CHAPTER I

I N T R O D U C T I O N

1.1. What is Pattern Recognition ?

We are all the time performing the act of recognition, which is regarded as a basic attribute of human beings as well as of other living animals. We recognize the objects around us and we move and act in relation to them. Yet, when asked to define what is recognition, we come out with a vague answer. According to the nature of the patterns to be recognized, we may divide our recognition action into two major types : (1) the recognition of concrete items and (2) the recognition of abstract items. It is the first type of recognition capability which has been the objective of various recognition systems and studies over past two decades.

Over such a short period of time, there has been a phenomenal growth of interest and activities in the field of pattern recognition encompassing a wide variety of disciplines. This has led to the application of pattern recognition techniques in such diverse tasks as land and terrain classification, weather prediction, analysis of seismic data, medical diagnosis, finger-print classification, system identification, speech recognition, target detection, character recognition etc. Because of such diversity, most of the research workers have tried to avoid any formal definition for pattern recognition. However, from engineering point of view, pattern recognition can be considered as the categorization of input data into

identifiable classes via extraction of significant characteristics of the data from a background of irrelevant details[116].

1.2. Aim of the Present Work :

The work reported here deals with the problem of character recognition, which has been of interest to many research workers in the field of pattern recognition. The main reasons for the interest have been the large possible commercial applications and ease of availability of raw data for testing and implementing various propositions in the general framework of pattern recognition. In the present work also, this has been one of the motivating factors for undertaking the problem of handprinted Devanāgari script* recognition. Moderately complex nature of the Devanāgari script has been another motivating factor posing a sufficient challenge and demanding a good amount of ingenuity on the part of a pattern recognition worker. Unlike simple juxtaposition in Roman script, a word in Devanāgari script is composed of composite characters joined by a horizontal line at the top. A composite character is made up of legal attachment of a vowel sign, called mātrā, and/or diacritical marks on a character, or a combination of half letters with a character; or both. This results in an extremely large number of combinations and complexity in the script, therefore asking for the use of innovative ideas to tackle this problem.

---

* Devanāgari is the script for Hindi, the official language of India. For familiarization with the script, see Appendix A.

The different existing techniques used to solve the pattern recognition problem can be grouped into two broad approaches : decision theoretic approach and structural (linguistic) approach. In the decision theoretic approach, a set of characteristic measurements (features) is extracted from the pattern and each pattern is viewed as a point or vector in the measurement space. The recognition of each pattern is usually made by partitioning the measurement space into a number of disjoint regions, each associated with a pattern class label and by determining the region to which the given unknown pattern belongs. A number of partitioning procedures making use of either statistical decision theory or information theory or decision making mechanism based on exposure to training sequence of labelled patterns have been developed and quite satisfactory performances have been obtained[7,24,40,67,81].

However, for a number of pattern recognition problems like analysis of bubble chamber tracks, scene analysis etc., the above approach has been found inadequate and use of structural approach has been argued[65,80,85,86,87]. The approach draws an analogy between the structure of patterns and the syntax of languages. Pattern primitives are first selected and their relationships in the pattern are described by a set of syntactic rules. The recognition is then accomplished by performing a syntax analysis (parsing) on the sentence describing the given pattern. But by and large, a great deal of pattern recognition schemes under the category of structural approach do not rely on the parsing of the input pattern using

the syntactic rules[48,88,110,117]. Quite often, the relationships between the different pattern primitives are coded into a numeric string form and then string matching or template matching as it is called, is employed to categorise the input data. This is more so in the case of character recognition techniques making use of structural properties of the different character classes. Some times a decision tree is employed which successively tests for different pattern primitives and their locations in the pattern. The design of decision tree or choice of templates plays an important role in such situations and they must be optimally selected in order to get the best performance out of the recognition system. However, not much thought seems to have been given to the choice of templates or decision trees in such recognition schemes. Most of the time, the choice is governed by the designer's knowledge about the problem at hand and his intuition. A part of the present work deals with this aspect of the problem, i.e., how to obtain a set of effective templates or efficient categorisation tree for pattern classification when a limited set of labelled patterns are available for such design.

1.3. Proposed Method of Attack :

As pointed out above, the recognition of Devanāgari script appears to be a formidable task when one considers the extremely large number of variations and complexities arising in the script due to the concatenation rules for forming words. However, some solace can be derived from the fact that there are certain subpatterns which occur recurringly in the Devanāgari

script and these can be identified or separated out without much processing. Therefore, if one is ready to base the recognition scheme on the structure of Devanāgari Alphabet and their juxtaposition rules for forming words, a viable recognition technique can be developed. But also in such an approach, some liberties must be taken with the Devanāgari script so as to keep any recognition scheme within manageable complexities. Keeping this in view, the proposed recognition scheme imposes certain limitations on the method of writing Devanāgari script so that it can be acceptable for classification. The main limitation is regarding the use of composite characters formed by a half letter with another character. For example : instead of writing चट्टान (chattān), the scheme requires it to be written as चटटान . It is the author's contention that the amount of tedious processing required for the decomposition of such composite characters and for the subsequent successful recognition of half letters is too heavy a price to pay for the retention of such style of writing and some sort of compromise must be made.

The recognition scheme for Devanāgari script basically consists of four phases besides the usual phase of preprocessing. These are : 1) extraction of main body, upper and lower mātrā regions from a given word pattern, 2) segmentation of the main body of the word pattern into individual letter patterns, 3) recognition of individual letters; and 4) recognition of vowel signs or mātrās. The heart of the whole scheme lies in the recognition of letters. A multistage decision process has

been developed for this, where each stage of the decision process eliminates a set of character labels for further consideration. A set of extremely simple primitives is used and all the characters are looked upon as concatenation of these primitives.

Regarding the other aspect of the work dealing with the choice of templates and efficient categorization trees, a particular class of pattern recognition problems is considered where the features or measurements are discrete in nature. Such types of features are very common in a number of character recognition techniques and other pattern recognition problems like medical diagnosis, analysis of data in social and behavioral sciences etc. Specifically, the problem considered can be posed as follows : Let a set of discrete, multivalued d-dimensional measurement vectors with labels from N classes be given as

$$X_1^i , X_2^i , \ldots, X_{m_i}^i , \quad i = 1,2,\ldots, N$$

where each

$$X_k^i = (x_{k_1}^i, x_{k_2}^i, \ldots, x_{k_d}^i).$$

The cost of each measurement is assumed to be known. The objective is to design an efficient decision tree so as to classify all these samples correctly.

A two phase solution to this problem is proposed where in the first phase a set of suitable templates is obtained. These are then viewed as the entries of a decision table which

is converted to an optimal decision tree in the second phase of the design. For the generation of templates, use of Stoffel's prime event theory[113] is proposed by developing a new algorithm which makes use of learning. In the learning algorithm, the given set of labelled patterns are treated as vertices in the d-dimensional measurement space. An implicant $W = [e_1, e_2, \ldots, e_d]$ is defined to represent a set of vertices in such measurement space with each $e_i \in Q_i$ and Q_i being the set of possible values which the ith measurement can take. In general, each e_i is defined to be a m-tuple so that the implicant W represents a set of vertices. For example : $W = [e_1, e_2, e_3] = [1, (0,2), 2]$ represents two vertices (patterns), $\{1,0,2\}$ and $\{1,2,2\}$ in the three dimensional measurement space.

Given two implicants $W_r = [e_{r1}, e_{r2}, \ldots, e_{rd}]$ and $W_s = [e_{s1}, e_{s2}, \ldots, e_{sd}]$, an inclusion relation is defined such that $W_r \subseteq W_s$ if every vertex $v \in W_r$ is also included in W_s. Thus $W_r \subseteq W_s$ iff $e_{ri} \subseteq e_{si}$ for all i. This inclusion property is found very significant as it partitions the set of vertices in the measurement space into subsets of included and non-included vertices and is used for obtaining a set of templates from the given set of samples.

Once the entries of the decision table, i.e., templates for different pattern classes are available, the design of efficient decision tree corresponding to it will involve a large amount of search over all possible trees. To cut down the search, a criterion has been developed whereby an estimate for

the minimum expected cost of a decision tree can be obtained. This criterion is then used to establish a search direction for obtaining an efficient decision tree. Use of this approach has been made to design a decision tree for the recognition of handprinted Devanāgari numerals. The decision tree so obtained is found to be more efficient as it requires a maximum of four tests compared to a maximum of eight tests required in an earlier proposed decision tree by the author [104] which was developed on the ad hoc basis.

1.4. Organisation of the Thesis :

The work reported in this dissertation has been organised into five chapters besides the present introductory chapter. In the second chapter, a brief review of the character recognition schemes developed over the past two decades is presented. The main emphasis in the review is on the feature extraction methods since the material on the classification is currently available in a number of books on pattern recognition [26,37,38,77,124]. In the third chapter, we describe a recognition system for handprinted Devanāgari script. The various characteristics of the Devanāgari script are brought out first to have a feel of the problems while trying to design a recognition system for Devanāgari script.

In the fourth chapter, the problem of designing efficient decision trees has been considered. First a need is established for the systematic design of the templates and decision trees which is later followed by a short review of the existing work

in multistage decision making with special reference to binary decision trees. Some of the existing methods for template selection are also pointed out. The later sections of the chapter discuss a learning classifier scheme for template selection and a search strategy for finding the efficient decision tree. A few illustrative examples and an application to the handprinted Devanāgeri numerals recognition are also discussed.

The fifth chapter presents the simulation details of the proposed Devanāgari script recognition system. Functions of different subroutines used in the simulation are discussed along with some flowcharts and suitable examples. Also presented are the method of data gathering and results of different recognition experiments carried out by using simulated Devanāgari script recognition system.

In the last chapter, a few remarks are made about the work presented in this dissertation and some suggestions are given for further work in this direction. An appendix on introduction to Devanāgari script has also been included for the benefit of those who are not familiar with this script.

1.5. Concluding Remark :

In this chapter, we have tried to outline the aims and objectives of this dissertation, which gives a new generalised approach to the learning classifier scheme for template generation and the design of efficient decision trees for a class of pattern recognition problems involving discrete variables.

Moreover, the design algorithm is equally applicable to the problem of converting decision tables into optimal decision trees. The classification technique used for the recognition of Devanāgari script, although specific in aim, underlines certain ideas which are also equally applicable to other pattern recognition tasks.