European Journal of Scientific Research
ISSN 1450-216X Vol.35 No.4 (2009), pp.570-582

http://www.eurojournals.com/ejsr.htm

A Comparative Study of Optical Character Recognition for Tamil Script

R. Jagadeesh Kannan
RMK Engg College, Chennai, India

R. Prabhakar
CIT, Coimbatore, India

Abstract

Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. Now, the rapidly growing computational power enables the implementation of the present CR methodologies and also creates an increasing demand on many emerging application domains, which require more advanced methodologies. Researchers for the recognition of Indic Languages and scripts are comparatively less with other languages. This paper gives an overview of the ongoing research in optical character recognition (OCR) systems for Tamil scripts and comparative study of our previous work. This survey paper has been felt necessary when the work on developing OCRs for Indian scripts, mainly focused on Tamil script is very promising, and is still in emerging status. The aim of this paper is to provide a starting point for the researchers entering into this field. Peculiarities in Tamil scripts, present status of the OCRs for Tamil scripts, techniques used in them, recognition accuracies, and the resources available, are discussed in detail.

Keywords: Optical character recognition, Printed, Handwritten, Cursive, Offline, Online, Tamil Scripts.

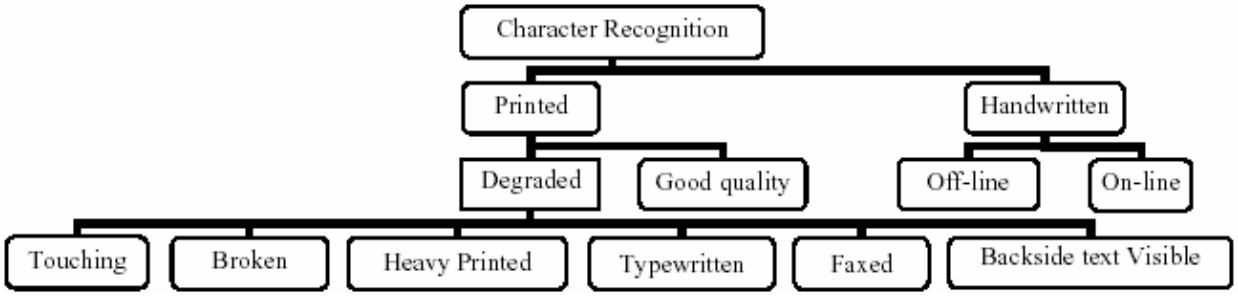
1. Introduction

During the past thirty years, substantial research efforts have been devoted to character recognition that is used to translate human readable characters to machine-readable codes. Immense effort has been made on character recognition, as it provides a solution for processing large volumes of data automatically in a large variety of scientific and business applications. Optical character recognition (OCR) deals with the recognition of optically processed characters rather than magnetically processed ones. OCR is a process of automatic recognition of characters by computers in optically scanned and digitized pages of text [3]. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical applications. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications [1] [2].

The field of Document Analysis and Recognition is vast and it contains many applications. Character recognition is one of the branches of DAR. As shown in Fig.1 the problem of character recognition can be divided into printed and handwritten character recognition. Handwritten character recognition has been further divided into off-line and online handwritten character recognition [4]. Off-

line handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in grey scale format. After being stored, it is conventional to perform further processing to allow superior recognition. In the on-line case, the handwriting is captured and stored in digital form via different means. The data generated by users writing with a stylus on an electronic device is known as digital ink, and the process of writing is called on-line handwriting [5].

Figure 1: Hierarchy of character recognition problems



It is generally accepted that the on-line method of recognizing handwritten text has achieved better results than its off-line counterpart. This may be attributed to the fact that more information may be captured in the on-line case such as the direction, speed and the order of strokes of the handwriting. On the other side machine-printed character recognition can be on good quality documents or degraded printed documents.

Character and handwriting recognition has a great potential in data and word processing for instance, automated postal address and ZIP code reading, data acquisition in bank checks, processing of archived institutional records, etc. Combined with a speech synthesizer, it can be used as an aid for people who are visually handicapped. As a result of intensive research and development efforts, systems are available for English language [6], [7] Chinese/Japanese languages [8], [9], [10] and handwritten numerals [11], [12]. However, less attention had been given to Indian language recognition. Main reasons for this slow development could be attributed to the complexity of the shape of Indian scripts, and also the large set of different patterns that exist in these languages, as opposed to English. Some of the peculiarities in Indian scripts are explored.

Indian scripts are different from Roman script in several ways. Indian scripts are two-dimensional compositions of symbols: core characters in the middle strip, optional modifiers above and/or below core characters. Two characters may be in shadow of each other. While line segments (strokes) are the predominant features for English, most of the Indian language scripts are formed by curves, holes, and also strokes. In Indian language scripts, the concept of upper case and lower-case characters is absent; however, the alphabet itself contains more number of symbols than that of English. Some efforts have been reported in the literature for Telugu, Devanagari [13], [14], [15], [16] and Bangla [17] scripts.

Tamil, the native language of a southern state in India, is one of the oldest language has several million speakers across the world and is an official language in countries such as Sri Lanka, Malaysia, Singapore and Tamil Nadu State of India. As it is the case with all Indic scripts, Tamil has a large alphabet size and hence text entry through QWERTY keyboard is cumbersome. The penetration of Information Technology (IT) becomes harder in a country such as India where the majorities 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.

An OCR implementation consists of a number of preprocessing steps followed by the actual recognition. The number and types of preprocessing algorithms employed on the scanned image depend on many factors such as age of the document, paper quality, resolution of the scanned image, the amount of skew in the image, the format and layout of the images and text, the kind of script used

and also on the type of characters - printed or handwritten (Anbumani & Subramanian 2000). Typical preprocessing includes the following stages, such as Binarization, Noise removing, Thinning, Skew detection and correction, Line segmentation, Word segmentation, and Character segmentation Recognition consists of Feature extraction, Feature selection, and Classification.

We briefly review the prior works on Recognition of Tamil characters in the following sections of this paper, which will examine each of the approaches, and give a summary of the progress made in each area. Along with that we briefly elucidate our two previous works on Tamil character recognition in the coming section.

2. Literature Survey

Siromoney et al. [18] described a method for recognition of machine printed Tamil characters using an encoded character string dictionary. The scheme employs string features extracted by row- and column-wise scanning of character matrix. The features in each row (column) are encoded suitably depending upon the complexity of the script to be recognized. A given text is presented symbol by symbol and information from each symbol is extracted in the form of a string and compared with the strings in the dictionary. When there is agreement the letters are recognized and printed out in Roman letters following a special method of transliteration. The lengthening of vowels and hardening of consonants are indicated by numerals printed above each letter.

Chinnuswamy et al. [19] proposed an approach for hand-printed Tamil character recognition. Here, the characters are assumed to be composed of line-like elements, called primitives, satisfying certain relational constraints. Labeled graphs are used to describe the structural composition of characters in terms of the primitives and the relational constraints satisfied by them. The recognition procedure consists of converting the input image into a labeled graph representing the input character and computing correlation coefficients with the labeled graphs stored for a set of basic symbols. This algorithm uses topological matching procedure to compute the correlation coefficients and then maximizes the correlation coefficient.

Suresh et al. [20] describes an approach to use the fuzzy concept on handwritten Tamil characters to classify them as one among the prototype characters using a feature called distance from the frame and a suitable membership function. The unknown and prototype characters are preprocessed and considered for recognition. The theory of fuzzy set provides an approximate but effective means of describing the behavior of ill-defined systems. Patterns of human origin like handwritten characters are to some extent found to be fuzzy in nature. It is decided to use fuzzy conceptual approach effectively. The algorithm is tested for about 250 samples for numerals and seven chosen Tamil characters and the success rate obtained varies from 76% to 94%.

Hewavitharana, S, and H.C. Fernando [21] a system is described to recognize handwritten Tamil characters using a two-stage classification approach, for a subset of the Tamil alphabet, which is a hybrid of structural and statistical techniques. In the first stage, an unknown character is pre-classified into one of the three groups: core, ascending and descending characters. Structural properties of the text line are used for this classification. Then, in the second stage, members of the pre-classified group are further analyzed using a statistical classifier for final recognition. The main recognition errors were due to abnormal writing and ambiguity among similar shaped characters. This could be avoided by using a word dictionary to look-up for possible character compositions. The presence of contextual knowledge will help to eliminate the ambiguity. We strongly feel that the method of pre-classification would have much higher recognition accuracy if applied to Optical Character Recognition, since printed characters preserve the correct positioning on three-zone frame.

In [22] described a recognition system for offline unconstrained handwritten Tamil characters based on support vector machine (SVM). SVM is a new type of pattern classifier based on a novel statistical learning technique. Due to the difficulty in great variation among handwritten characters, the system is trained with 106 characters and tested for 34 selected Tamil characters. The characters are chosen such that the sample data set represents almost all the characters. Data samples are collected

from different writers on A4 sized documents. They are scanned using a flat bed scanner at a resolution of 300 dpi and stored as grey scale images. Various preprocessing operations are performed on the digitized image to enhance the quality of the image. Random sized preprocessed image is normalized to uniform sized image. Pixel densities are calculated for different zones of the image and these values are used as the features of a character. These features are used to train and test the support vector machine.

Shivsubramani et al. [23] presents an efficient method for recognizing printed Tamil characters exploring the interclass relationship between them, which should be accomplished using Multiclass Hierarchical Support Vector Machines. A new variant of Multi Class Support Vector Machine constructs a hyper plane that separates each class of data from other classes. Character recognition, thus, involves classification of characters into multi-classes. Of the 126 unique characters identified in Tamil language, inter-class dependencies were found within many characters due to the similarity in their shapes. This enabled them to be organized into hierarchies, thus enhancing and simplifying the process classification. Taking advantage of the inter-class dependencies within the character a hierarchical based classification is possible based on the views put forth by [Szedmak et al., 2005] [24]. Combining both the views together, a Multiclass Hierarchical SVM algorithm was devised and is understood to be very efficient methods for character classification. The algorithm did prove more efficient than some of the commonly used classifiers. Some merits of our algorithm are:

- Strong mathematical model foundation rather than heuristics and analogies.
- Efficient in terms of accuracy (96.85%) in comparison with many commonly used classifiers.

In [25] the author presents a complete document image analysis system for Tamil newsprint. The system 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. Experience with OCR problems teaches that for most subtasks (text block identification, and character recognition.) involved in OCR, there is no single technique that gives perfect results for every type of document image. We have used the strength of artificial neural networks in empirical model building for solving the key problems of segmentation and character recognition. The final document is reconstructed in HTML document format and it is currently done manually. An important lacuna in our present system is absence of a suitable document model. A document is treated as a collection of disparate items without any logical structure connecting all of them. Future efforts will be focused on embedding various document components into a logical structure. This will help to make the reconstruction process as automatized. 94% recognition rate is obtained, when the text block having a few touching characters present.

In [26] the author gives a comparison of elastic matching schemes for writer dependent on-line handwriting recognition of isolated Tamil characters. Three different features are considered namely, preprocessed x-y co-ordinates, quantized slope values, and dominant point co-ordinates. Seven schemes based on these three features are compared using an elastic distance measure. The comparison is carried out in terms of recognition accuracy, recognition speed, number of training templates, and dynamic time warping based distance measure has been presented. The results show that dominant points based two-stage scheme, and combination of rigid and elastic matching schemes perform better than rest of the schemes, especially from the point of view of implementing them in a real time application. Efforts are underway to devise character grouping schemes for hierarchical classification, and classifier combination schemes so as to obtain a computationally more efficient recognition scheme with improved accuracy.

In [27] the author discusses the various strategies and techniques involved in the recognition of Tamil text and they refer Optical Character Recognition (OCR) for the process of converting printed Tamil text documents into software translated Unicode Tamil Text. The printed documents available in the form of books, papers, magazines, etc. are scanned using standard scanners which produce an image of the scanned document. As part of the preprocessing phase the image file is checked for skewing. The skewed image is corrected by a simple rotation technique in the appropriate direction,

and then it is passed through a noise elimination phase and is binarized. The preprocessed image is segmented using an algorithm, which decomposes the scanned text into paragraphs using special space detection technique and then the paragraphs into lines using vertical histograms, and lines into words using horizontal histograms, and words into character image glyphs using horizontal histograms. Each image glyph is comprised of 32×32 pixels. Thus a database of character image glyphs is created out of the segmentation phase. Then all the image glyphs are considered for recognition using Unicode mapping. Each image glyph is passed through various routines, which extract the features of the glyph. The various features that are considered for classification are the character height, character width, the number of horizontal lines (long and short), the number of vertical lines (long and short), the horizontally oriented curves, the vertically oriented curves, and the number of circles, number of slope lines, image centroid and special dots. The extracted features are passed to a Support Vector Machine (SVM) where the characters are classified by Supervised Learning Algorithm. These classes are mapped onto Unicode for recognition.

In [28] a generalized framework for Indic script character recognition is proposed and Tamil character recognition is discussed as a special case. Unique strokes in the script are manually identified and each stroke is represented as a string of shape features. The test stroke is compared with the database of such strings using the proposed flexible string-matching algorithm. The sequence of stroke labels is then converted into "horizontal block" using a rule list and the sequence of horizontal blocks is recognized as a character (with its IISCI code) using a Finite State Automaton (FSA). Online HWR studies typically handle smaller number of stroke classes since many of them deal with Latin script (26 or 52 classes). However, a study by Yeager et al presents results with 95 classes achieving 86.1% performance. Our results with 96 Tamil stroke classes compare favorably with the results of Yeager et al.

In [29, 30] a subspace-based method using Principal Component Analysis (PCA) is applied for Tamil character recognition. The input is a temporally ordered sequence of (x, y) pen coordinates corresponding to an isolated character obtained from a digitizer. The input is converted into a feature vector of constant dimensions following smoothing and normalization. Each class is modeled as a subspace, and for classification, the orthogonal distance of the test sample to the subspace of each class is computed. The effort published in [30] compares the performance of DTW and PCA for three modes of recognition: writer independent, writer dependent and writer adaptive. DTW is shown to outperform PCA in all the three modes of recognition. Although the performance of DTW based method is marginally better, in terms of speed, subspace based method wins over. Also classifier combination schemes for the two methods are proposed. DTW based method is computationally expensive; the disadvantage may be overcome by using prototype selection/reduction methods.

In [31], the problem of high interclass similarity in the case of Tamil characters is addressed by finding appropriate features. Angle features, Fourier coefficients and Wavelet features are compared using a Neural Network classifier. In the absence of smoothing, angle features are susceptible to noise and may fail to capture the intra-class similarity. Fourier coefficients do not capture subtle differences between two similar-looking characters because a change in the values of x and y over a small interval of time gets nullified over the entire frequency domain. On the other hand, Wavelet features are shown to retain the intra-class similarity and inter-class differences, resulting in high recognition accuracy. A Single hidden layer network was used for classification. The network gave excellent performance. The classification accuracy is 96.54% for 12-character problem and 94.30% for 135-character problem.

In [32] the author proposed a Data-driven HMM-based online handwritten word recognition system for Tamil. A symbol set consisting of 84 symbols was defined for the word recognition task and each symbol was modeled using a left-to-right HMM. Intersymbol pen-up strokes were modeled explicitly using two state left-to-right HMMs to capture the relative positions between symbols in the word context. Independently built symbol models and inter-symbol pen-up stroke models were concatenated to form the word models. The relatively low performance in the case of high lexicon size can be improved by the use of statistical language models, which are commonly applied in Western cursive recognition.

In [33] the author describes an approach to recognize handwritten Tamil characters using a multilayer perception with one hidden layer. The feature extracted from the handwritten character is Fourier descriptors. Also an analysis was carried out to determine the number of hidden layer nodes to achieve high performance of back propagation network in the recognition of handwritten Tamil characters. The system was trained using several different forms of handwriting provided by both male and female participants of different age groups. Test results indicate that Fourier descriptors combined with back propagation network provide good recognition accuracy of 97% for handwritten Tamil characters.

There are few works going on in Tamil OCR, the accuracy of the approaches still remains a challenging area of research. Many of the works related to Tamil OCR have not concentrated or dealt enough with the accuracy parameter. In the previous works, we had proposed two methods that are contributed to increase the performance of Tamil OCR. The methods are Accuracy Augmentation of Tamil OCR Using Algorithm Fusion [34] and Off-Line Cursive Handwritten Tamil Character Recognition [40].

2.1. Accuracy Augmentation of Tamil OCR Using Algorithm Fusion

The accuracy or efficiency of OCR purely depends on the algorithm we deployed. The efficiency decreases when an algorithm fails to identify a character or if the algorithm detects an unrelated character. We have proposed a method where we have fused two pattern recognition algorithms to get the advantage of both the algorithms and evaluate the efficiency of OCR. Before fusing, the scanned document is preprocessed. The processes involved in preprocessing are Histogram equalization and Gabor Filtering, Binarisation, ROI extraction and Region Probe Algorithm. Then we have fused two algorithms meaning that both the algorithms are taken into consideration. The advantages of this fusion process is given as

1) If one algorithm fails to identify a character, another algorithm may support in identifying the character.
2) If one algorithm gives wrong character another may give a correct one.
3) The possibility for same wrong identification by both the algorithms is less.
4) If one algorithm gives wrong result the decision of choosing the correct result is done by neural network which is discussed later in the paper

We have chosen Support Vector machine (SVM) and Hidden Markov Model (HMM) for the fusion and finally used neural network to predict the correct character when there arises a situation where two algorithms yield two different characters.

SVMs have achieved excellent recognition results in various pattern recognition applications [35]. Also in offline optical character recognition (OCR) they have been shown to be comparable or even superior to the standard techniques like Bayesian classifiers or multilayer perceptrons [36]. SVMs are discriminative classifiers based on Vapnik's structural risk minimization principle. They can implement flexible decision boundaries in high dimensional feature spaces. The implicit regularization of the classifier's complexity avoids overfitting and mostly this leads to good generalizations. Some further properties are commonly seen as reasons for the success of SVMs in real-world problems: the optimality of the training result is guaranteed, fast training algorithms exist and little a-priori knowledge is required, i.e. only a labeled training set.

Hidden Markov Models are suitable for handwriting recognition for a number of reasons [37].In the meantime, HMMs have also been successfully applied to image pattern recognition problems such as shape classification [38] and face recognition [39]. 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. Moreover, using an HMM-based approach, the segmentation problem, which is extremely difficult and error prone, can be avoided.

To improve the accuracy, we have trained Radial Bass Function Neural Network (RBFNN) with the output of both the algorithms. In that different samples of Tamil Characters are taken and given as input to both HMM and SVM. If HMM or SVM gives a false character, the neural network is trained with the weightage of both the algorithms and the actual character. This process is one for all the possible false recognition of the two algorithms. During OCR When both the algorithms not giving same character, trained RBFNN is used to retrieve the actual character. This way we can increase the accuracy of OCR.

We chose "Thirukural" OCR to test the proposed methodologies efficiency. Every "kural" has got 2 lines or 7 words. We tested the efficiency with set of "kural"s. The experimental results show that the accuracy is really improved than the previous works.

2.2. Off-line Cursive Handwritten Tamil Character Recognition

It is worth noticing that OCR deals with off-line recognition while handwriting recognition may be required for both on-line and off-line signals. Cursive Handwriting recognition is one of the very challenging problems.

The most crucial stage in the process of Optical Character Recognition (OCR) [41] is that of recognizing the characters and classifying them. The other processes involved include preprocessing activities like binarization and skew estimations. Preprocessing is primarily used to reduce variations of handwritten characters. It is followed by major phases like Segmentation and Feature Extraction. The process of segmentation and recognition pose quiets a lot of challenges especially in recognizing cursive hand-written scripts of different languages. The process of handwriting recognition involves extraction of some defined characteristics called features to classify an unknown character into one of the known classes. 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.

We have chosen a complete off-line OCR system for cursive handwritten Tamil characters using Hidden Markov Model. HMMs qualify as suitable tool for cursive script recognition for a number a reasons. In that 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. We have used discrete Hidden Markov Models proposed by Hewavitharana et al. [42] in our approach. 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 [43]. 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.

We have tested the proposed system with many handwritten documents of different individuals, Olaichuvadi samples, Machine printed, Scanned documents. 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.

3. Performance Evaluation

Suresh et al. [20], in these unconstrained handwritten numerals with invariant position and size are considered. The strings are obtained by considering the trace in clockwise direction. A 20 x 20 frame is used for writing alphabets. Prototype Numerals or Tamil Characters are used to infer the grammar. The algorithm was applied for about 250 samples for each of the Numerals seven chosen Tamil characters

like KA, THU, CHA, RA, TA, MA, YA and the percentage of successful recognition varies from 76% to 94%.

Hewavitharana et al. [21], the system was trained with 1000 characters belonging to all the classes. The testing data contained a separate set of 800 characters. A total of 50 text lines were subjected to segmentation and reference line identification. In all the cases, every character in each text line was correctly segmented. The reference line identification was almost 99% accurate resulting only 1% pre-classification error. In the trained set, a recognition rate of 89.5% was achieved for the 1st choice and 98.5% for the top 3 choices. In the test set, a recognition rate of 79.9% was achieved for the 1st choice and 96.9% for the top 3 choices. Understandably, the training set produced much higher recognition rate than the test set.

Shanthi et al. [22], the system was trained with 35441 characters belonging to 106 different characters written by 117 different users. The testing data contained a separate set of 6048 characters belonging to 34 different characters. A portion of the training data is also used to test the system, to check how well the system responds to the data it has been trained on. In that the pixel densities are calculated for different sized normalized image for the unknown characters and the features are given to the SVM classification module. The characters are classified based on the highest match and the recognized characters are stored in a word file and the characters can be viewed using the available TAM (TAmil Monolingual) font. The recognition rate for 32X32-image size is between 62.84 to 98.9%. Recognition rate for 48X48-image size is between 65.71 to 99.5%. For 64X64-image size the recognition rate is between 67 to 99.5% and for overlapping zone the recognition rate is between 71.7 to 98.9%. The overall recognition rate also increased from 85.5 to 87.35%.

Shivsubramani et al. [23], in this system 126 unique commonly occurring Tamil characters in shape have been identified. Hierarchy is built based on the 126 characters. Training data set was generated by labeling the features extracted from the test character image, with the corresponding class. A training dataset for a particular class, on average, contains 20 sample training data. Multiclass Hierarchical SVM turned out to be a very efficient method in process of classification. The accuracy of the algorithm depended on two parameter settings (RBF Kernel parameter σ and regularization parameter C).Based on a comparative study performed, Multiclass Hierarchical SVM showed better accuracy rate than many other classifiers used like Multilayer perceptron, KNN, Naive Bayes, decision tree and other rule based classifiers. The system was tested thrice using 3, 7 and 20 most similar characters respectively. The accuracy rate for Multiclass Hierarchical SVM is between 96.23 to 96.86%. For Multilayer perceptron the accuracy rate is between 91.8 to 95.45%. For KNN the accuracy rate is between 89.40 to 90.05%. For naïve Bayes the accuracy rate is between 84.5 to 88.90%. For decision tree the accuracy rate is between 91.0 to 93.23%.

Aparna et al. [25], in this system for Tamil character recognition, the text blocks have to be initially segmented into lines, words and characters. For the text block segmentation and character recognition the inverted binarized document (i.e. 0 for back ground and 1 for foreground) is being taken. For character recognition the original document (without any scaling) is being taken. A Radial basis function neural network is trained for the recognition of characters. All the possible 157 characters of Tamil script including English numerals are taken for training the neural network. This trained neural network is used for the recognition of the segmented characters. When the text block having a few touching characters is sent for character recognition, 94% recognition rate is obtained. In general, for other documents it varied from 85 to 90 percent depending on the touching characters present in the text part.

Joshi et al. [26], this is for on-line Tamil handwriting recognition is based on template based elastic matching algorithms. Advantage of elastic matching algorithms is that they do not require a relatively large amount of training data, making them suitable for writer dependent recognition. Character recognition based on DTW with seven different recognition schemes. It is an elastic matching technique and hence it allows comparing two sequences of different lengths. The performance of these schemes with respect to three criteria: average recognition accuracy, average recognition speed, and number of training templates used per class. Average recognition accuracy is

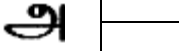
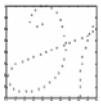
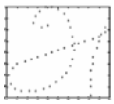
found out by dividing the number of correctly recognized test patterns by the total number of test patterns. Average recognition speed is calculated by dividing the number of test patterns recognized by the total time taken and its unit is characters per second. Our results show that dominant points based two-stage scheme (scheme 6), and combination of rigid and elastic matching schemes (scheme 7) perform better than rest of the schemes, especially from the point of view of implementing them in a real time application. Scheme 6 gives 94.8% recognition accuracy with recognition speed of 14.45 chars/sec where as scheme 7 gives 95.89% recognition accuracy with recognition speed of 32.65 chars/sec.

Sundaresan et al. [32], in this system we considered only 12 Tamil characters for recognition. After finding a good character representation (wavelet features) and neural network architecture, we focused on the complete set of Tamil characters. We used the sequence of angle features for representing the Tamil characters. A Single hidden layer network was used for classification. The network gave excellent performance. The classification accuracy is 96.54% for 12 character problem and 94.30% for 135 character problem. In the case of 12 character problem, there are 50 examples per class for training purpose and 200 examples for testing purpose and for 135 characters problem there are 40 examples per class for training and 20 examples for testing.

Bharath et al. [33], the evaluation of the word recognition system was carried out on different lexicon sizes such as 1K, 2K, 5K, 10K and 20K words to assess the performance of the word models in terms of recognition accuracy. Even though a Tamil word is normally written as a sequence of syllabic units, the writing order of symbols within a syllabic unit may change with writers. However, from the experience of data collection and by manual inspection of a few collected samples, it was observed that the majority write the base consonant first and then the matra (if any), except for matras 78, 79 and 80. These matras are written before writing the base consonant for two reasons: (i) these matras are horizontally separate from the base consonant and occur on the left of the consonant and (ii) writing them after the base consonant will considerably interrupt the flow of writing. These facts were taken into account while determining the expected order of symbols for any given Unicode string. Manual inspection of the samples also reveal d that handwritten Tamil words rarely suffer from the problem of delayed strokes when compared to western cursive writing. This alleviates the need to capture delayed strokes in the word model. During evaluation, it is ensured that the truth of an input test sample is always present in the lexicon in the expected order of symbols. The accuracies obtained ranged from 98% to 92.2% with different lexicon sizes (1K to 20K words).

The comparison of recognition rates of our method with all the available methods is presented in Table 1.

Table 1: Comparison of Recognition rate

Existing Method		% Recognition Rate		Our Methods		% Recognition Rate
Suresh et al.[20]		88			Algorithm Fusion Off-Line Cursive Handwritten	88.12 87.13
Hewavitharana et al. [21]		Trained set	98.5		Algorithm Fusion	98.51 97.01
		Test set	96.9		Off-Line Cursive Handwritten	98.32 96.72
Shanthi et al. [22]		32x32 48x48 64x64	81.32 90.66 94.51		Algorithm Fusion	81.33 90.68 94.56 94.56
		64x64 overlapping	94.51		Off-Line Cursive Handwritten	81.03 90.55 94.43 94.42
Shivsubramani et al. [23]		Characters 3	96.85		Algorithm Fusion	96.88 96.22 96.89
		7 20	96.23 96.86		Off-Line Cursive Handwritten	96.81 96.12 96.80
Joshi et al. [26]		Speed (chars/sec) 1.69	99.69		Algorithm Fusion	99.70 98.89 99.18
		3.31	98.87			98.51 95.92 94.83 95.91
		5.83 67.39 32.65 14.45 32.65	99.15 98.49 95.89 94.80 95.89		Off-Line Cursive Handwritten	99.66 98.82 99.11 98.45 95.87 94.77 95.88
Sundaresan et al. [32]		Character problem 12	96.54		Algorithm Fusion	96.56 94.31
		135	94.30		Off-Line Cursive Handwritten	96.51 94.29
Bharath et al. [33]		Lexicon size 1K	97.96		Algorithm Fusion	97.97 95.83 94.51 93.18 92.16
		2K 5K 10K 20K	95.82 94.49 93.17 92.15		Off-Line Cursive Handwritten	97.94 95.80 94.47 93.14 92.12

4. Conclusion

We have given an overview of the ongoing research in optical OCR systems for Tamil scripts and comparative study of our previous work. This paper has been felt necessary when the work on developing OCRs for Indian scripts, mainly focused on Tamil script is very hopeful, and is still in promising status.

References

[1] Mantas, J., 1986. An overview of character recognition methodologies, Pattern recognition, 19 (6): 425-430.

[2] Govindan, V.K. and A.P. Shivaprasad, 1990. Character Recognition-A Review, Pattern Recognition, 23 (7): 671-683.

[3] Pal, U., and B.B. Chaudhuri, 2004. Indian Script Character Recognition: a Survey, Pattern Recognition, 37: 1887-1899.

[4] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: a comprehensive survey", IEEE Transactions on PAMI, Vol. 22(1), pp. 63–84, 2000.

[5] R. Plamondon, D. Lopresti, L.R.B. Shoemaker and R. Srihari, "On-line Handwriting Recognition," Encyclopedia of Electrical and Electronics Eng., J.G. Webster, ed., vol. 15, pp.123-146, New York: Wiley, 1999.

[6] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition", IEEE Trans. on Pattern Anal. Mach. Intell., vol. 11, no. 1, pp. 68-83, Jan. 1989.

[7] Hu, M. K. Brown and W. Turin, "HMM based on-line handwriting recognition", IEEE Trans. on Pattern Anal. Mach. Intell., vol. 18, no. 10, pp. 1039-1045, Oct. 1996.

[8] D. Deng, K. P. Chan, and Y. Yu, "Handwritten Chinese character recognition using spatial Gabor filters and self-organizing feature maps", Proc. IEEE Inter. Confer. On Image Processing, vol. 3, pp. 940-944, Austin TX, June 1994.

[9] C-H. Chang, "Simulated annealing clustering of Chinese words for contextual text recognition", Pattern Recognition Letters, vol. 17, no. 1, pp. 57-66, 1996.

[10] H. Yamada, K. Yamamoto, and T. Saito, "A non-linear normalization method for hand printed Kanji character recognition–line density equalization", Pattern Recognition, vol. 23, no. 9, pp. 1023-1029, 1990.

[11] S-W Lee, "Off-line recognition of totally unconstrained handwritten numerals using multiplayer cluster neural network", IEEE Trans. on Pattern Anal. Mach. Intell., vol. 18, no. 6, pp. 648-652, June 1996.

[12] J. Cai and Z-Q Liu, "Integration of structural and statistical information for unconstrained handwritten numeral recognition," IEEE Trans. on Pattern Anal. Mach. Intell., vol. 21, no. 3, pp. 263-270, Mar. 1999.

[13] V. Bansal and R.M.K. Sinha, "On how to describe shapes of Devanagari characters and use them for recognition", Proc. 5th Int. Conf. Document Analysis and Recognition, Bangalore, India, pp. 410-413, Sept. 1999.

[14] R Bajaj and S Chaudhary. 'Devanagari Numeral Recognition using Multiple Neural Classifiers.' Indian Conference on Pattern Recognition, Image Processing and Computer Vision (ICPIC), December 1995.

[15] K K Biswas and S Chatterjee. 'Feature based Recognition of Hindi Characters.'Indian Conference on Pattern Recognition, Image Processing and Computer Vision (ICPIC), December 1995, pp 182-187.

[16] S Palit and B Chaudhary. 'A Feature based Scheme for Machine Recognition of Printed Devanagari Script.' Indian Conference on Pattern Recognition, Image Processing and Computer Vision, (ICPIC), December 1995.

[17] B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system", Pattern Recognition, vol. 31, no. 5, pp. 531-549, 1997.

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

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

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

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

[22] [1]N. Shanthi and K. Duraiswamy, "Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM", 1Department of Information Technology, 2Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, India

[23] 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.

[24] [Szedmak et al., 2005] Sandor Szedmak, John Shawe- Taylor, Learning Hierarchies at Two-class Complexity, Kernel Methods and Structured domains, NIPS 2005

[25] K.H.Aparna, Sumanth Jaganathan, P.Krishnan, V.S.Chakravarthy, "Document Image Analysis: with specific Application to Tamil Newsprint", Department of Electrical engineering, IIT Madras, Chennai-600036.

[26] N. Joshi, G. Sita, A. G. Ramakrishnan, and S. Madhvanath. Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition. Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, 2004.

[27] 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.

[28] H. Aparna, V. Subramanian, Kasirajan, V. Prakash, V. Chakravarthy, and S. Madhvanath. Online Handwriting Recognition for Tamil. Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition, 2004.

[29] V. Deepu and S. Madhvanath. Principal Component Analysis for Online Handwritten Character Recognition. Proceedings of the 17th International Conference on Pattern Recognition, 2004.

[30] N. Joshi, G. Sita, A. G. Ramakrishnan, and S. Madhvanath. Tamil Handwriting Recognition Using Subspace and DTW Based Classifiers. Proceedings of the 11th International Conference on Neural Information Processing, 2004.

[31] C. S. Sundaresan and S. S. Keerthi. A Study of Representations for Pen based Handwriting Recognition of Tamil Characters. Proceedings of the 5th International Conference on Document Analysis and Recognition, 1999.

[32] "Hidden Markov Models for Online Handwritten Tamil Word Recognition." Bharath A, Sriganesh Madhvanath, HP Laboratories India, HPL-2007-108, July 6, 2007*

[33] Neural Network Based Offline Tamil Handwritten Character Recognition System, Sutha, J. Ramaraj. N., Sethu Inst. of Technol., Virudhunagar.

[34] R.Jagadeesh Kannan, R. Prabhakar, "Accuracy Augmentation of Tamil OCR Using Algorithm Fusion", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5, May 2008

[35] N. Cristianini and J. Shawe-Taylor. Support Vector Machines. Cambridge University Press, 2000.

[36] D. DeCoste and B. Schölkopf. Training invariant support vector machines. Machine Learning, 46(1/3):161, 2002.

[37] H. Bunke, M. Roth, and E. G. Schukat-Talamazzini. Offline Cursive Handwriting Recognition using Hidden Markov Models. Pattern Recognition, 28(9):1399–1413, 1995.

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

[39] 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.

[40] R.Jagadeesh kannan and R.Prabhakar, "Off-Line Cursive Handwritten Tamil Character Recognition" RMK Engineering College, Chennai, INDIA.

[41] G. Nagy, On the Frontiers of OCR, Proceedings of the IEEE, vol. 40, #8, pp. 1093-1100, July 1992.

[42] 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.

[43] K. Khatatneh, "Probabilistic Artificial Neural Network for Recognizing the Arabic. Hand Written Characters", Journal of Computer Science 3 (12), 881-886, 2006.