Def Sci J, Vol 35, No 3, July 1985, pp 327-351

Trends in Pattern Recognition and Machine Learning

D. Dutta Majumdar
Indian Statistical Institute, Calcutta

Abstract. This paper is tutorial in nature introducing the statistical and syntactic pattern recognition technique. The problem of pattern recognition has special reference with image analysis and some aspects of modern methods and application of the area of shape analysis and detection of objects included.

1. Introduction

During the past two and a half decades, there has been a considerable growth of interest in problems of pattern recognition. This is because a tremendous amount of non-numerical information is being generated as a result of continued interaction between science and technology on the one hand and the emerging man-machine civilization of the modern human society on the other. This demand has created an increasing need of methods and techniques for use in the design of pattern recognition systems. Many different approaches have been proposed for diverse types of applications and a large number of research papers including some books have been published.

In this paper, an attempt has been made to present the problems of pattern recognition (and image analysis which is the largest sub-group of pattern recognition) and modern trends in its methods and applications. Trends of research in frontier areas of this subject, in the author's view are being directed towards some major objectives some of which are application dependent as is enunciated in the following sections. The other objective is from futuristic developments of fifth generation computer communication technology and related knowledges based architecure (Fig. 1) which is beyond the scope of this paper. Though one may look at pattern recognition as signal processing in two and more dimensions and signal processing as pattern recognition in one dimension such a fusion is yet to take place in practice. So an attempt has been made to organise presentation in a fashion covering preliminaries of the methods and applications and then indicate the recent trends and their implications in a nutshell as follows : (1) PR and ML application areas and conceptualisation of the methodologies involved towards a generalised model (Fig. 2); (2) Mathematical and statistical approaches so far experimented with e.g. (a) statistical decision theory, (b) syntactic approach (Fig. 3), and fuzzy set theoretic approach and their recent

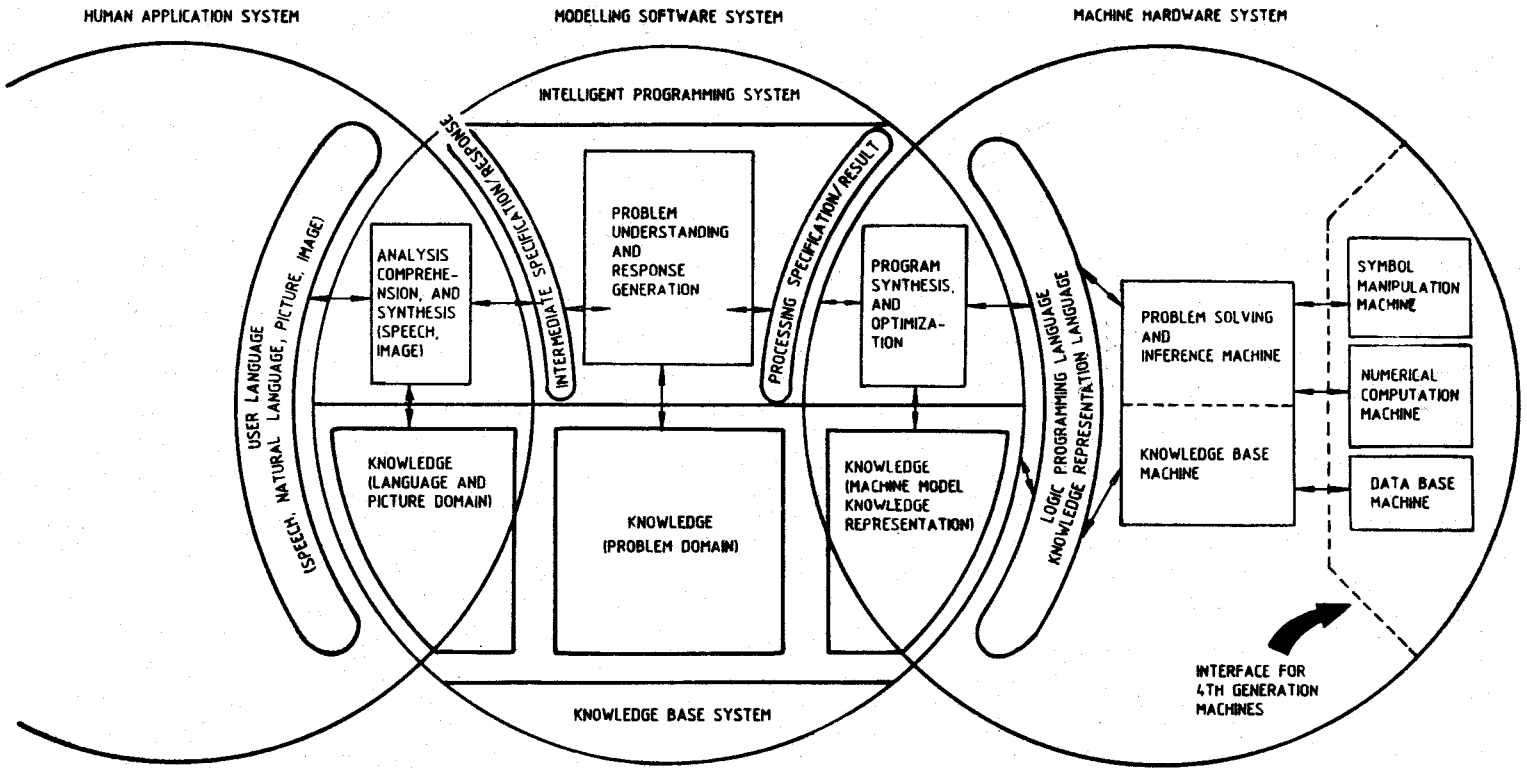

Figure 1. Conceptual diagram of the fifth generation computer system.

trends (Fig. 4); (3) Pre-processing and feature selection problem; (4) Problems of shape analysis; (5) Stochastic approach in PR and machine learning; and (6) Some applications of PR techniques in undersea search.

2. Syntactic Approach

A syntactic pattern recognition system can be considered as consisting of two major parts—analysis and recognition. The analysis part consists of primitive

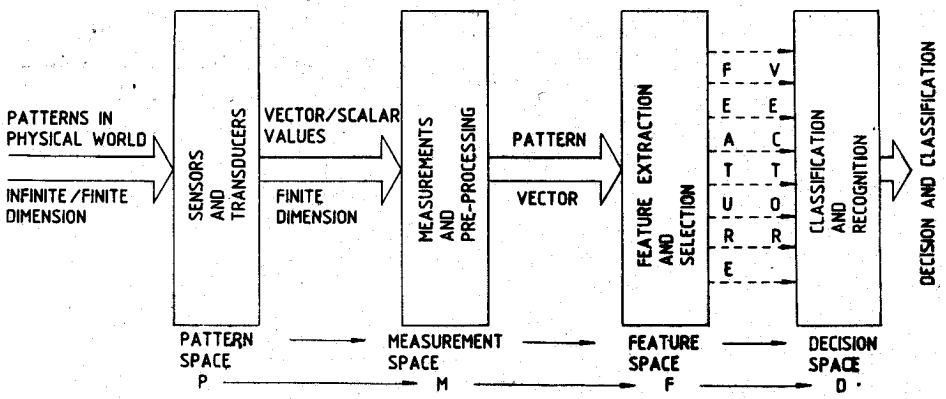

Figure 2. Conceptualization of a pattern recognition system in block diagram.

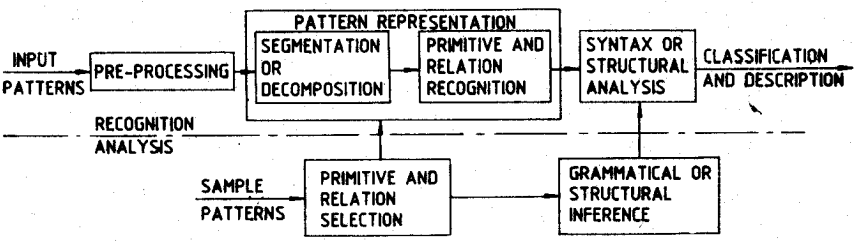

Figure 3. Block diagram of a syntactic pattern recognition system.

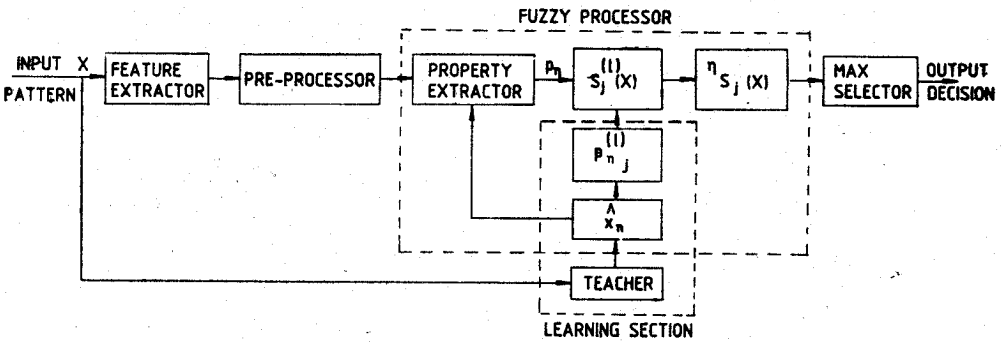

Figure 4. Block diagram of a fuzzy classifier.

selection and grammatical or structural inference and the recognition part consists of pre-processing, segmentation or decomposition, primitive and relation recognition and syntax analysis. A simple block diagram of such a system is shown in Fig. 3.

The pre-processing stage is very important in any pattern recognition system and is common to statistical, syntactic and fuzzy approach. To make the processing in the later stages more efficient, some sort of data compression should also be attempted at this stage. Each pre-processed pattern is then represented by a language-like structure e.g. string. This pattern representation process consists of pattern segmentation or decomposition and primitive, and relation recognition.

To represent a pattern in terms of its sub-patterns, we must segment or decompose the pattern and, in the meantime, identify or recognize the primitives and the relations in it. In other words, each pre-processed pattern is segmented or decomposed into sub-patterns and pattern primitives based on pre-specified syntactic operations. In turn, each sub-pattern is identified with a given set of pattern primitives. Each pattern at this point is represented by a set of primitives with specified syntactic operations, and in more recent sophisticated systems should also be able to detect the syntactic relations within the pattern. For example, in terms of concatenation operation, each pattern is represented by a string of concatenated primitives. The decision as to whether or not the representation (pattern) is syntactically correct (i.e., belongs to the class of patterns described by the given syntax or grammar) will be made by the syntax analyser or parser. When performing the syntax analysis or parsing, the analyser can usually produce a complete syntactic description, in terms of a parse or parsing tree, of the pattern, provided that the latter is syntactically correct. Otherwise, the pattern is either rejected or analysed on the basis of other given grammars which presumably describe other possible classes of patterns under consideration.

In order to have a grammar describing the structural information about the class of patterns under study, a grammatical inference machine is required that can infer a grammar from a given set of training patterns in language-like representation. Current trend is to perform this task primarily by the designer or through an interactive procedure. This machine's function is analogous to the 'learning process' in the decision theoretic pattern recognition system. The structural description of the class of patterns under study is learned from the actual sample patterns from that class. The learned description, in the form of a grammar, is then used for pattern description and syntax analysis. A more general form of learning might include the capability of learning the best set of primitives and the corresponding structural description for the class of patterns concerned.

3. Training and Learning in Recognition System Design

3.1. *The Process of Learning*

The classifiers so far described can be implemented with complete a priori knowledge relevant to pattern classes, namely, weighting co-efficients in linear discrimi-

nant classifier, reference vectors in minimum distance classifier and $P(C_k)$ (*p.d.f.*) and $P(X/C_k)$ (conditional *p.d.f.*) in Bayes' classifier etc. In practice, an infinite number of samples of classes are not available. We have, instead, a finite and usually small number of samples so that information required for optimal design of feature extractor or classifiers is often partially known. Under such circumstances, we must assume at best, that these samples are representative of those that would be obtained by examining a much larger sample size. If this requirement is satisfied—and we must usualy satisfy this requirement through engineering judgement exercised in the selection of samples—the classifier can be designed to have the capability of learning the best values of the statistical information from the training patterns to result in nearly the minimum number of misclassification.

Learning is a task of constructing the regions or templates in the N-dimensional space in which labelled samples of the classes are contained. By observing the patterns with known classification, a linear discriminant classifier, for example, can automatically adjust the weighting co-efficients associated with its discriminant function. The performance of the classifier is supposed to improve as the number of training patterns is increased. Under the assumption that the patterns from different classes are linearly separable, it is also possible to develop several algorithms referred to as 'error-correction' training procedure[1-4] to find the linear hyperplanes which properley separate the data and to have the property of converging to the solution which linearly separates the prototypes into their correct classifications if indeed the data is so separable.

In the statistical classification approach, if the unknown information is the parameter values of a known distribution function $P(X/C_k)$, the parametric learning technique can be applied. Suppose $\phi, \phi_2, \ldots\phi_j, \ldots\phi_m$ are the m training subsets of patterns corresponding to the m classes. Then with the knowledge that $P(X/C_j)$ are normal, the estimates of the parameters S_j and $\bar{X}_j$ are defined by the following sample statistics

$$\langle X \rangle_j = E(X) = \frac{1}{M_j} \sum_{X \in \phi_j} X \tag{1}$$

$$\langle S \rangle_j = \frac{1}{M_j} \sum_{X \in \phi_j} (X - \langle X \rangle_j)\ (X - \langle X \rangle_j)^T \tag{2}$$

where $E(X)$ denotes expected value of X, M_j is the number of patterns in the training subset ϕ_j and $\langle X \rangle_j$ and $\langle S \rangle_j$ denote the sample mean (or centre of gravity) and sample covariance matrix respectively of the jth class. If, on the other hand, both the function and parameter values are unknown, the non-parametric techniques in general should be used.

The two stages of pattern recognition—deriving the decision rule (learning) and using it to recognise a pattern—can be performed in two ways : (a) learning before recognition, and (b) learning and recognition concurrently. In the first method, all

the labelled pattern samples are collected and the best decision rule based on those samples is derived. The fixed decision rule is then applied without change to classify unlabelled patterns. The decision rule in the latter method is adaptive and is updated according to output decision. If the learned information gradually approaches the true information, then the decisions based on the learned information will eventually approach the optimal decision as if all the information required is known. Therefore, during the system's operation, the performance of the system is gradually improved. Learning process can be termed as 'supervised' or 'non-supervised' depending on whether the correct classification of the input patterns observed is known or not. Learning methods based on Bayesian estimation techniques and stochastic approximation are discussed in the following sections.

3.2. *Bayesian Learning*

Since the estimates of parameters are random vectors, it will be shown in this section how the density function of the estimate can be calculated by a successive process. Both the supervised and unsupervised techniques using Bayes' theorem are treated separately.

3.2.1. *Supervised learning* : Supervised estimation schemes using Bayesian learnings are discussed here to successively estimate an unknown parameter ϕ in a known form of feature distributior of each class $P(X/C_j)$. Let us assume that on a priori density function for the unknown parameter ϕ be $P_0(\phi)$ which reflects the initial knowledge about ϕ (N-dimentional vector). Let X_1, X_2,. ...X_n be a sequence of independent identically distributed feature vectors observed from the same pattern class C_j. Then according to Bayes' theorem, the function $P_0(\phi)$ changes to the a posteriori density function $P(\phi/X_1, \ldots X_n)$. For example, the a posteriori density function of ϕ given the first observation X_1 is

$$P(\phi/X_1) = \frac{P(X_1/\phi)\, P_0(\phi)}{P(X_1)} \tag{3}$$

After X_1 and X_2 are observed, the a posteriori density function of ϕ is

$$P(\phi/X_1, X_2) = \frac{P(X_2/X_1, \phi)\, P(\phi/X_1)}{P(X_2/X_1)} \tag{4}$$

After nth observation is observed

$$P(\phi/X_1, \ldots X_n) = \frac{P(X_n/X_1 \ldots X_{n-1}, \phi)\, P(\phi/X_1, \ldots X_{n-1})}{P(X_n/X_1, \ldots X_{n-1})} \tag{5}$$

With the knowledge of $P(\phi/X_1, \ldots X_n)$ one can compute the required probability density function

$$P(X_{n+1}/X_1, \ldots X_n, C_j) = \int P(X_{n+1}/X_1, \ldots X_n, C_j, \phi)\, P(\phi/X_1, \ldots X_n, C_j)\, d\phi$$

$$n = 1, 2, \ldots \tag{6}$$

where $P(X_{n+1}/X_1, \ldots X_n, C_j, \phi)$ is known. The a posteriori density function on the average becomes more concentrated and converge to the true value of the parameter so long as the true value is not excluded by the a posteriori density function of the parameter[5] ϕ.

Consider for example, the case of learning the mean vector $\bar{X}$ with known covariance matrix S for a Gaussian distribution. Let $\bar{X}_0$ represent the initial estimate of the mean vector and S_0 be the initial covariance matrix which reflects the uncertaintity about $\bar{X}_0$ so that

$$P_0(\phi) = P_0(\bar{X}) = \frac{1}{(2\pi)^{N/2} \mid S_0 \mid^{1/2}} \exp\left[-\tfrac{1}{2}(\bar{X} - \bar{X}_0)^T S_0^{-1} (\bar{X} - \bar{X}_0)\right] \tag{7}$$

Then after successive application of Bayes' theorem, the a posteriori density function $P(\bar{X}/X_1, \ldots X_n)$ is again a Gaussian density function with $\bar{X}_0$ and S_0 replaced by the new estimates $\bar{X}_n$ and S_n, where

$$\begin{aligned} \bar{X}_n &= E\,[\bar{X}_{n+1}/X_1, \ldots X_n] \\ &= (n^{-1}S)\,(S_0 + n^{-1}S)^{-1}\,\bar{X}_0 + S_0(S_0 - n^{-1}\,S)^{-1}\,\langle X \rangle \end{aligned} \tag{8a}$$

$$\begin{aligned} S_n &= \mathrm{Cov}\,[\bar{X}_{n+1}/X_1, \ldots X_n] \\ &= (n^{-1}S)\,(S_0 + n^{-1}\,S)^{-1}\,S_0 \end{aligned} \tag{8b}$$

and

$$\langle X \rangle = \frac{1}{n} \sum_i X_i,\ i = 1, 2, \ldots n \tag{8c}$$

In a special case, where

$$S_0 = \alpha^{-1}\,S,\ \alpha > 0 \tag{9a}$$

$$\bar{X}_n = \frac{\alpha}{n + \alpha}\,\bar{X}_0 + \frac{n}{n + \alpha}\,\langle X \rangle \tag{9b}$$

and

$$S_n = \frac{1}{n + \alpha}\,S \tag{9c}$$

where $\bar{X}_n$ and S_n are respectively, the conditional mean and covariance of $\bar{X}$ after n learning observations. As $n \to \infty$, $\bar{X}_n \to \langle X \rangle$, an unbiased estimate of true mean vector $\bar{X}$ and $S_n \to 0$.

There are other cases such as estimating mean vector and covariance matrix and learning the covariance matrix with zero mean vector etc.[2-4,6] using Bayesian iterative process.

3.2.2. *Non-supervised learning* : In non-supervised learning, the learning observations (as their correct classifications are not known) are considered as coming from the mixture distribution with the probability distribution of each class as component distributions. The problem of learning is then reduced to a process of successive

estimation of some unknown parameters in either a mixture distribution of all possible pattern classes or of a known decision boundary.

The mixture distribution is defined as

$$P(X) = \sum_i P(X/Z_i^n)\, P(Z_i^n),\ i = 1, 2, \ldots W \tag{10}$$

where $P(X/Z_i^n)$ denotes the ith-partition conditional distribution, $P(Z_i^n)$ the mixing parameter for ith-partition Z_i^n and W ($= m^n$, m = number of class distribution) is the number of ways Z_1^n, Z_2^n, $\ldots Z_W^n$ in which the set of training observations X_1, X_2, $\ldots X_n$ can be partitioned constituting an overall mixture distribution.

If $P(X/\phi, P)$ represents the parameter conditional mixture distribution where $\phi = \{\phi_1, \phi_2, \ldots \phi_W\}$ and $P = \{PZ_1^n), P(Z_2^n), \ldots P(Z_W^n)\}$ are the two sets of parameters and $P(X/\phi_i, Z_i^n)$ the ith parameter conditional distribution, then in terms of the set of parameters the above equation becomes

$$P(X/\phi, P) = \sum_i P(X/\phi_i, Z_i^n), P(Z_i^n), i = 1, 2, \ldots W \tag{11}$$

The problem of non-supervised learning is, therefore, reduced to that of finding a unique solution for ϕ and P, given $P(X/\phi, P)$.

Let us now assume that there are two pattern classes C_1 and C_2 having the respective known form of the probability density functions $P(X/C_1)$ and $P(X/C_2)$ and the parameter ϕ of the mixture distribution is unknown. Then the a posteriori density to estimate the parameter φ is obtained by the Eqn. (12).

$$P(\phi/X_1, \ldots X_n) = \sum_i P(\phi/X_1, \ldots X_n, Z_i^n)\, P(Z_i^n/X_1, \ldots X_n) \tag{12}$$

$$i = 1, 2, \ldots W,\ W = 2^n$$

The problem is therefore reduced to that of supervised learning for each of the 2^n partitions.

To estimate the parameter of an optimum decision boundary of a two-class pattern recognition problem, consider that learning observations X_1, X_2, . X_n are drawn from one of the classes C_1 and C_2 having univariate Gaussian distribution with some unknown parameters. If the variance of each class is the same and $P(C_1) = P(C_2) = 1/2$, the optimum decision boundary in non-sequential Bayes' classification process is known to be the mean of the two means.

For supervised learning, the two means can easily be learned from the classified learning observations. In case of non-supervised learning, the problem is one of estimating the mean of the mixture distribution $P(X)$ where

$$P(X) = \sum_j P(C_j)\, P(X/C_j),\ j = 1, 2. \tag{13a}$$

$$= \tfrac{1}{2}\,[N(\bar{X}_1, \sigma^2) + N(\bar{X}_2, \sigma^2)] \tag{13b}$$

$$= \frac{1}{(2\pi)^{1/2}\sigma} \exp\,[-a^2/2\sigma^2] \exp\,[-(1/2\sigma^2) \times$$

$$\times (X - \bar{X})^2]\ \text{Cosh}\ [(a/\sigma^2)\,(X - \bar{X})] \tag{14a}$$

where

$$a = (\bar{X}_2 - \bar{X}_1)/2 \tag{14b}$$

$$\bar{x} = (\bar{X}_1 + \bar{X}_2)/2 \tag{14c}$$

and the optimum decision boundary is seen to be simply the mean $\bar{X}$ of the mixture distribution. The simplest estimate of $\bar{X}$ is the sample mean

$$\langle X \rangle = \frac{1}{n} \sum_i X_i,\ i = 1, 2, \ldots n \tag{15}$$

For the case of N-dimensional spherically symmetric multivariate Gaussian distributions which differ only in the mean vector location, the optimum decision boundary is the $(N-1)$-dimensional hyperplane which perpendicularly bisects the vector connecting the two means[2,3]. Several other approaches for learning of recognition systems are available in literature[1-10].

3.3. *Learning Using Stochastic Approximation*

Stochastic approximation is a recursive technique that has been developed as an optimization technique for random environments. This approximation can be used for successive estimation of an unknown parameter when due to the stochastic nature of the problem the measurements are having certain errors. The technique guarantees the convergence of the algorithm even when the observation vectors are not linearly separable. Details about stochastic approximation along with the several applications such as in communication theory, control theory and pattern recognition, are available in texts.[1-9,11-16] This section relates only to some of the learning methods in pattern recognition problems using stochastic approximation. Before that discussion, let us first consider a simple example which leads to a basic approach to successive estimation and then describe the Robbins-Monro algorithm[13] in which a simplest form of stochastic approximation is used in finding a root of a regression function.

Suppose we have n observation vectors $X_1, X_2, \ldots X_n$ and we want to make a successive estimate of the mean vector from these observations. The non-successive estimate $\bar{X}_n$ of the mean vector based on these observations is given by

$$\bar{X}_n = \frac{1}{n} \sum_i X_i, \qquad i = 1, 2, \ldots n \tag{16}$$

This can be written as

$$\bar{X}_n = \frac{1}{n} \sum_i X_i + \frac{1}{n} X_n, \quad i = 1, 2, ..(n-1) \tag{17a}$$

$$\bar{X}_n = \left(\frac{n-1}{n}\right)\left(\frac{1}{n-1}\right) \sum_i X_i + \frac{1}{n} X_n, i = 1, 2, \ldots (n-1) \tag{17b}$$

$$= \left(\frac{n-1}{n}\right) \bar{X}_{n-1} + \frac{1}{n} X_n = \bar{X}_{n-1} + \frac{1}{n} (X_n - \bar{X}_{n-1}) \tag{17c}$$

Therefore, if we store n and $\bar{X}_{n-1}$, the mean vector estimated from $(n-1)$ samples, we can compute $\bar{X}_n$ with a new incoming nth observation X_n using Eqn. (17c). It also shows that as n increases, the effect of the new sample X_n on the expected vector decreases as follows

$$X_1, \frac{1}{2} X_2, \frac{1}{3} X_3, \ldots \frac{1}{i} X_i, \ldots \frac{1}{n} X_n$$

The sequence $1, \frac{1}{2}, \frac{1}{3}, \ldots \frac{1}{i}, \ldots \frac{1}{n}$ is known as a 'harmonic sequence'. The above findings therefore suggest that if we have an expression for non-successive estimate of a parameter from n samples, the expression for its successive estimate may be obtained by separating the estimate in two parts, one of which corresponds to the estimate obtained from $(n-1)$ samples and the other is the contribution of the nth sample. The effect of the nth sample can also be made reduced by using a co-efficient which is a decreasing function of n.

3.3.1. *The Robbins-Monro Algorithm* : Let $f(\theta)$ be a function of $\hat{\theta}$ having a single root $\hat{\theta}$ so that

$$f(\hat{\theta}) = 0 \tag{18a}$$

$$f(\theta) > 0 \text{ for } \theta > \hat{\theta} \tag{18b}$$

$$f(\theta) < 0 \text{ for } \theta < \hat{\theta} \tag{18c}$$

The last two conditions are assumed with little loss of generality since most functions of a single root not satisfying these conditions can be made to do so by multiplying the function by -1. If $g(\theta)$ denotes the noisy value of the regression function $f(\theta)$ then $(f(\theta) - g(\theta))$ will denote the error between these true and noisy random observation at a particular value of θ. These are illustrated in Fig. 5.

Let us now assume that (a) the random observations $g(\theta)$ are unbiased and (b) variance of the random observation $g(\theta)$ from $f(\theta)$ is finite for all values of θ.

The first assumption does mean that if there are numerous observations $g(\theta)$ for a fixed θ, the expected value of these observations should approach $f(\theta)$ at that

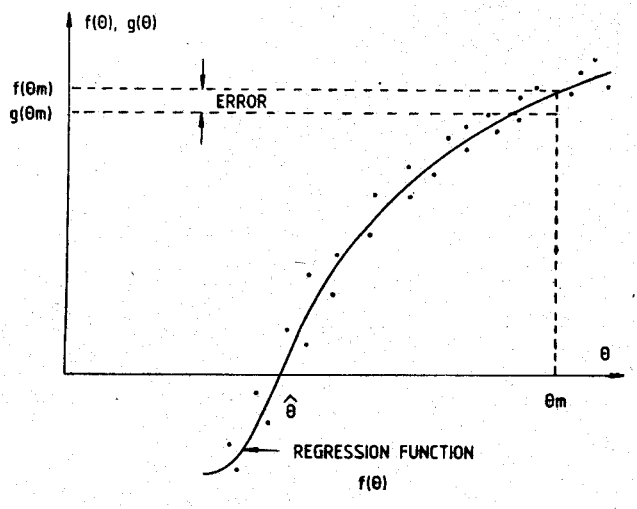

Figure 5. Finding a root of a regression curve representing a discriminant function.

point as the sample size increases. The second assumption on the other hand needs the noisy observation to be reasonably well behaved. In other words, it precludes noisy observations which are so far from $f(\theta)$ that the root-finding algorithm would never be able to recover. These two assumptions therefore imply that

$$E\{g(\theta)\} = f(\theta) \tag{19a}$$

$$\sigma^2(\theta) = E\{(f(\theta) - g(\theta))^2\} < A, \text{ for all } \theta \tag{19b}$$

where A is a finite positive constant.

Under these assumptions, the Robbins-Monro algorithm can be used to successively estimate the root θ of the function $f(\theta)$. The algorithm says that if θ_n represents the estimate of the root at nth iteration and $g(\theta_n)$ is the nth random observation, then the $(n + 1)$th estimate of θ is given by

$$\theta_{n+1} = \theta_n - \lambda_n g(\theta_n) \tag{20}$$

where λ_n is a member of a sequence of positive numbers satisfying the following conditions,

$$\lim_{n \to \infty} \lambda_n = 0 \tag{21a}$$

$$\sum_n \lambda_n = \infty, \; n = 1, 2, \ldots \infty \tag{21b}$$

$$\sum_n \lambda_n^2 < \infty, \; n = 1, 2, \ldots \infty \tag{21c}$$

These three conditions (21a, 21b, 21c) showed that the algorithm (20) will converge to $\hat{\theta}$ in the mean square sense[13], i.e.,

$$\underset{n\to\infty}{\text{Lt}} \quad E\{(\theta_n - \hat{\theta})^2\} = 0 \tag{22}$$

The Eqn. (22) says that as the number of iteration approaches infinity, the variance of θ_n $\hat{\theta}$ will approach zero, i.e., θ_n will approach $\hat{\theta}$.

The physical meaning and the role of the Eqns. (21) for the convergence of the algorithm can be described as follows. Eqn. (21a) allows the process to settle down in the limit. The Eqn. (21b) insures on the other hand that there is enough corrective action to avoid stopping short of the process. The finite variance of the accumulated noise is guaranteed by Eqn. (21c) so that we can correct the effect of noise. For instance, $\lambda_n = \frac{1}{n}$, a harmonic sequence and more generally $\lambda_n = 1/n^k$, $1 \geqslant k > \frac{1}{2}$ satisfies equations (21a) through (21c). It is to be mentioned here that these are not the only possible sequences.

The convergence of the algorithm (20) with probability 1, i.e.,

$$\text{Prob}\{\underset{n\to\infty}{\text{Lt}}\ \theta_n = \hat{\theta}\} = 1 \tag{23}$$

was established by Blum[14]. This stronger form of convergence indicates that, in the limit, it is guaranteed that θ_n will equal $\hat{\theta}$. It was Dvoretzky[15] who gave a generalised form of these proofs of Robbins-Monro and Blum and showed that the convergence criteria Eqns. (22) and (23) held for any stochastic approximation procedures satisfying the conditions of his theorem.

Consider Eqns. (20) and (21). Since any sequence λ_n satisfying Eqns. (21) must decrease with increasing n, the decreasing significance of the correction factor λ_n has the effect of decreasing the magnitude of the adjustments with successive iterations. As a result, the Robbins-Monro algorithm, as well as other similar stochastic approximation schemes, is found generally slow to converge. Of course, this is the price for guaranteed convergence. One way to accelarate this Robbins-Monro algorithm is to change λ_n to the next value only when a sign change of $g(\theta_n)$ is observed. Because change in sign of $g(\theta_n)$ tend to occur more often in the vicinity of the root $\hat{\theta}$. For the points away from $\hat{\theta}$, convergence speed is more important than guaranteed convergence and large corrections are desired. When a sign change of $g(\theta_n)$ is observed near $\hat{\theta}$, we have to start worrying about convergence and therefore the corrections should be smaller and smaller.

3.3.2. *Supervised learning* : Let X_1, X_2, $\ldots X_n \ldots$ be a sequence of classified training feature vectors from which $P(C_j)$, the unknown a priori probability of jth class, $j = 1, 2, \ldots m$ is to be estimated. If the initial estimate of $P(C_j)$ is $P_0(C_j)$ such that

$$0 \leqslant P_0(C_j) \leqslant 1 \tag{24a}$$

and

$$\sum_j P_0(C_j) = 1,\ j = 1, 2, \ldots m \tag{24b}$$

then the successive estimates of $P(C_j)$ can be obtained from the following stochastic approximation algorithm

$$P_{n+1}(C_j) = P_n(C_j) + \lambda_{n+1}\left[\frac{n_j}{n} - P_n(C_j)\right] \tag{25a}$$

where $$\sum_j n_j = n, \quad j = 1, 2, \ldots m \tag{25b}$$

n_j denotes the number of times the observations are from jth class, $P_n(C_j)$ represents the value of $P(C_j)$ estimated after nth iteration and $\{\lambda_n\}$ is a sequence of numbers satisfying the convergence conditions. Since $E(n_j) = n\,P(C_j)$, the conditions

$$E\left(\frac{n_j}{n}\right) = P(C_j) \tag{26a}$$

and $$E\left[\left(\frac{n_j}{n}\right) - P(C_j)\right]^2 < \infty \tag{26b}$$

will always exist. Therefore, the successive estimates $P_n(C_j)$ will approach the true value $P(C_j)$ in the mean square sense and with unity probability.

Let us now consider the case of learning an unknown probability density function $P(X)$ from the observations $X_1, X_2, \ldots X_n$. Let $P(X)$ be denoted by a finite series approximation in the form[2].

$$P(X) = \sum_j a_j \xi_j(X), \quad j = 1, 2, \ldots m \tag{27}$$

where $\xi_j(X)$ is a system of orthonormal functions such that

$$\int \xi_j(X)\, \xi_i(X)\, dX = 0, \; i \neq j$$

$$= 1, \; i = j \tag{28}$$

The parameters $\{a_j\}$ can be estimated from the stochastic approximation algorithm given below

$$a_{j(n+1)} = a_{j(n)} + \lambda_n[\xi_j(X_n) - a_{j(n)}] \tag{29}$$

$$j = 1, 2, \ldots m$$

where $a_{j(n+1)}$ denotes the estimated value of a_j after $(n+1)$th iteration and $\{\lambda_n\}$ satisfies the conditions (21). Therefore, as $n \to \infty$, $a_{j(n)} \to a_j$, in the mean square sence and with probability 1.

Consider for example, the case of learning the mean vector $\bar{X}$ of a Gaussian distribution. It is a special case of the above algorithm where the form of the density function is known but some parameters are unknown.

Now in terms of recursive relationships, (8a) and Eqns. (8b) can be written as

$$\bar{X}_n = S(S_{n-1} + S)^{-1}\, \bar{X}_{n-1} + S_{n-1}\,(S_{n-1} + S)^{-1}\, X_n \tag{30a}$$

and $$S_n = S(S_{n-1} + S)^{-1}\, S_{n-1} \tag{30b}$$

By adding and subtracting $\bar{X}_{n-1}$ at the right hand sides, Eqn. (30a) can be written as

$$\bar{X}_n = \bar{X}_{n-1} + S_{n-1}(S_{n-1} + S)^{-1}(X_n - \bar{X}_{n-1}) \tag{31}$$

In a special case, where Eqn. (9a) is satisfied i.e.,

$$S_0 = \alpha^{-1} S, \alpha > 0$$

$$\bar{X}_n = \bar{X}_{n-1} + (n + \alpha)^{-1}(X_n - \bar{X}_{n-1}) \tag{32}$$

Let us now assume that

(a) X_n is an N-dimensional noise vector satisfying the conditions of zero mean and finite variance for each component and for all n i.e.,

$$X_n = \bar{X} + \eta_n \tag{33a}$$

where $\bar{X}$ is the true mean vector to be learned and

$$\eta_n = (\eta_n^1, \eta_n^2, \ldots \eta_n^N) \tag{33b}$$

and (b) $$\lambda_n = (n + \alpha)^{-1} \tag{34}$$

Then using Eqns. (33a) and (34), Eqn. (32) can be written as

$$\bar{X}_n = \bar{X}_{n-1} + \lambda_n(\bar{X} + \eta_n - \bar{X}_{n-1}) \tag{35a}$$

$$= (1 - \lambda_n)\bar{X}_{n-1} + \lambda_n \bar{X} + \lambda_n \eta_n \tag{35b}$$

$$= F_n \bar{X}_{n-1} + \lambda_n \bar{X} + \lambda_n \eta_n \tag{35c}$$

$$= T_n(\bar{X}_1, \bar{X}_2, \ldots \bar{X}_{n-1}) + \lambda_n \eta_n \tag{35d}$$

where

$$F_n = 1 - \lambda_n = 1 - (n + \alpha)^{-1} \tag{36a}$$

and $$T_n(\bar{X}_1, \bar{X}_2, \ldots \bar{X}_{n-1}) = F_n \bar{X}_{n-1} + \lambda_n \bar{X} \tag{36b}$$

Therefore,

$$\| T_n(\bar{X}_1, \bar{X}_2, \ldots \bar{X}_{n-1}) - \bar{X} \| = \| F_n \bar{X}_{n-1} - (1 - \lambda_n)\bar{X} \|$$

$$= F_n \| \bar{X}_{n-1} - \bar{X} \| \tag{37}$$

and F_n satisfies the conditions,

$$F_n > 0 \text{ and } \prod_n F_n = 0, \; n = 1, 2, \ldots \infty$$

Again, since

$$\| \bar{X}_0 \| < \infty, \; \bar{X}_0 \text{ denotes the initial estimate of } \bar{X} \tag{38a}$$

and $$E[\|\eta_n\|^2] \leqslant B < \infty, \; B \text{ denotes some boundary} \tag{38b}$$

We may write

$$E[\|\bar{X}_0\|^2] + \sum_n E[\|\lambda_n \eta_n\|^2] \leqslant E[\|\bar{X}_0\|^2] + B \sum_n \lambda_n^2 < \infty \tag{39}$$

$$n = 1, 2, \ldots \infty$$

Now according to Dvoretzky's theorem for normal linear space[2,15], the condition (39) implies that

$$\underset{n\to\infty}{\text{Lt}}\ E[\|\bar{X}_n - \bar{X}\|^2] = 0 \tag{40a}$$

and

$$\text{Prob}\{\underset{n\to\infty}{\text{Lt}}\ \bar{X}_n = \bar{X}\} = 1 \tag{40b}$$

Therefore, Eqn. (32) represents an expression of a special case of stochastic approximation and the convergence of $\bar{X}_n$ to the true mean vector $\bar{X}$ in the mean square sense and with unity probability is implied by Dvoretzky's theorem.

3.3.3. *Non-supervised learning* : As mentioned in Section 3.2.2., the problem of non-supervised learning will be considered as the problem of estimating parameters in a mixture distribution. The unknown parameters are estimated by the stochastic approximation algorithm.

Consider the Eqn. (15) which can be rewritten as in Eqn. (17c)

$$\bar{X}_n = \bar{X}_{n-1} + n^{-1}(X_n - \bar{X}_{n-1}) \tag{41a}$$

$$= \bar{X}_{n-1} + \lambda_n(X_n - \bar{X}_{n-1}) \tag{41b}$$

with the initial estimate $\bar{X}_0 = 0$ and $\lambda_n = n^{-1}$ satisfying the condition (21). Let $\bar{X}$ be the true value of the mean to be estimated and

$$T_n(\bar{X}_1, \bar{X}_2, \ldots\bar{X}_{n-1}) = (1 - \lambda_n)\,\bar{X}_{n-1} + \lambda_n\bar{X} \tag{42}$$

Then

$$|\,T_n(\bar{X}_1, \bar{X}_2, \ldots\bar{X}_{n-1}) - \bar{X}\,| = |\,(1 - \lambda_n)\,\bar{X}_{n-1} + (1 - \lambda_n)\,\bar{X}\,| \tag{43a}$$

$$= (1 - \lambda_n)\,|\,\bar{X}_{n-1} - \bar{X}\,| \tag{43b}$$

$$= F_n\,|\,\bar{X}_{n-1} - \bar{X}\,| \tag{43c}$$

Since $F_n = (1 - \lambda_n) = (1 - n^{-1})$ satisfies the Dvoretzky's condition

$$\prod_n F_n = 0,\ n = 1, 2, \ldots\infty$$

it can therefore be seen that Eqn. (15) also represents a special algorithm of stochastic approximation with λ_n being a harmonic sequence. The estimates converge to the true mean in the mean square sense and with unity probability.

Let us now take the following assumptions[17,18] to estimate the unknown parameters of a mixture distribution using stochastic approximation.

(a) Learning samples (observations) are drawn from the mixture distribution defined by

$$P(X/\phi, P) = \sum_j P(X/\phi_j, C_j)\,P(C_j) \tag{44}$$

$$j = 1, 2, \ldots m$$

where $\phi = \{\phi_1, \phi_2, \ldots\phi_j, \ldots\phi_m\}$ and $P = \{P(C_1), P(C_2), \ldots P(C_j), \ldots P(C_m)\}$ are the sets of parameters characterising the mixture distribution function $P(X/\phi, P)$. The probability

density function of the class C_j is characterised by a set of parameters ϕ_j. The a priori probabilities $P(C_1)$, $P(C_2)$, ...$P(C_j)$, ...$P(C_m)$ corresponding to m classes are fixed but unknown and

$$\sum_j P(C_j) = 1, \quad j = 1, 2, ...m \tag{45}$$

(b) There exist unbiased estimates of certain statistics (e.g., first moment, second moment, etc.) for the mixture. The functional relationship between these statistics $H = \{H(X)\}$ and the parameter sets P and ϕ is known, i.e. an equation of the form

$$G_1(H, \phi, P) = 0 \tag{46}$$

is available at each stage of the learning process,

(c) An equation of the form

$$G_2(\phi, P) = 0 \tag{47}$$

relating ϕ and P is also known to result in a unique solution for the unknown parameters ϕ and P.

If these above conditions are satisfied with probability 1, then the true values of the parameters ϕ and P are defined in the limit by Eqns. (46) and (47). The process of learning is then reduced to that of finding the unique solution for ϕ and P through the functional relationships G_1 and G_2 where G_1 can be obtained from the successive estimates $\{H(X)\}$ and G_2 is given a priori or sought by some auxiliary estimation procedures. Several examples for estimation of ϕ and P using stochastic approximation are given[2].

3.3.4. *Non-linear stochastic approximation algorithm :* The relationship in Eqn. (32) or Eqn. (41b) defines the linear algorithm where only the first order difference $(X_n - \bar{X}_{n-1})$ is used in modifying the previous estimate and to arrive at the upto-date estimate $\bar{X}_n$. An extension of this algorithm to non-linear case had been defined by Chien[19] as

$$\bar{X}_n = \bar{X}_{n-1} + \gamma_n(D_n) \tag{48a}$$

where

$$D_n = X_n - \bar{X}_{n-1} \tag{48b}$$

$\gamma_n(D_n)$ is a non-linear function of D_n. This non-linear stochastic approximation algorithm is defined to handle spurious patterns like character samples with various imperfections due to the poor quality in writing or printing encountered in practice. To determine the functional form of λ_n so that the resulting $\bar{X}_n$ will give a minimum mean square error, the following assumption were made.

(a) Each learning sample is composed of an unknown mean vector under estimation and an independent Gaussian noise with zero mean and finite variance, i.e., condition (33a) is assumed.

(b) The noise component is considered as a mixture distribution consisted of two parts namely, the inherent measurement variation associated with reliable samples and

the spurious learning samples those carry unreliable information. Let S_1 and S_2 denote the convariance matrices of these noises with zero mean, and $(1 - r)$ and r the corresponding probability of occurrences such that

$$S_2 >> S_1$$

and $$r << 1$$

Under the above conditions, $\gamma_n(D_n)$ is found to be[19]

$$\gamma_n(D_n) = \frac{S_{n-1}\,[(1 - r)\,P_1(D_n)\,S_1^{-1} + rP_2(D_n)\,S_2^{-1}]}{(1 - r)\,P_1(D_n) + rP_2(D_n)}\,D_n \tag{49}$$

where $P(\cdot)$ denotes the probability density of the noise component discussed above. S_{n-1} represents the covariance matrix of the error vector $S_{n-1} = \bar{X}_{n-1} - \bar{X}$ at the $(n - 1)$th iteration. Now in order to explain the behaviour of $\gamma_n(D_n)$, let us define a parameter

$$d(D_n) = D_n^T\,S_1^{-1}\,D_n \tag{50}$$

which gives a generalised distance-measure between X_n and $\bar{X}_{n-1}$.

For genuine learning samples having reliable information for $\bar{X}$, $d(D_n)$ is small and $P_1(D_n) >> P_2(D_n)$. Thus

$$\gamma_n(D_n) \approx S_{n-1}\,S_1^{-1}\,D_n \tag{51a}$$

For spurious samples on the other hand, $d(D_n)$ is large such that $P_1(D_n) \to 0$ and $P_2(D_n) >> P_1(D_n)$. As a result,

$$\gamma_n(D_n) \approx S_{n-1}\,S_2^{-1}\,D_n \tag{51b}$$

Since $S_2 >> S_1$, the factor $(S_{n-1}\,S_2^{-1}\,D_n)$ will comparatively be insignificant for a fixed S_{n-1} and D_n. Therefore, one can select a threshold or a boundary B_n such that

$$\begin{aligned}\gamma_n(D_n) &= S_{n-1}\,S_1^{-1}\,D_n \text{ for } d(D_n) \leqslant B_n \\ &= 0 \qquad\qquad\qquad\quad > B_n\end{aligned} \tag{52}$$

This threshold effect thus gives a good approximation which implies that one can simply discard the learning samples those are found unreliable (large D_n) and at the same time carry out a linear transformation on the remaining reliable samples having small D_n values. The non-linear algorithm (48a) can, therefore, be written as the following quasi-linear form

$$\begin{aligned}\bar{X}_n &= \bar{X}_{n-1} + S_{n-1}\,S_1^{-1}\,D_n \text{ for } d(D_n)\ B_n \\ &= \bar{X}_{n-1} \qquad\qquad\qquad > B_n\end{aligned} \tag{53}$$

A comparison of both the linear Eqn. (32) and non-linear Eqn. (53) algorithms as implemented on a set of unreliable patterns of English characters along with the algorithms for selection of B_n and S_{n-1} is reported[19].

In order to discard the unreliable samples used in updating the parameters of a pattern recognition system, a self-supervised learning algorithm had also been developed[10,20]. Since for all practical problems, the distribution of the members of a class in the feature space has a central tendency, the algorithm assumed that the probability of misclassification near these central tendencies is substantially low. Different 'guard zones' were then defined in these regions to play the role of a supervisor whose purpose is to postpone the updating programme for doubtful unreliable samples (lying outside the guard zones). As a result, the minimisation of the effect of wrong classification on the initial parameters of the algorithm is ensured. As the system used some inherent properties of the distribution of the same parameters as used by the classifier itself, it is called a 'self-supervisory system'. The details of the algorithm with selection of an optimum guard zone is described as an adaptive recognition of vowel sounds and can be applied in other forms of pattern.

4. The Problem of Shape Analysis

The problem of shape analysis, discrimination, recognition, and description is a central one in the field of pattern recognition and image analysis. The recent spurt of activities in this field is partly because of its application in computer vision, robotics, and other diverse fields including military requirements, apart from simple applications such as character recognition. The problem of shape analysis, like other PR problems is a two-stage process. Firstly, the information reduction process in which a compact set of invariant feature of 'shape descriptors' of the object are extracted from the binary or grey level image of the object. Secondly, these are used for discrimination, classification, recognition and description by computers employing statistical, syntactic, fuzzy and other mathematical techniques. The methodologies employed can be classified according to many criteria as shown in Fig. 6. Special mention should be made of the review and the book by Pavlidis on Shape Analysis, however these are also seven years old.

The two major approaches for developing algorithms for shape analysis are (a) boundary based techniques and (b) region based techniques. Boundary based techniques are based on the philosophy that if the boundary of an object contains enough information about its shape then it is natural to retain only the boundary for further analysis of the object. In certain cases of this technique we need to extract local features, in certain cases we may need global features, and in some other cases we may need both. One of the earliest shape descriptors obtained from the boundary of an object are the elements of Freeman's chain code where the pattern is represented as a string of octal digits. Because these are dependent on size and starting point on the boundary, they cannot be used for shape discrimination. But useful information such as area, perimeter, curvature, convexity and concavity tree can be obtained from the

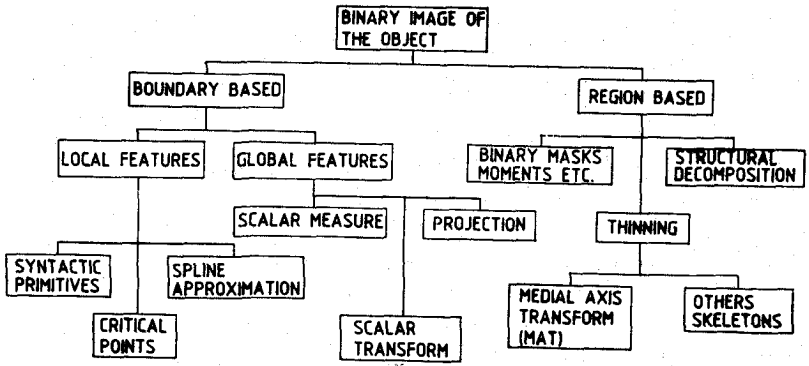

Figure 6. Classification of shape analysis techniques.

chain code elements. Extensive studies have been made by several authors such as Knoke and Wiley, Ladely and Fu.

One important technique that needs to be mentioned is the method of 'Critical Points', which according to classical geometry are maxima, minima, and points of inflection. The importance of curvature maxima in shape perception have been proved by psychological studies and a number of algorithms have been developed on these. Obviously, local features are sensitive to noise, dirt etc., and global features like elongation, area etc. are better suited in a noisy environment. The distinction between local and global feature extraction techniques are somewhat like that of a context free and context sensitive grammars in syntactic approach.

Global feature extraction methods are based mainly on different orthogonal transforms of the boundary of a two-dimensional object into a one-dimensional periodic function. In one approach by Granland and Roskies, the boundary is continuously scanned from a starting point on the boundary and a periodic function is obtained with period equal to the perimeter. In the second approach by Searle and Schwartz *et al.*, the length of the ratio chords from the centre of gravity of the object to the boundary at discrete intervals of angular inclination with respect to a reference vector are used as the input sequence to the orthogonal transform to extract shape descriptors. These approaches are extensively used to obtain Fourier Descriptors (FD's) of different shapes as explained by Person and Fu. The main disadvantage is that FD's require normalization to make them invariant to position, orientation and size of the object. The other technique is that of $R(\theta)$ transform. The function $R(\theta)$ is expanded using Fourier Series expansion or Walsh Series expansion and the resultant coefficients are used as shape discriminants. A disadvantage in using Walsh functions was the variance of both amplitude and phase angle spectra to shift in the position of the reference vector or the change in orientation of the object shape. Two recent methods which try to reduce this dependence are by Sethi & Sarvarayudu and Dinstein & Silberberg of which the former one is information preserving and can be used to reconstruct the shape.

We have already mentioned about the Region Based Techniques, consisting of Mask Matching, Structural Decomposition and Thinning Techniques. In shape recognition problem, a great deal of information lies in the locus of medial lines of strokes. One of the widely studied techniques to obtain the skeleton of a shape is the symmetrical axis transform better known as Medial Axis Transform (MAT). Most of the research work on shape analysis by our group at the Indian Statistical Institute, Calcutta is based on region based techniqnes. Two important contributions are (1) Theoretical development of a new definition of shape and (2) Fuzzy recognition and description of sides and symmetrics. A list of relevant references mentioned in the above discussion is given at the end[21-31].

5. Detection of Man-Made Objects Using Local Shape Descriptors—Application of PR Techniques in Automatic Undersea Search

Automatic undersea search operations for man-made objects on the ocean bed is an extremely important and sensitive subject. Though for a successful operation it is expected that multiple sensors such as side scan sonar, scanning sonar, magnetometers, television and underwater photographs will be needed but so far the reported work is limited to optical photographs only and the methodology involved is automatic shape analysis in distinguishing between man-made object and empty ocean floor from the photographs.

It is well known that in the past undersea search operations have been carried out by using men to recognise the target being sought. The Naval Research Laboratory of Washington has some reports of some search operations including the search of the Thresher and Scorpion submarines as well as the hydrogen bomb, which was lost off Spain. It was reported that the data from these operations were often hard to interpret and subject to human errors. But modern techniques are being developed on automating such operations. High resolution sensor is useful in obtaining detailed shape. However, its range is greatly limited if the necessary high frequency is used to obtain adequate details. Recognition of man-made objects by sonar is particularly difficult in areas which contain a large number of boulders and rocks. Magnetometers are useful in locating steel sheets. However, their detection ranges are relatively high with existing equipment. Besides, false alerts are frequent in areas where the bottom contains some natural object that exhibits magnetic properties. Though TV underwater photographs are good for such inspection purposes, but its range is limited to approximately 35 feet due to back-scatter problems. Sides-scan-sonar records are reported to be very difficult to interpret and analyses of photographs is also subject to human error. It is reported that the photographs which displayed the Thresher submarines draft markings were initially observed and set aside as of no value. It was only on later reflection that an operator discovered the draft markings.

Optical sensors are reported to be successful in numerous search missions. On the Thresher and Scorpion search operation film cameras were the primary sensor used to

locate the submarines. The major limitation of conventional cameras is that their range is limited to ten meters from the bottom. As a result, the area coverage is relatively small per exposure. The major limit factor to conventional camera system also in sea water is the back-scatter, and the maximum distance is of the order of ten meters only. The first method used to reduce back-scatter was to move the camera laterally away from the light source. A second technique puts the light beneath the camera. A third technique known as light behind the camera (Libec) system (Fig. 7) was developed by Naval Research Laboratory, Washington. In this case the light source is above the camera and the coverage per frame increases by a factor of four over a conventional camera. A system was designed to achieve underwater photographs at a distance of 70 meters with a resolution of 2.4 milli radium and field view of 64°. The fourth technique is known as Range Gating and is also reputed to be very successful in avoiding back-scatter problem.

Besides the back-scatter problem, there is a fundamental resolution limit which results from forward scattering from particle and which creates blurring of the image.

With the help of different shape descriptors a decision can be made on whether a given image contains man-made objects. The assumption behind this method is that man-made objects tend to have longer and smoother edges than natural objects, and that the orientation of their edges tends to concentrate in a few directions. So the edge

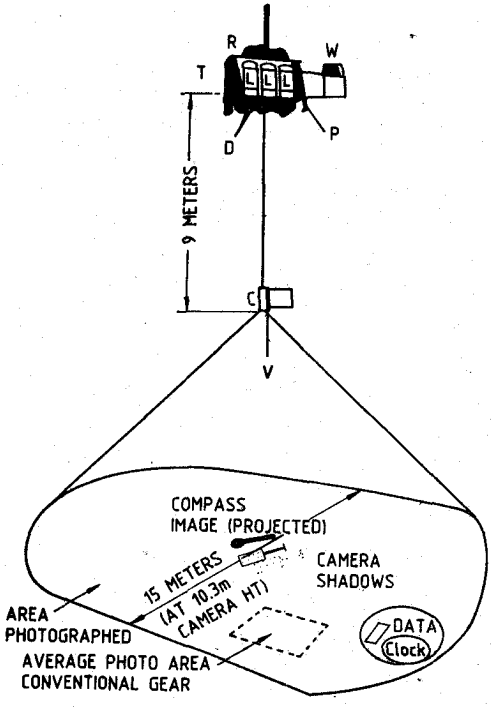

Figure 7. Geometry of the LIBEC system.

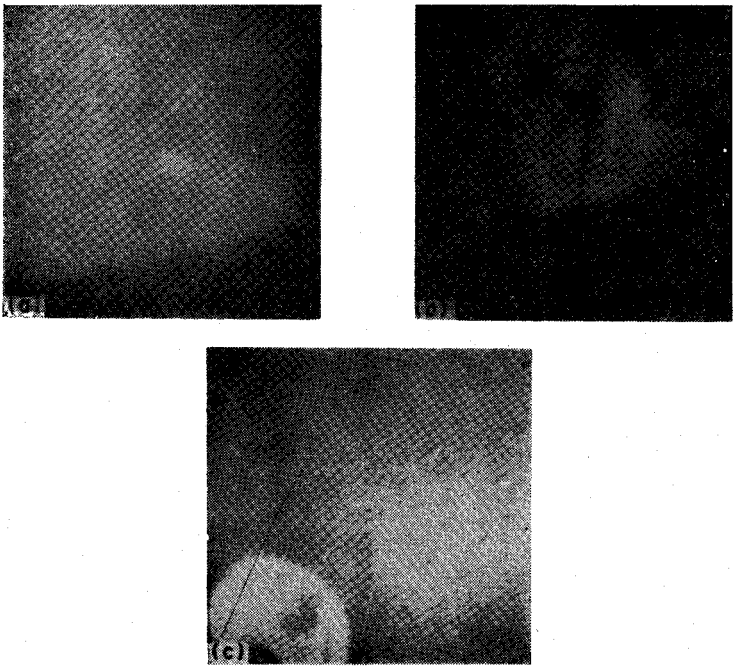

Figure 8. Optical underwater photographs :
(a) Ship on ocean floor (taken at 9 m),
(b) Ship on ocean floor (taken at 14 m),
(c) Empty ocean floor.

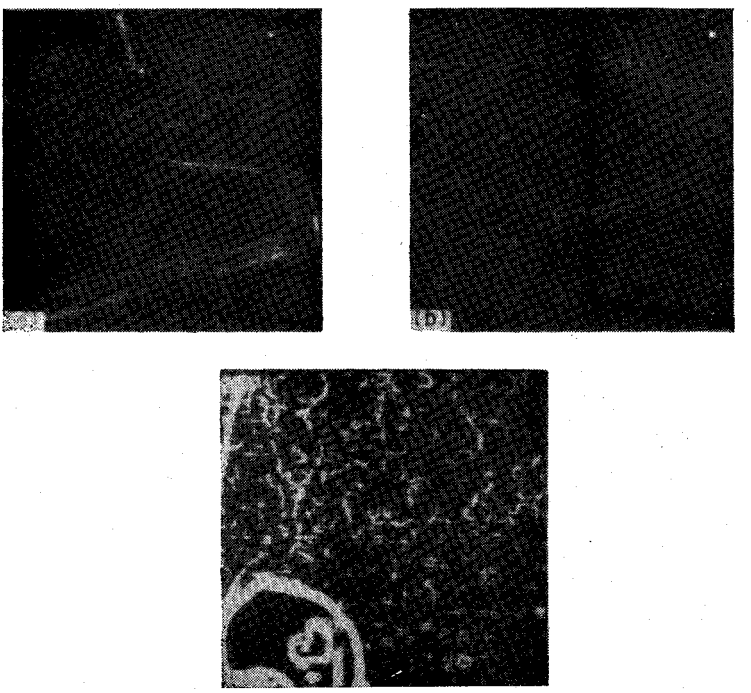

Figure 9. Results of edge detection.

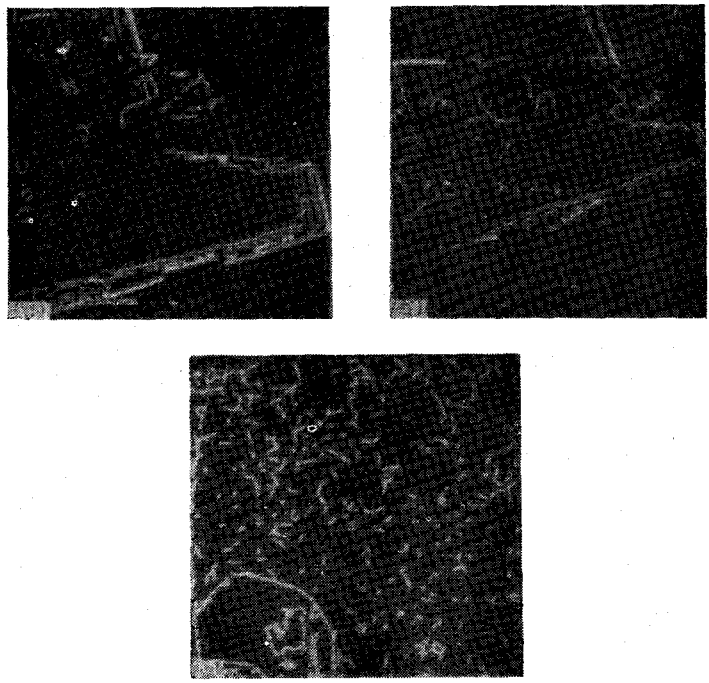

Figure 10. Results of piecewise linear approximation.

extraction problem through (a) edge detection, (b) thinning, (c) linking, and (d) some sort of piecewise linear approximation of the edges, shape descriptor such as average length per edge limit, average angle change per unit length and edge orientations histogram, can be calculated. With the help of this local shape descriptor, some texture features and global shape features, one can distinguish between man-made objects and ocean floor. However, there are many different types of ocean floors. Some of them contain sea plants or large rocks that contain edges which are also relatively longer and smoother and tend to orient in a few major directions. For such cases some other features than local shape are needed.

Fig. 8 shows optical underwater photographs with (a) ship on the ocean floor (taken at 9m), (b) ship on the ocean floor (taken at 14m) and (c) empty ocean floor. Fig. 9 shows the results of edge detection algorithm and Fig. 10 shows the results of piecewise liner approximation algorithm for linking.

References

1. Fukunaga, K., 'Introduction to Statistical Pattern Recognition' (Academic Press, N.Y.), 1972.
2. Fu, K. S., 'Sequential Methods in Pattern Recognition and Machine Learning' (Academic Press, London), 1968.
3. Andrews, H.C., 'Mathematical Techniques in Pattern Recognition' (Wiley Interscience, N.Y.), 1972.

4. Nilson, N. J., 'Learning Machines—Foundation of Trainable Pattern Classifying Systems' (McGraw-Hill, N.Y.), 1965.
5. Anderson, T. W., 'An Introduction to Multivariate Statistical Analysis' (Wiley, N. Y.), 1958.
6. Mendel, J. M. & Fu, K. S., 'Adaptive, Learning and Pattern Recognition Systems—Theory and Applications' (Academic Press), 1917.
7. Tou, J. T. & Gonzalez, R. C., 'Pattern Recognition Principles' (Addison-Wesley Publishing Company Inc., London), 1974.
8. Fu, K. S., 'Pattern Recognition and Machine Learning' (Planum Press, N. Y.), 1971.
9. Young, T. W. & Calvert, T. W., 'Classification, Estimation and Pattern Recognition' (American Elsevier Publishing Co. Inc., N. Y.), 1974.
10. Pal, S. K., 'Studies on the Application of Fuzzy Set Theoretic Approach in Some Problems of Pattern Recognition and Man-Machine Communication by Voice', Ph.D. Thesis, University of Calcutta, 1978.
11. Wilde, D. J., 'Optimum Seeking Methods' (Prentice Hall, Englewood Cliffs, N. J.), 1964.
12. Ullman, J. R., 'Pattern Recognition Techniques' (Butterworths, London), 1973.
13 Robbins, H. & Monro, S., *Ann. Math. Stat.*, **22** (1951), 400-407.
14. Blum, J. R., *Ann. Math. Stat.*, **25** (1954), 382-386.
15. Dvoretzky, A., 'On Stochastic Approximation', *Proc. 3rd. Berkeley Symp. on Mathematical Statistics and Probability*, (J. Neyman, ed.), University of California Press, Berkeley, pp. 39-55, 1956.
16. Davis, M. H. A., 'Linear Estimation and Stochastic Control' (Chapaman and Hall, London), 1977.
17. Chien, Y. T. & Fu, K. S., *IEEE Trans. Syst., Man aud Cyberns.*, **SMC-3** (1967), 28-38.
18. Nikolic, Z. J. & Fu, K. S., *Proc. Nat. Electron. Conf.*, **22** (1966), 607-612.
19. Chien, Y. T., *Inform. Sci.*, **2** (1970), 351-368.
20. Pal, S. K., Datta, A. K. & Dutta Majumdar, D., A Self-supervised Vowel Recognition System, *Pattern Recognition*, (to appear).
21. Dudani, S. A., Freeding, K. J. & McGhee, R. B., *IEEE Trans. Comput.*, **C-26** (1977), 39-46.
22. Zahn, C. T. & Roskies, R. Z., *IEEE Trans. Comput.*, **C-21** (1972), 269-281.
23. Person, E. & Fu, K. S., *IEEE Trans. Syst., Man. Cybern.*, **SMC-7** (1977), 170-179.
24. Freeman, H. & Davis, L. S., *IEEE Trans. Comput.*, **C-26** (1977), 297-303.
25. Freeman, H. & Saghri, A., 'Generalized Chain Codes for Planar Curves', Proc. 4th Int. Joint Conf. Pattern Recognition, Kyoto, Japan, 7-10 Nov., 1978, pp. 701-703.
26. Pavlidis, T. & Ali, F., *IEEE Trans. Syst., Man. Cybern.*, **SMC-5** (1975), 610-614.
27. Parui, S. K. & Dutta Majumdar, D., *Pattern Recognition Letters*, **1** (1982), 37-42.
28. Parui, S. K. & Dutta Majumdar, D., 'Symmetry Analysis by Computer', to appear in Pattern Recognition.
29. Pavlidis, T., *Comput. Graphics Image Processing*, **7** (1978), 243-258.
30. Pavlidis, T., *IEEE Trans. PAMI*, **2** (1980), 301-312.
31. Dutta Majumdar, D., Parui, S. K. & Chaudhuri, B. B., *Jour. IETE*, **26** (1980), 21-28.
32. Dutta Majumdar, D., Pattern Recognition Methods and Applications—A State of the Art Review. Recent Trends of Res. Pattern Recognition and Digital Techniques, February 1977.
33. Dutta Majumdar, D. & Pal, S. K., *IEEE Trans. on System Man and Cybernetics*, (Aug. 1977), 625–629.
34. Dutta Majumdar, D., Pal, S. K. & Chaudhuri, B. B., Some Experiments on Computer Recognition of Speech Pattern. Proc. Int., Conf. on Information Sciences and Systems, August 19-24, 1976. Patres, Greece, Hemisphere Publishing Corporation, USA.
35. Dutta Majumdar, D., *Int. Jour. of System Sciences*, **9** (1978), 837-886.

36. Dutta Majumdar, D. & Pal, S. K., *IEEE Trans.*, **SMC-8**, (1978), 4.

37. Dutta Majumdar, D., Parui, S. K. & Chaudhuri, B. B., *Journal of the Institution of Electronics and Telecommunication Engineers*, **26** (1980), 21-28.

38. Dutta Majumdar, D., Parui, S. K. & Chaudhuri, B. B., Computer Recognition of Handwrittern Numerals by Syntactic Method, *IEEE Trans. PAMI* (to be published).

39. Dutta Majumder, D., Mathematical and Statistical Tools in Discriminant Analysis for Pattern Recognition Problems, Proc. Int. Discriminant Analysis Conf., Czechoslovak Academy of Sciences, Liblice, Prague, 27th Sept.-1st Oct. 1982.

40. Dutta Majumdar, D., Chanda, B. & Chaudhuri, B. B., A Neighbour Weighting Algorithm for Image Smoothing. *Proc. Int. Conf. on Comm. Circuitry and Systems*, J. U., December 1981.

41. Dutta Majumdar, D., Choudhury, B. B. & Murthy, C. A., *IEEE Trans.*, **SMC-13**, (1983), 1.