ 3667032.stm

[22] The Hindu. Sanskrit to be declared classical language. October 28, 2005. Retrieved on 2007-08-16.http://www.hindu.com/2005/10/28/stories/2005102809281200.htm

[23] D Dhanya and A G Ramakrishnan, "Simultaneous Recognition of Tamil and Roman Scripts", Proc. Tamil Internet 2001, Kuala Lumpur, Aug 26-28, 2001, pp. 64-68.

[24] K.H. Aparna & V.S. Chakravarthy, "A complete OCR system development of Tamil Magazine Documents," Tamil Internet 2003, Chennai, August, 22-24, 2003.

[25] "Multiclass Hierarchical SVM for Recognition of Printed Tamil Characters"; R Loganathan, Shivsubramani K Moorthy, V Ajay, K P Soman Selected in IJCAI- AND2007 International Workhop.

[26] B.M. Sagar, Dr. Shobha G, Dr. Ramakanth Kumar P "OCR for printed Kannada text to Machine editable format using Database approach" WSEAS TRANSACTIONS on COMPUTERS May 30, 2008.

[27] Hani Khasawneh "A New Algorithm for Arabic Optical Character Recognition" Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (Pp211-224).

[28] Nadia Ben Amor, Najoua Essoukri Ben Amara " A Hybrid Approach For Multifont Arabic Characters Recognition" Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp194-198)