CHAPTER 1

INTRODUCTION

1.1 PREAMBLE

Recognition is regarded as a basic attribute of human beings, as well as other living organisms. A pattern is the description of an object. Performing the acts of recognition in every walk of our lives, such as recognition of objects around us, recognizing a friend in a crowd, recognizing voice of a known individual and recognizing handwritten characters.

According to the nature of the patterns to be recognized, we may divide our acts of recognition into two major types, namely, the recognition of concrete items and the recognition of abstract items. The concrete items of recognition include spatial patterns like characters, finger prints, weather maps, objects etc and temporal patterns like speech waveforms, electrocardiograms, target signatures, and time series Tou et al (1974). An example of recognition of an abstract item is the recognition of solution to a problem when your eyes and ears are closed. However we are concerned only with recognition of concrete items. Recognition of concrete patterns by human beings may be considered as a physiological problem which involves a relationship between a person and a physical stimulus. When a person perceives a pattern, he makes an inductive inference and associates this perception with some general concepts or clues which he has derived from his past experience. Human recognition in reality is a question of estimating the relative odds that the input data can be associated with one of a set of known statistical populations which depend on our past experience and which form
the clues and the a priori information for recognition. Thus, the problem of pattern recognition may be regarded as one of discriminating the input data not between individual patterns but between populations, via the search for features or invariant attributes among members of a population.

Hence, Pattern Recognition can be defined as the classification of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant details. A Character recognition system is a pattern recognition system which receives scanned preprocessed data as input and identifies the name of the character by comparing it with the prototype characters already in the memory by Annadurai et al (2007).

The study of pattern recognition problems may be logically divided into two major categories Ullman (1982)

- The study of the pattern recognition capability of human beings and other living organisms.
- The development of theory and techniques for the design of devices capable of performing a given recognition task for a specific application.

The first area is concerned with disciplines such as psychology, physiology and biology. The second area deals primarily with Engineering, Computer and Information Sciences. With the rapid proliferation of computers and with newer and newer technology, scientists are trying to communicate with machines in their natural mode. Pattern recognition scientists are working with a vision to build a machine with new technology and methodology, which can identify and understand speech and process pictorial information for human use with more and more efficiency.
A lot of research has been going on since 1928 in the area of Man-
Machine communication such as

a. Automatic speech recognition
b. Cursive script recognition
c. Natural Language Processing
d. Optical Character Recognition Systems
e. Speaker/Speech Identification/Understanding

These researches, since 1956, led to the development of a number
of OCR, Voice Recognizer and Voice Synthesizer. But the performances of
these systems are controlled by many constraints. Any deviations from these
constraints cause a large deterioration in the specified performance Avi-Itzhak
et al (1995). The true success, that is, the ability to recognize patterns with the
same fluency as human beings, remains to be a distant dream.

In this thesis the problem of recognizing handwritten Tamil
characters has been taken up and various approaches are presented.

1.2 PROBLEM DEFINITION

The problem considered is more specifically the recognition of the
subset of handwritten Tamil characters and numerals.

A Hidden Markov Model (HMM), Octal Graph approach and
Structural Hidden Markov Model (SHMM) based recognition schemes are
developed and tested on three databases including the ones from Hewlett
Packard Labs (HPL - http://www.hpl.hp.com/india/research/penhw-
resources/tamil-iso-char.html), National Institute of Standard and Technology
(NIST - http://yann.lecun.com/exdb/mnist/) and on the dataset generated
locally. The results are seems to be better than the previously reported ones.
1.3 OBJECTIVE

- The primary objective of this research work is to design and develop a character recognition system, which can further be developed to recognize all printed and handwritten Tamil characters & numerals.
- The system could be extremely useful to digitize innumerable ancient documents available both on paper and palm leaves.
- Therefore, this attempt here is a humble effort towards developing an efficient recognition system, in public domain, for some Tamil characters and numerals.
- The system designed and developed is a constrained character recognition system.

1.4 MOTIVATION

Optical Character Recognition (OCR) systems have been under research for few decades. But still it remains a highly challenging task to implement an OCR that works under all possible conditions and gives highly accurate results. In this research work, a simple character recognition system for subset of handwritten Tamil Characters and numerals is developed.

The motivation for this work is mainly due to the amount of effort that has to be spent in achieving unambiguous decision logic, and less effort on structural logic for the development of an automatic recognition system. The two other components of the system namely feature extractor and recognizer can be made structurally flexible using proper control structures so that the changes in the structure of the features can be handled by the use of appropriate procedures from a procedure inventory of the automatic design. It is the class description (Decision Logic) which is to be changed according to the character sets. So, if it is possible to learn the reference class descriptions
automatically when samples of each character classes are given, the automation of the recognition system software for the cursive characters is not difficult.

1.5 SIGNIFICANCE

In recent days the computers are used in Government and non-government organizations for variety of applications. Therefore it is necessary for the software developers to meet the requirements of the government officers such as operating the system in their regional languages. Hence the language translators and machine translation is very much required. Work on Indian languages is not as easy as it looks it requires different strategies compared with western languages. Some of the issues of Indian languages made this work significant. Such are:

- All major Indian scripts are mixtures of syllabic and alphabetical scripts.
- Each language has vowels and consonants separately and these forms more than 250 compound characters.
- There is a very high variance in the width of fonts to be considered for recognition due to the presence of ligatures and modifiers in the font set.
- There are many multi-component characters in Dravidian text lines. The components like matras and halants occupy positions over or under the main character.
- There are several characters that touch each other. Some touching is due to dark printing or low resolution.
• Though there is a rule regarding the font style, every individual writes in his own way.

• The century old Tamil character fonts are very different from the one currently being used.

• Hence shape discrimination becomes very difficult.

• Deformation of image such as disconnected line segments, change of orientation needs to be attended.

• Variation in size and pitch are to be taken care.

1.6 PATTERN RECOGNITION APPROACHES

The different methods for automatic pattern recognition are motivated by the way in which pattern classes are characterized and defined. Three such ways are described now. Firstly, Characterization of a pattern class by a roster of its members suggests automatic pattern recognition by template matching. On the other hand, characterization of a pattern class by common properties shared by all of its members suggests automatic pattern recognition via the deduction and processing of similar features. If all the features of the class can be determined from sample patterns then it is called as feature matching design. Finally, when the patterns of a class are vectors whose components are strings, then a pattern class can be recognized by its clustering properties.

These basic design concepts for automatic pattern recognition described above are being implemented by the following principal categories of methodology.
1.6.1 Heuristic Approach

It is based on human intuition and experience, making use of template matching and feature matching concepts. System design using this principle generally consists of a set of adhoc procedures Hagita et al (1987), Suresh et al (1999). The structure and performance of a heuristic system depends on a large degree on the cleverness and experience of the system designers.

1.6.2 Mathematical Approach

This is based on classification rules which are formulated and derived in a mathematical framework making use of the feature matching and clustering concept. The main difference between heuristic approach and this approach is that the former is based on adhoc rules and the later is based on classification rules.

1.6.3 Linguistic (Syntactic or Structural) Methods

Characterizing of patterns by primitive elements (Sub-patterns) and their relationships suggests automatic pattern recognition by the linguistic or syntactic or structural approach, making use of the common – property concept. A pattern can be described by a hierarchical structure of sub-patterns analogous to the syntactic structure of languages. This permits application of formal language theory to the pattern recognition problems. A pattern grammar is considered as consisting of finite sets of elements called variables, primitives, and productions. The rules of production determine the type of grammars. The well referred grammars are Regular Grammars, Context-Free Grammars, and Context-Sensitive Grammars Edwards et al (1964). The essence of this approach lies in the selection of pattern primitives, the assembling of the primitives and their relationships into pattern grammars and
analysis and recognition in terms of these grammars. This approach is particularly useful in dealing with patterns which cannot be conveniently described by numerical measurements or are so complex that local features cannot be identified and global properties must be used.

The earliest advocates to this approach are Narasimhan et al (1964), and Eden (1968). Eden is the first to demonstrate that a set of primitive elements together with a set of rules for combining them could be used to generate script, which was virtually indistinguishable from human handwriting.

The design of a classifier, of course, requires previous knowledge of feature space or the parameters for classification. Even if the feature space is not known, the boundary of class estimation is important. Hence, it is important to note that learning or training takes place only during the design phase of a pattern recognition system. All learning activities require the assistance of a set of samples from the feature space, which is called the training set. If the correct labels of the samples in the training set are known, the learning that takes place with the help of these samples is called supervised learning. In some applications, only a set of training patterns of unknown classification may be available. In these situations unsupervised pattern recognition techniques are applicable.

Supervised pattern recognition Biwas et al (1983) is characterized by the fact that the correct classification of every training pattern is known. In the unsupervised case, however, one is faced with the problem of actually learning the pattern classes present in the given data. Of course, the quality of recognition performance will be largely determined by how closely the training patterns resemble the actual data with which the system will be confronted during normal operation.
The design of an automatic pattern recognition system generally involves several major problem areas. The first one is concerned with the representation of input data which can be measured from the objects to be recognized.

The second problem concerns the extraction of characteristic features or attributes from the received input data and the reduction of dimensions of pattern vectors.

The third problem involves the determination of the optimum decision procedures, which are needed in identification and classification process. After the observed data from patterns to be recognized have been expressed in the form of pattern points or measurement vectors in the pattern space, the machine would decide to which pattern class these data belong.

Solving the preprocessing Chaudhri et al (1993) and feature extraction Calvert (1970) problem and the optimum decision Fu et al (1967) and classification Mohiuddin et al (1994) problem generally involve a set of parameters that must be estimated and optimized. This gives rise to the parameter estimation problem. Furthermore, it is conceivable that both the feature extraction process and the decision-making process may be considerably improved by making use of contextual information in the patterns. Contextual information can be measured by contingent probabilities, language statistics and neighboring variation. In some applications contextual information is indispensable in achieving accurate recognition. For instance, fully automatic speech recognition is possible only when contextual and linguistic information is available to supplement the information of the speech sound wave. Similarly in the recognition of handwritten, cursive handwritten characters and finger prints, contextual information is extremely desirable Sinha et al (1993).
The dichotomy of syntactic and statistical approaches Kam-Fai Chan et al (1999), Lam et al (1988) are more apparent. The preprocessing techniques like feature extraction / primitive extraction are similar in both approaches. The basic difference between the two approaches is that in statistical approach the features are a set of numerical measurements on the pattern and sub-pattern parameters whereas the primitives in syntactic approaches are sub-patterns themselves.

When the patterns are very rich in structural information and the recognition problem requires classification and description, syntactic approach is a helpful complement to the decision theoretic or any other approach.

1.6.4 Fuzzy Set Theoretic Approach

Fuzzy set theory, first presented to the scientific society by Zadeh (1965) has since become an area of fairly vigorous research. It is a generalization of classical set theory, in the sense that given an universe X and a subset A of it, any element x of X, instead of having a degree of membership either 0 or 1 in A as postulated under the classical set theory, can have a membership value $\mu_A(x) \in [0,1]$, which represents the degree of its belonging to A. Fuzzy set theory plays at least two roles in pattern recognition. In the first, it serves as an interface between linguistic variables seemingly performed by humans and quantitative characterizations appropriate for machines. In this role, it might also serve as a bridge between the symbolic processing of Artificial Intelligence and the parallel distributed processing approaches performed by adaptive pattern recognition Kochi Wada et al (1998). In the second role, it emphasizes the possibility-distribution interpretation of the concept of fuzziness that it legitimizes and
provides a meaningful interpretation of certain distributions believed to be useful, but difficult to justify on the basis of objective probabilities.

The most obvious manner in which fuzzy set theory is applied to pattern recognition Mazumdar et al (1982), is to treat the classes as fuzzy sets. Hence each pattern does not necessarily belong to one class. Rather there is a certain degree of possibility that it belongs to each of the classes, the extent being measured in terms of the membership functions. If the information processing ceases immediately after classification, then nothing extra can be gained by using fuzzy set theory, since one has to ultimately decide in favor of just one class, possibly the one for which the unknown pattern has the maximum membership value.

Fuzzy set membership values can also be used to help, provide estimates or incomplete information. Kandel et al (1979), Pal et al (1986) and Bezde (1992) among others provide a good survey of the extent to which this fusion has been made.

1.6.5 Neural Network Classifiers

In general neural networks are composed of many nonlinear computational elements operating in parallel and arranged in patterns reminiscent of biological neural nets Wasserman (1990), Young (1988). The computational elements or nodes are connected via weights that are typically adapted during use to improve performance. The ability to adapt and continue learning is not only a highly desirable characteristic as far as practical pattern recognition is concerned; it also provides a degree of robustness by compensating for minor variations in the characteristics of processing elements. Neural net classifiers are also non parametric and make weaker assumptions regarding the underlying distributions than traditional statistical classifiers. As such they can be expected to be more robust in Non-Gaussian situations.
Of the commonly known neural networks, the most important ones that can be used as classifiers apart from Single-layer Perceptrons and Multi-layer Perceptrons networks Stringa (1989) are:

- The Hopfield nets
- The Hamming nets
- The Carpentor / Groosberg classifiers
- The Kohonen Self-organizing feature maps

The Hopfield and Hamming nets can be trained with supervision, but are generally used with fixed weights. The Hamming net is neural network implementation of the optimum classifier for binary patterns corrupted by random noise. The Carpentor/Grossberg classifiers do unsupervised learning, the former by the leader algorithm and the latter by K-means algorithm.

1.7 HISTORY OF OPTICAL CHARACTER RECOGNITION

The world is filling with computers. They are becoming ubiquitous. As ever more people are forced into contact with computers and our dependence upon them continues to increase, it is essential that they become easier to use. As more of the world’s information processing is done electronically, it becomes more important to make the transfer of information between people and machines simple and reliable. In spite of the major effort that has been expended to bring about a paper-free society, a very large number of paper-based documents are processed daily by computers all over the world in order to handle, retrieve, and store information. The problem is that the manual process used to enter the data from these documents into computers demands a great deal of time and money. The field of Document Analysis and Recognition (DAR) has played a very important role in the attempt to overcome this problem Flavio Bortolozzi et al (2005). The general
objective of DAR research is to fully automate the process of entering and understanding printed or handwritten data into the computer.

Handwriting is a means of communication which everyone nearly learns at an early stage. Handwriting has continued to persist as a means of communication and recording information in day-to-day life even with the introduction of new technologies. Writing is considered to have made possible much of culture and civilization. Each script has a set of icons, which are known as characters or letters, having definite basic shapes. There are rules for combining letters to represent shapes of higher level linguistic units. Automated recognition of handwritten text has been an active subject of research since the early days of computers. Despite the age of the subject, it remains one of the most challenging and exciting areas of research in computer science. In recent years it has grown into a mature discipline, producing a huge body of work.

Despite long standing predictions that handwriting, and even paper itself, would become obsolete in the age of the digital computer, both persist. Whilst the computer has hugely simplified the process of producing printed documents, the convenience of a pen and paper still makes it the natural medium for many important tasks. The reason that handwriting persists in the age of the digital computer is the convenience of paper and pen as compared to keyboards for numerous day-to-day situations. Widespread acceptance of digital computers seemingly challenges the future of handwriting. However, in numerous situations, a pen together with paper or a simple notepad is much more convenient than a keyboard. Rejean Plamondon et al. (2000). A brief survey of students in any lecture theatre confirms the dominance of handwritten notes over those typing on laptops. However, the ease and convenience of having information in digital form provides a powerful incentive to find a way of quickly converting handwritten text into its digital
The task of handwriting recognition is the transcription of handwritten data into a digital format. The goal is to process handwritten data electronically with the same or nearly the same accuracy as humans. By doing the processing with computers, a large amount of data can be transcripted at a high speed.

Writing, which has been the most natural mode of collecting, storing and transmitting the information through the centuries, now serves not only for the communication among humans, but also, serves for the communication of humans and machines. The intensive research effort on the field of Character Recognition (CR) was not only because of its challenge on simulation of human reading, but also, because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents. Historically, CR systems have evolved in three ages Nafiz Arica et al (2001):

1870 – 1980 Early ages - The origins of character recognition can be found in 1870 when Carey invented the retina scanner, which is an image transmission system using a mosaic of photocells, and later in 1890 when Nipkow invented the sequential scanner which was a major break through both for modern television and reading machines. However, character recognition first appeared as an aid to the visually handicapped and the first successful attempts were made by the Russian scientist Tyurin in 1900 Mantas (1986). The first character recognizers appeared in the middle of the 1940s with the development of the digital computers. The early work on the automatic recognition of characters has been concentrated either upon machine printed text or upon small set of well-distinguished handwritten text or symbols. Machine-printed CR systems in this period generally used template matching in which an image is compared to a library of images. For handwritten text, low level image processing techniques have been used on
the binary image to extract feature vectors, which are then fed to statistical classifiers. Successful, but constrained algorithms have been implemented mostly for Latin characters and numerals. However, some studies on Japanese, Chinese, Hebrew, Indian, Cyrillic, Greek and Arabic characters and numerals in both machine-printed and handwritten cases were also initiated Nafiz Arica et al (2001).

The commercial character recognizers were available in 1950s, when electronic tablets capturing the x-y coordinate data of pen-tip movement was first introduced. This innovation enabled the researchers to work on the on-line handwriting recognition problem. A good source of references for on-line recognition until 1980 can be found in Tappert (1984).

**1980 – 1990 Developments** - The studies until 1980 suffered from the lack of powerful computer hardware and data acquisition devices. With the explosion on the information technology, the previously developed methodologies found a very fertile environment for rapid growth in many application areas, as well as CR system development. Structural approaches were initiated in many systems in addition to the statistical methods. These systems broke the character image into a set of pattern primitives such as lines and curves. The rules were then determined which character most likely matched the extracted primitives. However, the CR research was focused on basically the shape recognition techniques without using any semantic information. This led to an upper limit in the recognition rate, which was not sufficient in many practical applications. Historical review of CR research and development during this period can be found in Mori et al (1992) and Suen et al (1990) for off-line and on-line case, respectively.

**After 1990 Advancements** - The real progress on CR systems is achieved during this period, using the new development tools and methodologies, which are empowered by the continuously growing
information technologies. In the early nineties, Image Processing and Pattern Recognition techniques are efficiently combined with the Artificial Intelligence methodologies. Researchers developed complex CR algorithms, which receive high-resolution input data and require extensive work in the implementation phase. Nowadays, in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets, efficient modern methodologies such as Artificial Neural Networks (ANN), Fuzzy Set Reasoning (FSR) and Natural Language Processing (NLP) are available. The recent systems for the machine-printed off-line and limited vocabulary, user dependent on-line handwritten characters are quite satisfactory for restricted applications. However, there is still a long way to go in order to reach the ultimate goal of machine simulation of fluent human reading, especially for unconstrained on-line and off-line handwriting.

### 1.8 INTRODUCTION TO CHARACTER RECOGNITION

Character Recognition techniques associate a symbolic identity with the image of a character. This problem of replication of human functions by Computer Systems involves the recognition of both printed (Machine and hand) and handwritten characters.

The problem of handwritten character recognition is an interesting and challenging task to researchers because even the human beings, who possess the best trained optical readers, namely, their eyes, make about 4 percent mistakes when reading in the absence of context. The machine replication of human reading has been the subject of intensive research for more than four decades. A large number of research papers and reports have already been published on this topic. Many commercial establishments have manufactured recognizers of varying capabilities Srihari (1992). Hand-held, desktop, medium-size and large systems costing as high as half a million
dollars are available, and are in use for various applications Siromoney et al (1978). However, the ultimate goal of developing a reading machine having the same capability as human being is unachievable till date. Hence an effort is required to either bridge or narrow-down this gap.

1.9 CLASSIFICATION OF CHARACTER RECOGNITION SYSTEMS

Character recognition systems can be broadly classified according to the data acquisition techniques, text type and the recognition type. The classification of CR systems is shown in Figure 1.1.

![Figure 1.1 Classification of Character Recognition Systems](image)

1.9.1 Systems Classified according to the Data Acquisition Techniques

The progress in CR methodologies evolved in two categories according to the mode of data acquisition.

- On-line character recognition systems.
- Off-line character recognition systems.
Online Character Recognition System

On-line recognition refers to methods and a technique dealing with the automatic processing of a message as it is written using a digitizer or an instrumental stylus that captures information about the pen-tip, generally its position, velocity, or acceleration as a function of time Rejean Plamondon et al (2000). The digitizers are mostly electromagnetic – electrostatic tablets, which send the coordinates of the pen tip to the host computer at regular intervals. Some digitizers use pressure-sensitive tablets, which have layers of conductive and resistive material with a mechanical spacing between the layers. There are also, other technologies including laser beams and optical sensing of a light pen. The on-line handwriting recognition problem has a number of distinguishing features, which must be exploited to get more accurate results than the off-line recognition problem.

Advantages of on-line character recognition

1. **It is a real time process.** Since the digitizer captures the data during the writing, the CR system with or without a lag makes the recognition.

2. **It is adaptive in real time.** The writer gives immediate feedback to the recognizer for improving the recognition rate, as he keeps drawing the symbols on the tablet and observes the results.

3. **It captures the temporal and dynamic information of the pen trajectory.** This information consists of the number and order of pen-strokes, the direction of the writing for each pen-stroke and the speed of the writing within each pen stroke.
4. **Very little pre-processing is required.** The operations, such as smoothing, de-slanting, de-skewing, detection of line orientations, corners, loop and cusps are easier and faster with the pen trajectory data than on pixel images.

5. **Segmentation is easy.** Segmentation operations are facilitated by using temporal and pen-lift information, particularly, for hand-printed characters.

**Disadvantages of on-line character recognition**

1. The writer requires special equipment, which is not as comfortable as pen and paper.
2. It cannot be applied to documents printed or written on papers.
3. The available systems are slow and recognition rates are low for handwriting that is not neat.

Applications of on-line character recognition systems include small hand-held devices, which call for a pen-only computer interface and complex multimedia systems, which use multiple input modalities including scanned documents, speech, keyboard and electronic pen. On-line character recognition systems are useful in social environments where speech does not provide enough privacy. They provide an efficient alternative for the large alphabets where the keyboard is cumbersome. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the on-line character recognition techniques.
Off-line Character Recognition System

Off-line character recognition is known as Optical Character Recognition (OCR), because the image of writing is converted into bit pattern by an optically digitizing device such as optical scanner or camera. The recognition is done on this bit pattern data for machine-printed or handwritten text. The research and development is well progressed for the recognition of the machine-printed documents. In recent years, the focus of attention is shifted towards the recognition of hand-written script. The major advantage of the off-line recognizers is to allow the previously written and printed texts to be processed and recognized. The drawbacks of the off-line recognizers, compared to on-line recognizers are summarized as follows:

1. Off-line conversion usually requires costly and imperfect preprocessing techniques prior to feature extraction and recognition stages.

2. The lack of temporal or dynamic information results in lower recognition rates compared to on-line recognition.

Some applications of the off-line recognition are large-scale data processing such as postal address reading, cheque sorting and office automation for text entry, automatic inspection and identification. Off-line character recognition is a very important tool for creation of the electronic libraries. It provides a great compression and efficiency by converting the document image from any image file format into more useful formats like HTML or various word processor formats. Recently, content based image or video database systems make use of off-line character recognition for indexing and retrieval, extracting the writings in complex images. Also, the wide spread use of web necessitates the utilization of off-line recognition systems for content based Internet access to paper documents. Automatic
assessment of essays is another area where off-line recognition systems are useful.

1.9.2 Systems Classified According to the Text Type

Considering the text type, hand-written and machine-printed character recognition systems are two main areas of interest in the CR field. Machine-printed text includes the materials such as books, newspapers, magazines, documents and various writing units in the video or still image. The problems for fixed-font and multi-font character recognition is relatively well understood and solved with little constraint. When the documents are generated on a high quality paper with modern printing technologies, the available systems yield as good as 99% recognition accuracy. However, the recognition rates of the commercially available products are very much dependent on the age of the documents, quality of paper and ink, which may result in significant data acquisition noise. On the other hand, hand-written character recognition systems still have limited capabilities. The problem can be divided into two categories: cursive and normal script. In practice, however, it is difficult to draw a clear distinction between them. A combination of these two forms can be seen frequently. Based on the nature of writing and the difficulty of segmentation process, there are five stages for the problem of handwritten word recognition Nafiz Arica et al (2001) as indicated in Figure 1.2.

1. Boxed discrete characters
2. Spaced discrete characters
3. Run-on discretely written characters
4. Run cursive script writing
5. Mixed cursive and discrete

Figure 1.2 Five Stages of Handwritten Word Recognition Problem
Boxed discrete characters require the writer to place each character within its own box on a form. The boxes themselves can be easily found and dropped out of the image or can be printed on the form in a special color ink that will not be picked up during scanning, thus eliminating the segmentation problem entirely. Spaced discrete characters can be segmented reliably by means of horizontal projections, creating a histogram of gray values in the image over all the columns and picking the valleys of this histogram as the points of segmentation. This has the same level of segmentation difficulty as is usually found with clean machine-printed documents. Characters at the third stage are usually discretely written, however they may be touching, therefore making the points of segmentation less obvious. Degraded machine printed characters may also be found at this level of difficulty. There has been a fairly extensive and successful research on the first three stages.

Cursive or mixed written texts require more sophisticated approaches compared to the previous cases. First of all, advanced segmentation techniques are to be used for character based recognition schemes. In pure cursive handwriting, a word is formed mostly from a single stroke. This makes segmentation by the traditional projection or connected component methods ineffective. Secondly, shape discrimination between a character that look alike, are also difficult and requires the context information. In Tamil language mostly the first three types of handwriting is used. Few users will write using the other two approaches. This thesis mainly focuses on the second type of handwriting. A good source of references in hand-written character recognition can be found in Steinherz et al (1999), Rejean Plamondon et al (2000), Mohamed Cheriet et al (2007).
1.9.3 Systems Classified According to the Recognition Type

There are two main approaches for recognizing handwriting recognition system namely global approach and analytical approach. In global approach the words are identified directly, instead of reading them letter by letter. In analytical approach each word is partitioned into discrete segments, letters or digits, and recognizing them through classification process. Though the global approach is more human manner to identify word directly, instead of reading them letter by letter, the number of classes increases substantially in the case of standard vocabularies. This research work focuses on offline handwriting recognition using analytical approach.

1.10 STEPS IN HANDWRITTEN CHARACTER RECOGNITION

The major steps involved in handwriting recognition system suggested by Annadurai et al (2007) are shown in Figure 1.3.

The steps involved are

1. Preprocessing
2. Feature Extraction
3. Recognition
4. Post processing
1.10.1 Preprocessing

The raw input of the digitizer typically contains noise due to erratic hand movements and inaccuracies in digitization of the actual input. Original documents are often dirty due to smearing and smudging of text and aging. In some cases, the documents are of very poor quality due to seeping of ink from the other side of the page and general degradation of the paper and ink. Preprocessing is concerned mainly with the reduction of these kinds of noise and variability in the input. The number and type of preprocessing algorithms employed on the scanned image depend on many factors such as paper quality, resolution of the scanned image, the amount of skew in the image and the layout of the text. Various preprocessing operations are performed prior to recognition to enhance the quality of the input image.

1.10.2 Feature Extraction

Feature Extraction is the problem of extracting from the preprocessed data, the information, which is most relevant for classification.
purposes, in the sense of minimizing the within-class pattern variability, while enhancing the between-class pattern variability Trier et al (1996). In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. Due to the nature of handwriting with its high degree of variability and imprecision, obtaining these features are a difficult task. The task of the human expert is to select features that allow effective and efficient recognition of patterns. There are different feature extraction methods available. The features are categorized into geometric features, structural features Mohamed Cheriet et al (2007), and features extracted from transformation method.

1.10.3 Recognition

The final goal of character recognition is to obtain the class codes of character patterns. On segmenting character patterns or words from document images, the task of recognition becomes assigning each character pattern or word to a class out of a pre-defined class set. Pattern classification has been the main theme of pattern recognition field and is often taken as a synonym for “Pattern Recognition (PR)”. Traditionally, pattern recognition techniques are classified as template and feature-based approach Pal et al (2004). In the template-based approach, an unknown pattern is superposed directly on the ideal template pattern and the degree of correlation between the two is used for the decision about classification. Early OCR systems employed only template-based approach, but modern systems combine this with feature-based approaches to obtain better results.

Feature-based approaches derive important properties (features) from the test patterns and employ them in a more sophisticated classification model. The feature-based approaches can be of two types, namely spatial domain and transform domain approaches. Spatial domain approaches derive
features directly from the pixel representation of the pattern. In a transform
domain technique, the pattern image is at first transformed into another space
using Fourier, Cosine, Slant or Wavelet transform and useful features are
derived from the transformed images.

Syntactic or formal grammars, moment-based as well as graph
theoretic approaches were also tested for OCR problems. A modern group of
techniques have evolved that do not explicitly derive any feature from the
patterns. During training phase either raw or normalized patterns are fed to
such a system, and the system adjusts itself to minimize the misclassification
error of these patterns. Artificial neural network is an example of such a
system which adjusts the weights of its links from the training patterns. These
weights implicitly work as features for classification.

Another example of non-explicit feature-based method is Hidden
Markov Model. Statistically derived parameters play vital role in this
approach and it needs a very large number of training samples to estimate the
probability parameters in a reliable manner.

1.10.4 Post Processing

Until this point no semantic information is considered during the
stages of CR. It is well known that humans read by context up to 60% for
careless handwriting. While preprocessing tries to “clean” the document in a
certain sense, it may remove important information, since the context
information is not available at this stage. The lack of context information
during the segmentation stage may cause even more severe and irreversible
errors, since it yields meaningless segmentation boundaries. It is clear that if
the semantic information were available to a certain extent, it would
contribute a lot to the accuracy of the CR stages. On the other hand, the entire
CR problem is for determining the context of the document image. Therefore
utilization of the context information in the CR problem creates a chicken and egg problem. The review of the recent CR research indicates minor improvements, when only shape recognition of the character is considered. Therefore, the incorporation of context and shape information in all the stages of CR systems is necessary for meaningful improvements in recognition rates. This is done in the post processing stage with a feedback to the early stages of CR. The incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates.

1.11 METHODOLOGY

An overview of the methodologies adopted in this research work is presented below.

A system is developed to recognize handwritten Tamil characters using a Hidden Markov Model (HMM) approach, for a subset of the Tamil characters. The image is scanned using standard scanner. The output of the scanned image is preprocessed. In preprocessing the image is digitized. The digitized image is segmented as line, word and characters. The segmented characters are pre-classified. A character image is considered for feature extraction using image glyph algorithms. The feature set (Image glyph) of each character is stored in the database. The database of the image glyph is created. The image glyphs of handwritten characters are considered for recognition. The glyph set is used for pre-recognition based on these features. The extracted features are passed to HMM where the characters are recognized.

The system uses Octal Graph conversion to recognize a subset of handwritten Tamil characters. In this work, the segmentation process separates the individual characters from the input. The segmented characters
are size normalized in to fixed dimensions. The normalized image is converted into an octal graph so that it represents the features of the character to a good extent. Then the features are extracted for the handwritten characters. Compare the features of each character with that of the characters in the repository. If the features are matched, then the character is recognized.

A novel solution is proposed to recognize handwritten Tamil characters and numerals using Structural Hidden Markov Model (SHMM). Here the image processing phase is skipped because the MNIST database is used for recognition. The feature extraction is done by the chain codes. The extracted features are the sequences of integers from 0 – 7. These features of sequences are used to find the structure of the character in terms of various strokes. After extracting the strokes from the chain code it is grouped. The extracted features are fed to SHMM where the characters are recognized.

1.12 CONTRIBUTIONS

The main theme of the work is to design and develop a recognition system and to automate such recognizers for different handwritten character sets, especially for Numerals, and characters of Tamil Language using structural approaches. This has been accomplished by doing the following:

- Automating the two tone conversion
- Automating the segmentation
- Automating the feature extraction
- Automating and adopting a recognition system for character recognition.
All the above steps have been implemented using Hidden Markov Models (HMM), Octal Graph approach and Structural Hidden Markov Model (SHMM) for handwritten Tamil characters.

However, some of the difficulties in the determination of reference class descriptions for recognition of characters have been overcome using the following steps.

1. Defined and generated the prototype character set with subset of characters.
2. Collected enormous samples of each class of the character set.
3. Carefully examined the samples for reliable/stable features of the each class.
4. Wrote the class descriptions in terms of the features selected.
5. Verified the ambiguity of each of the class descriptions.
6. If any of the classes are found ambiguous, define new features and repeat steps 2-4 until ambiguity becomes zero.

The steps described above are normally carried out manually by human employing his/her power of reasoning and capacity of inference. Depending upon the complexity of the character set it may even involve a number of days or a few months to arrive at the right class descriptions free of ambiguities. On the other hand, if steps 2 – 6 are made to carry out by machines, unambiguous class descriptions can be generated for a character set in minutes. Even though, the full reasoning capabilities of human beings cannot be incorporated with machines; they will carry out the job without causing any lapse on their part.
1.13 HARDWARE AND SOFTWARE REQUIREMENTS

Hardware

System Configuration: Pentium IV 2.6GHz, 512MB RAM.

Scanner: Cannon – Cano scan LiDE 25 (F910111) 300 DPI

Software Specifications

1. Operating System – Microsoft Windows XP
2. Compiler: VC++6.0.

1.14 APPLICATIONS OF CHARACTER RECOGNITION SYSTEMS

Optical character recognition technology has many practical applications. The importance of any work in the field of character recognition can be seen from the real world applications listed below.

- For blind people – as reading aid using photo sensor and tactile simulators, and as a sensory aid with sound output.
- Telecommunication aid for deaf.
- In postal department – for postal address reading and as a reader for handwritten and printed postal codes.
- For character print quality analysis/measurement, document reading and sorting, and dictionary generation.
- In publishing industry, as a reader for data communication terminal.
- In direct processing of documents – as a multipurpose document reader for large-scale data processing, as a micro-film reader
data input systems, for high speed data entry, for changing
text/graphics into a computer readable form, as electronic page
reader to handle large volume of mail.

- In customer billing, as in telephone exchange billing system,
- In Banking, for automated finger print identification,
- In business applications – financial business applications like
cheque sorting strategy optimization.
- For digital bar code reading.
- Handwriting analyzer for automatic writer recognition and
signature verification.
- Use in health insurance data acquisition.
- For mechanized document reading in textile and clothing
manufacture enterprises, automatic punching of industrial
telegraphs, retail data processing applications in food
enterprises, and for retail product code name and price reading
techniques.
- In law enforcement applications.
- In educational administrations – examinations assessment and
attendance record evaluation.
- As mark sheet reader for payroll accounting and book-keeping.
- In optical census.
- In automated cartography, metallurgical industries, computer
assisted forensic linguistic system, electronic mail, information
units and libraries and for facsimile.
- For shorthand transcription.
1.15 ORGANIZATION OF THE THESIS

Chapter 1 introduces the subject of character recognition, the history of OCR, various steps in character recognition, and its applications. Literature review and objectives of the research work are presented in this chapter.

Chapter 2 describes the recognition of printed and handwritten Tamil characters using Hidden Markov Models (HMM).

Chapter 3 describes the architecture of the Octal Graph approach to recognize the handwritten Tamil Characters.

Chapter 4 illustrates a method to recognize the handwritten characters using Structural Hidden Markov Models (SHMM).

The last chapter gives the conclusion of the work done and scope for future research work.