I J C T A, 9(3), 2016, pp. 1579-1587
© International Science Press

Tamil Handwritten Character Recognition: Progress and Challenges

K. Punitharaja*, and P. Elango**

***Abstract*:** An Optical Character Recognition (OCR) system may provide a solution to the data entry problems, a bottleneck for the data processing industry. Therefore, OCR systems are being developed for almost all major languages and Tamil language is no exception to it. During the past years, considerable research and development works have been done towards the development of an efficient Tamil character recognition system. In this paper present a comprehensive review of Tamil character recognition system techniques and evaluate the status of the Tamil character recognition system development.

***Keywords*:** Tamil Character Recognition, Tamil language, Optical Character Recognition, Digital document processing, Pattern recognition

1. INTRODUCTION

Optical character recognition (*OCR*) systems are being developed for almost all major languages and Tamil language is no exception to it. Like other languages, Tamil language poses its own challenging problems to the developers of Tamil character recognition (*TCR*) systems. During the past years, considerable research works have been done towards the development of an efficient T*CR* system. Many research articles and technical reports on this topic have appeared in leading conference proceedings and journals. The objective of this paper is to review the T*CR* literature and evaluate the status of the *TCR* system development.

We believe that the review should prove helpful in identifying and solving problems that are being faced in developing a practical T*CR* system. In this paper we have developed a set of criteria for categorizing the *TCR* techniques. The criterion set is based on the detailed lists of features definition and extraction, and classification methods that are frequently being experimented with by *TCR* researchers. Besides this, we have tabulated recognition result summary that can be readily used to compare and contrast the performance of respective recognition techniques. Result tabulation has been a problem as some research articles did not provide complete information about the test data that they have used and the environment under which the recognition experiments were conducted.

We have also analyzed characteristics of Tamil text with the recognition point of view. This analysis may give a better understanding of Tamil text and should prove beneficial in assessing the complexities that they may pose in developing a *TCR* system. General factors that affect *TCR* system development are discussed in Section 3. A set of criteria that we have used to categorize the existing *TCR* techniques is presented in Section 4. Recent advances in *TCR* system design are reviewed in Section 5. Recognition performance issues are discussed in Section 6. The current status of the test data that is being used and its relevance in developing a *TCR* system are examined in Section 7. A conclusion of our study is presented in section 8.

* Assistant Professor Department of Computer Science, Tranquebar Bishop Manikam Lutheran College, PORAYAR–609307, Nagapattinam District, Tamilnadu, India, *Email: punitharajak@gmail.com*

** Assistant Professor, Department of Information Technology Perunthalaivar Kamarajar Institute of Engineering and Technology (PKIET), Nedungadu, KARAIKAL-609603, Puducherry, India, *Email: elanalin74@gmail.com*

2. CHARACTERISTICS OF TAMIL LANGUAGE

India is a multi-lingual and multi-script country, where eighteen official scripts are accepted and have over hundred regional languages. Tamil is the most popular language in the world and particularly in Tamilnadu, India. More than 8 crore Tamils live in Tamil Nadu and Pondicherry. About one crore Tamils live in the other states of India. Outside India, Sri Lanka, Burma, Malaysia, Singapore, Indonesia, South Africa, Fiji, Mauritius islands are some of the countries having a large number of Tamil speaking people. Thus, the work on Tamil script is very useful for the Tamil community around the world. The alphabet of the modern Tamil script consists of 12 vowels, one Aaydham, 18 consonants and 216 consonantal vowels and hence there is a total of 247characters in Tamil. The basic characters of Tamil script are shown in Fig.1. Writing style in Tamil script is from left to right. The concept of upper/lower case is absent in Tamil script. In Tamil script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right (or both) or bottom of the consonant. These modified shapes are called modified characters. A vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. These compound letters are formed by adding a vowel marker to the consonant. Some vowels require the basic shape of the consonant to be altered in a way that is specific to that vowel. Others are written by adding a vowel-specific suffix to the consonant, yet others a prefix, and finally some vowels require adding both a prefix and a suffix to the consonant. In every case the vowel marker is different from the standalone character for the vowel. There are about 6 “Grantha” characters are also used in Tamil [8]. At present no dataset on Tamil handwritten compound characters is available and hence we consider only basic characters of Tamil script.

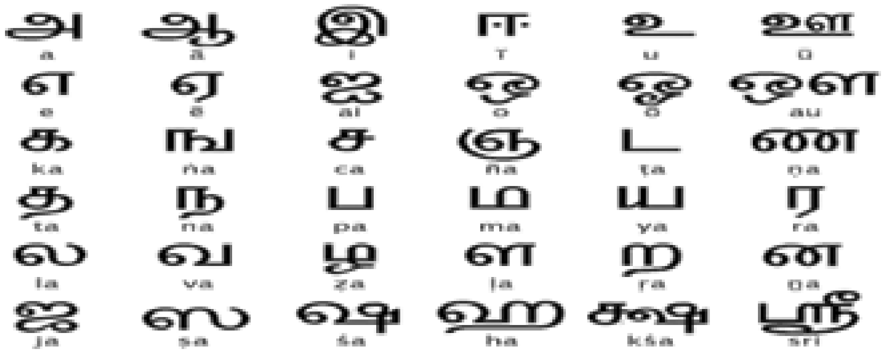
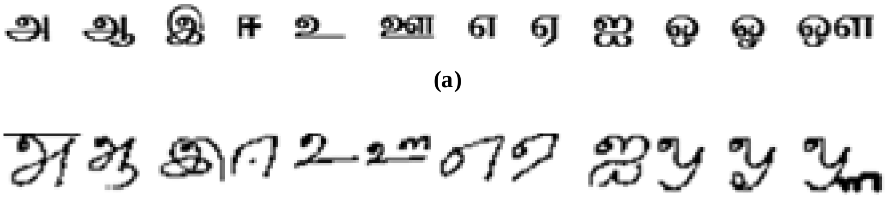


Figure 1: Samples of printed basic Tamil characters: Vowels and Consonants.

The complexity of a handwritten character recognition system increases mainly because of various writing styles of different individuals. Most of the errors in such system arise because of the confusion among the similar shaped characters. In Tamil there are many similar shaped characters. Examples of some groups of similar shaped characters are shown in Fig. 2. To get an idea of similar shaped printed as well as handwritten characters, we provide the samples of both printed and handwritten Tamil characters in Fig. 2. Although there are some differences between the samples of a group in the printed characters but the difference in the corresponding handwritten samples is very less. From the Fig. 2(b) it can be seen that



(a)

(b)

Figure 2: Examples of some similar shaped Tamil characters. (a) printed samples (b) handwritten samples of (a).

shapes of two or more characters of a group are very similar due to handwritten style of different individual and such shape similarity is the main reason of low recognition rate.

3. FACTORS AFFECTING THE DESIGN OF TAMIL CHARACTER RECOGNITION (*TCR*) SYSTEMS

3.1. Working Principle

Any character recognition system goes under following steps, i.e. Image acquisition, Preprocessing, Segmentation, Feature extraction, classification and post processing. Block diagram is of general character recognition system is shown in fig 3.

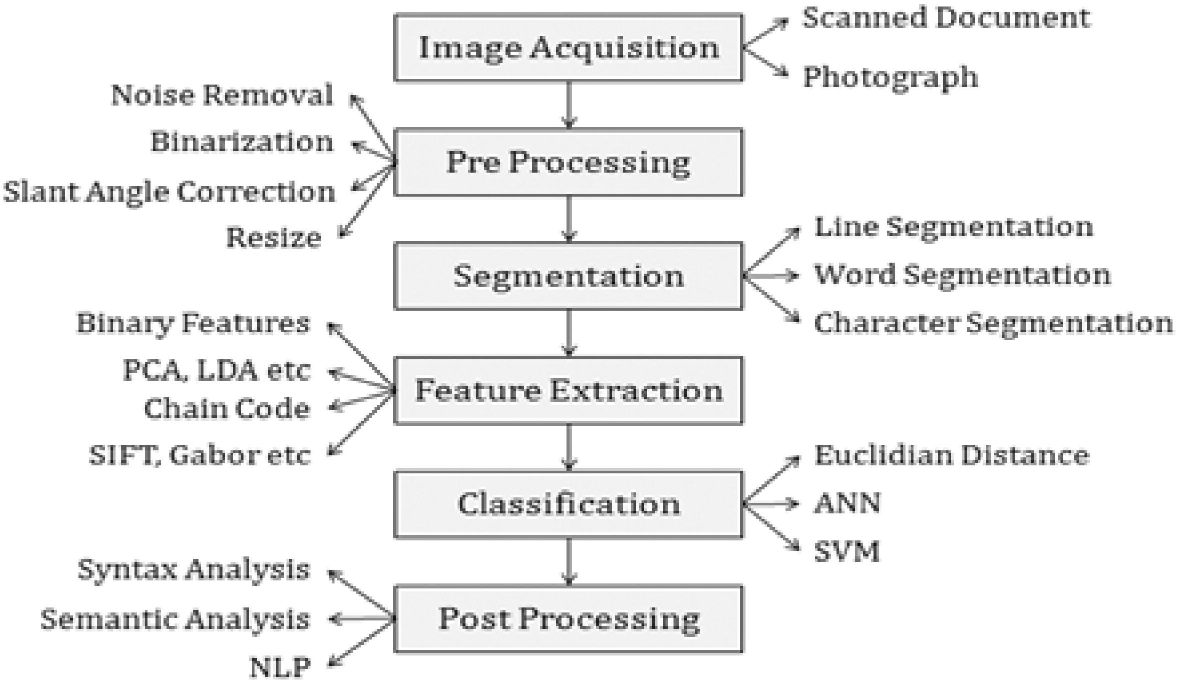


Figure 3: Block Diagram of Character Recognition System

Image acquisition: Images for HCR system might be acquired by scanning handwritten document or bycapturing photograph of document or by directly writing in computer using stylus. This is also known asdigitization process.

Preprocessing: Preprocessing involves series of operations performed to enhance to make it suitable forsegmentation [4]. Preprocessing step involves noise removal generated during document generation.Proper filter like mean filter, min-max filter, Gaussian filter etc may be applied to remove noise fromdocument. Binarization process converts gray scale or colored image to black and white image. Binarymorphological operations like opening, closing, thinning, hole filling etc may be applied to enhancevisibility and structural information of character. If document is scanned then it may not be perfectlyhorizontally aligned, so we need to align it by performing slant angle correction. Input document may beresized if it is too large in size to reduce dimensions to improve speed of processing. However reducingdimension below certain level may remove some useful features too.

Segmentation: Generally document is processed in hierarchical way. At first level lines are segmentedusing row histogram. From each row, words are extracted using column histogram and finally charactersare extracted from words. Accuracy of final result is highly depends on accuracy of segmentation.

Feature Extraction: Feature extraction is the heart of any pattern recognition application. Featureextraction techniques like Principle Component Analysis (PCA), Linear Discriminant Analysis

(LDA), Independent Component Analysis (ICA), Chain Code (CC), Scale Invariant Feature Extraction (SIFT), zoning, Gradient based features, Histogram might be applied to extract the features of individualcharacters. These features are used to train the system.

Classification: When input image is presented to HCR system, its features are extracted and given as aninput to the trained classifier like artificial neural network or support vector machine. Classifiers compare the input feature with stored pattern and find out the best matching class for input.

Post processing: This step is not compulsory; sometimes it helps to improve the accuracy ofrecognition. Syntax analysis, semantic analysis kind of higher level concepts might be applied to checkthe context of recognized character.

Support vector machine (SVM): Support vector machine is supervised learning tool, which is used for classification and regression. The basic SVM takes a set of input data and predicts, for eachgiven input. Given a set of training examples, each marked as belonging to one of two categories, aSVM training algorithm builds a model that assigns new examples into one category or the other. Moreformally, a support vector machine constructs a hyper plane or set of hyper planes in a highdimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point ofany class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.Whereas the original problem may be stated in a finite dimensional space, it often happens thatthe sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space, presumably making the separation easier in that space.

We classify factors affecting the development of a *TCR* system into two classes: *random* factors and *linguistic* factors. Random factors affect the document scanning process. Examples of random factors are: the document digitization errors, ink and dirt spattering, paper quality, the quality of writing tools, random distortions introduced by the scanning devices, etc. Linguistic factors are the intrinsic part of the Tamil language. The cardinality of the Tamil alphabet set is one of the linguistic factors that affect the design of *TCR* system. In Tamil character recognition literature, different cardinalities of Tamil alphabet set are considered by the researchers. Sometimes the cardinality is increased because different shapes of a character are considered as distinct characters. For human Tamil text readers this consideration may not make much difference, but for an optical character recognition (OCR) system different representations of such characters need to be considered as different characters because every shape once recognized is mapped into a distinct Unicode.

Character shape variation within a word is another linguistic factor that affects the design of a *TCR* system. Writing style and character connectivity are the two major sources of shape variations. Shapes of all characters that do not form connectivity with any character remain unchanged in every textual form. Shapes of connectivity forming characters vary drastically. Unlike the development of an OCR system for English alphabet based languages; this shape variation poses many problems in developing a *TCR* system for Tamil language.

Tamil text reproduction method-a linguistic factor-affects design of a *TCR* system. *Tamil* text is written from left to right and it is cursive in nature. Moreover, a *Tamil* word consists of one or more combined characters.

4. CLASSIFICATION OF TAMIL CHARACTER RECOGNITION TECHNIQUES

A combination of pattern recognition techniques is being devised to produce an efficient *TCR* system. We classify these techniques into different classes on the basis of the criteria: Textual data acquisition mode; Text style; Text segmentation; Feature definition, extraction and representation; Classification and Post processing.

4.1. Textual data acquisition mode

The on-line and off-line are the two possible modes of data acquisition. An online *TCR* system recognizes handwritten text by capturing the pen positions in real time. Research works for developing an efficient on-line character recognition system are described in section 5.1. An off-line *TCR* system recognizes an existing text. In such a system, digital images of existing texts are produced by scanning them line by line or page by page. Afterwards these images are processed and analyzed. An off-line *TCR* system may be built to recognize handwritten and typewritten Tamil texts. Recent advances in the development of an efficient off-line *TCR* system are presented in section 5.2.

4.2. Text style

A Tamil text can be produced in any of the following styles.

a) Unconstrained non-isolated handwritten text: The natural Tamil text-the cursive handwritten text produced without imposing any constraint on writers. The recognition scheme that recognizes this text requires both the word and character level segmentation (a process that breaks a sentence into isolated words and characters).

b) Unconstrained isolated handwritten text: The cursive unconstrained handwritten text produced where writers are expected to write isolated non-overlapping characters required for the mailing address.

c) Constrained non-isolated handwritten text: The text produced under some constraints like using writing guidelines or writing the text at a fixed position.

   For this type of text, word and character segmentation schemes are required to represent characters to the recognition module.

d) Constrained isolated handwritten text: Isolated characters produced under constrained environment. Filling a form where characters should be written in specified boxes is an example of such text.

e) Unifont typewritten text: Typewritten text that involves only one font.

f) Multifont typewritten text: Typewritten text involving many fonts.

Each text style mentioned above poses problems of its own and each requires a separate recognition scheme. The recognition schemes developed for solving the problems posed by various text styles are reviewed in section 5.2.1.

4.3. Text segmentation

A text segmentation process extracts the basic characters from a given text. In research works where the focus of the research was on the recognition technique only, the segmentation process was ignored. In such works we assume that the test data were manually segmented. However, the segmentation process is an essential step in automating a *TCR* system, therefore, several segmentation approaches were developed. Commonly used approaches for Tamil text segmentation are reviewed in section 5.2.2.

4.4. Feature definition, extraction and representation

A feature describes the characteristics of an underlying character, its partial structure or the structure of the whole word. Like OCR techniques, *TCR* techniques can also be distinguished from one another on the basis of feature definitions that they employ and the way they extract and represent features. In *TCR* research works, both the statistical (quantitative) and structural (qualitative) features are used.

4.5. CLASSIFICATION

In an attempt to obtain a good recognition score, almost all classification techniques: mathematical, statistical, syntactical, graph-theoretic, neural network based, heuristics and so on are used for Tamil character recognition. A comparative performance analysis of these techniques is presented in section 5.2.5. For some reasons, decision tree (a hierarchical graph-theoretic technique) based classification techniques are popular among *TCR* researchers; occasionally, some researchers have referred to this technique as syntactic technique which is a misleading term. Statistical techniques are the second frequent choice.

4.6. Post processing

The post processing is a process that uses properties of natural languages to enhance the recognition reliability. Recently, use of well established Markov model based post processing approaches for improving the reliability of an*TCR* system has appeared in *TCR* literature [83].

5. TAMIL CHARACTER RECOGNITION (*TCR*) SYSTEM DESIGN: RECENT ADVANCES

During the past years, attempts to design both the on-line as well as off-line *TCR* systems were intensively made, but as compared to research efforts for devising an off-line *TCR* system, the research efforts for devising an on-line *TCR* system are very little. Major on-line *TCR* techniques developed during the years and their performance are presented in Section 5.1. Off-line *TCR* techniques are reviewed and discussed in Section 5.2.

5.1. On-line *TCR* techniques

The first step in an on-line *TCR* technique is to extract features from the strokes that are formed using stylus as a writing tool in real-time. These features are extracted by segmenting the characters into strokes. Commonly used structural features are: open and closed curve segments, vertical and horizontal strokes. However, some researchers have also used hybrid features, i.e., a combination of structural and statistical features. They formed features by combining the structural feature: direction code with the statistical features: segment length, segment slope, coordinates of each segment point and dot counts.

5.2. Off-line TCR techniques

On the basis of information given in the research articles available to us, we categorize research works on off-line *TCR* systems into four categories. Category I consists of research articles that explicitly state that the technique was developed for offline *TCR* system. Categories II and III consist of research articles on typewritten and handwritten characters respectively. Although the term off-line was not explicitly mentioned, yet from the text we guess that techniques described in there were intended for the off-line recognition. Category IV consists of research articles where the textual style of the test data was not mentioned at all.

5.2.1. Text style

Research works to build *TCR* systems capable of recognizing text styles ranging from printed and standardized characters to totally unconstrained handwritten text are being carried out. Techniques are devised to recognize handwritten words handwritten characters as well as words. Similarly, recognition techniques are being devised for typewritten multifont character recognition and unifont character recognition. Some studies emphasize on the development of techniques, hence the text style is not specified.

5.2.2. Text segmentation

The text segmentation techniques are explicitly presented by some researchers. However, several researchers, who focus their attention on the development of feature extraction and classification modules only, have limited discussion on segmentation related issues. They assumed that the input to their systems were single isolated characters. There are cases where the entire segmentation process is completely ignored. Although text segmentation is a major issue, yet very little effort have been made to study the segmentation problem in isolation. Most of the segmentation processes are described as a preprocessing step. After analyzing the segmentation techniques, we group them into: stroke based segmentation histogram based segmentation outer contour analysis based segmentation connected component based segmentation projection based segmentation potential and actual column connection based segmentation.

5.2.3. Feature definition, extraction and representation

To obtain an accurate recognition performance, both the quantitative (statistical / numerical) and qualitative (structural / topological) features were defined and used in *TCR* research. Commonly used features are discussed below.

5.2.3.1. Quantitative features

The simplest feature is the black pixel count. In this category there are two features: one is the simple black pixel count in the entire region, another is the maximum and minimum pixel counts in a marked region like the quadrants. Other features are: character height, character width, character area, character weight above and below the base line, dot counts, aspect ratio (relationship between width and height of a character).

5.2.3.2. Qualitative features

The qualitative features represent the structure of the entire character or the stroke. Ideally, feature forming structures are assigned a code instead of a value. Examples of the qualitative features are: the branch point count codes, branch attributes, closed curves, open curves, corner point count codes, crossing code, crossing point; layout context (base line information and location of one character with respect to its neighbors) (character position relative to the base line); character’s position within a word (first, last, middle, or isolated character)

5.2.4. Classification

Classification schemes used in *TCR* researches are: Tree classifier, stage-wise classifier that uses three stages: primary, secondary and post-processing. Basic strokes are identified in the primary stage. Using the information of primary stage, characters are recognized in the secondary stage and finally a word is recognized in the post processing stage.

5.2.5. Post processing

After the character recognition process, there might be rejected character(s). In such cases, post processing can be used for further testing. They used a dictionary and a probability of observing a given lattice of characters using different models of a word.

6. PERFORMANCE ANALYSIS

In this section an analysis of off-line *TCR* systems is presented. Based upon the representation of results in the *TCR* literature, we categorize these test results into four categories described below.

6.1. Off-line character recognition

This category includes those research articles in which the word ‘off-line’ is explicitly mentioned. These articles describe systems that can recognize handwritten characters, and printed characters.

6.2. Typewritten character recognition

In this category, the research articles that explicitly describe systems for typewritten character recognition are included.

6.3. Handwritten character recognition

The research articles that explicitly describe techniques for handwritten character recognition are the part of this category.

7. DISCUSSION

The importance of the test data in the development of an OCR system for English based languages is very well documented and the same is true in the case of *TCR* system. Unfortunately, there is no standard test data set available that may be used to test and compare *TCR* techniques. In an attempt to locate a test data, we surveyed available research articles. Our findings are summarized in Tables 19-21 below. From Table 21, it can be seen that the largest test data set consists of 5000 characters only, which is insufficient for an authentic conclusion.

Table 1
Tamil Character Recognition

Training data set		*Test data set*			
Sample Size	*Correct Recognition %*	*Sample Size*	*CorrectRecognition %*	*Overall %*	*Test Data Attributes*
600	-	-	-	85	Handwritten
300	-	-	-	87	Handwritten
120	-	150	80	-	Handwritten
240	-	380	79.25	-	Handwritten
700	-	-	-	-	Typewritten
80	-	750	70.50	-	Typewritten
		1000	82.31	-	Typewritten

The quality and size of the test data set is the only resource which can be used to predict the reliability of a *TCR* technique. A large test data set (more than 1000 character/class) that reflects factors affecting the Tamil text production process is required to estimate the reliability of a *TCR* technique. The common factors that affect the text production process are: font type, pen type, paper texture, paper color, ink color, writing style, writing environment and writers' mood.

References

[1] Punitharaja K. and Elango P., "Effective Handwritten Tamil Character Recognition using SVM and MQDF" Malaya Journal of Matematik, S(2) 2015, S(2) pp. 504-512.

[2] AsmaaQasimShareef, "OCR-ANN Back-Propagation Based Classifier", *International Journal of Computer Science and Mobile Computing*, 2015. pp. 307-314.

[3] Aadesh Neupane, "Development of Nepali Character Database for Character Recognition based on Clustering", *International Journal of Computer Applications*, 2014. pp. 42-46.

[4] Ahmed T. Sahlol, "A Proposed OCR Algorithm for the Recognition of Handwritten Arabic Characters", *Journal of Pattern Recognition and Intelligent Systems*, 2014. pp. 90-104.

[5] Punitharaja K. "Hand Written Tamil Character Recognition Using Wavelet-LDA Feature Selection and PNN Classification" *International Conference on Advances in Engineering and Technology*, 2011.

[6] Kumar S. and Singh C., "A Study of Zernike Moments and its use in Tamil Handwritten Character Recognition", *In Proc. International Conference on Cognition and Recognition*, 2005, pp. 514-520.

[7] Pal U. and Chaudhuri B. B., "Indian script character recognition: A Survey", *Pattern Recognition*, Vol. 37, pp. 1887-1899, 2004.

[8] Pal U., Roy K. and Kimura F., "A Lexicon Driven Method for Unconstrained Bangla Handwritten Word Recognition", *In 10th International Workshop on Frontiers in Handwriting Recognition*, 2006, pp. 601-606.

[9] Pal U., Sharma N., Wakabayashi T. and Kimura F., “Off-Line Handwritten Character Recognition of Tamil Script”, *In Proc. 9th International Conference on Document Analysis and Recognition*, 2007, pp. 496-500.

[10] Pal U., Sharma N., Wakabayashi T. and Kimura F. Kimura, “Handwritten Numeral Recognition of Six Popular Indian [11] Scripts”, *In Proc. 9th International Conference on Document Analysis and Recognition*, 2007, pp. 749-753.

[12] Plamondon R and Srihari S. N., “On-Line and off-line handwritten recognition: A comprehensive survey”, *IEEE Trans on PAMI*, Vol. 22, pp. 62-84, 2000.

[13] Punitharaja K. “Hand Written Tamil Character Recognition Using Wavelet-LDA Feature Selection and PNN Classification” *International Conference on Advances in Engineering and Technology*, 2011

[14] Sharma N., Pal U., Kimura F. and Pal S., “Recognition of Offline Handwritten Tamil Characters using Quadratic Classifier”, *In Proc. Indian Conference on Computer Vision Graphics and Image Processing*, 2006, pp. 805-816.

[15] Shi M., Fujisawa Y., Wakabayashi T., and Kimura F., “Handwritten numeral recognition using gradient and curvature of gray scale images”, *Pattern Recognition*, Vol. 35, pp. 2051-2059, 2002.

[16] Muhammad Naeem Ayyaz, Imran Javed, Waqar Mahmood, “Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction”, Pakistan journal of Engineering and Application Science, Vol. 10, pp. 57-67, Jan–2012.

[17] Pranob K. Charles, V. Harish, M. Swathi, CH. Deepthi, “A Review on the Various Techniques used for Optical Character Recognition”, International Journal of Engineering Research and Applications, Vol. 2, Issue 1, pp. 659-662, Jan-Feb 2012.

[18] Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, “An Improved Zone Based Hybrid Feature Extraction Model for Handwritten Alphabets Recognition Using Euler Number”, International Journal of Soft Computing and Engineering, Vol. 2, Issue 2, pp. 504-58, May 2012.

[19] J. Pradeepa, E. Srinivasana, S. Himavathib, “Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten”, International journal of Engineering, Vol. 25, No. 2, pp. 99-106, May 2012.