CHAPTER 18

Pattern Recognition and Classification

Author-Editors: ROBERT M. HARALICK and KING-SUN FU

GENERAL CONTENTS: Basic pattern recognition concepts. Principles of spectral discrimination. Economic consequences of decisions. The Bayes decision rule. The maximum decision rule. The Gaussian assumption. Feature selection. Principle components. Bhattacharyya distance. Syntactic pattern recognition applied to remote sensing problems. Recognition of clouds and shadows. References.

INTRODUCTION

BASIC PATTERN RECOGNITION CONCEPTS

Pattern recognition as used here refers to the automatic machine determination of salient patterns in remotely sensed image data. From the pattern-recognition perspective, the world to be sensed is composed of units defined by the sensor. For digital imaging sensors, as a first approximation, the units can be thought of as small nonoverlapping areas on the ground: one such area for each picture element (pixel) in the image. The sensor makes an ordered set of measurements on each unit sensed. The ordered set of measurements is called a measurement vector or measurement pattern. Each value measured in this set is a number proportional to the energy received by the sensor in some band of the electromagnetic spectrum at some specified observation time. The basic pattern-recognition problem is first to automatically and consistently determine the information class or category of each distinct region on the ground using the set of sensor measurement-patterns and second to estimate the error rate for the automatically determined assignments.

Specific examples of pattern recognition for remote sensing applications include determining

1) tree-species composition in a forest
2) hot spots of incipient forest fires
3) natural vegetation cover-types
4) crop types
5) state of health or stressed vegetation
6) percent of sedimentation in a river or lake
7) percent of pollutant in a river or lake
8) geological formation and rock types
9) lineament patterns
10) degree of mineralization
11) number of small objects in a smooth background
12) urban land-use patterns

The automation of these tasks requires a corresponding variety of methods and techniques varying from simple to highly complex. It is the purpose of this chapter to describe the most commonly used techniques.

LITERATURE DEALING WITH PATTERN RECOGNITION CONCEPTS

Books describing the principles of pattern recognition have been written by Sebestyen (1962), Nilsson (1965), Arkadev and Braverman (1966), Fu (1968), Kanal (Ed.) (1968), Watanabe (Ed.) 1969), Mendel and Fu (1970), Fu (Ed.) (1971), Andrews (1972), Fukunaga (1972), Meisel (1972), Patrick (1972), Watanabe (Ed.) (1972), Chen (1973), Duda and Hart (1973), Ullman (1973), Tou and Gonzalez (1974), Batchelor (1974), Young and Calvert (1974), Fu and Whinston (Ed.) (1977), and Batchelor (1978). Some of these books have been reviewed and the reader might be interested in consulting the reviews listed in Table 18-1 before attempting to read any of these books.

Shorter reports and review articles include those by Nagy (1968), Ho and Aggrawala (1968), Fu, Landgrebe, and Phillips (1969), Casy and Nagy (1971), Nagy (1972), Kanal (1972), and Kanal (1974). Reprints of important pattern-recognition articles can be found in Sklansky (1973) and Aggrawala (1977). The May 1979 issue of the **IEEE Proceedings** was a special issue on pattern recognition and image processing. Journal papers on pattern recognition appear in the **IEEE Transaction on Computers, IEEE Transactions on Systems, Man and Cybernetics,** and **IEEE Transaction on Pattern Analysis and Machine Intelligence.** The Pattern Recognition Society publishes a journal called **Pattern Recognition.** Conference papers appear in the **International Joint Conference on Pattern Recognition, The Pattern Recognition and Image Processing Conference, The Purdue Symposium on Machine Processing of Remotely Sensed Data,** and the **Environmental Research Institute of Michigan Remote Sensing of Environment Conferences.**

SUMMARY RELATIVE TO PATTERN RECOGNITION CONCEPTS

To automate pattern recognition, we must define the classes of entities of interest, that is, the kinds of objects between which we must discriminate; we must choose instruments or sensors

TABLE 18-1

Listing of Various Books on Pattern Recognition Where They Have Been Reviewed

AUTHOR(S)	TITLE	WHERE REVIEWED
Harry Andrews	Introduction to Mathematical Techniques in Pattern Recognition; Prentice Hall, New Jersey, 1972, 504 pages.	IEEE Information Theory vol. IT-19, no. 6, November, 1973, p. 831
Richard Duda & Peter Hart	Pattern Classification & Scene Analysis; Wiley, New York, 1973, 482 pages.	IEEE Computer Transactions, vol. C-23, no. 2, February, 1974, p. 223 IEEE Information Theory, vol. IT-19, no. 6, November, 1973, p. 827–829.
King-Sun Fu	Syntactic Methods in Pattern Recognition; Academic Press, New York, 1974, 397 pages.	IEEE Systems Man Cybernetics, vol. SMC-6, no. 8, August, 1976, p. 590.
Keinosuke Fukunaga	Introduction to Statistical Pattern Recognition; Academic Press, New York, 1972, 382 pages.	IEEE Systems Man Cybernetics, vol. SMC-4, no. 2, March, 1974, p. 238. IEEE Information Theory, vol. IT-19, no. 6, November 1973, pp. 829–830.
William Meisel	Computer-Oriented Approaches to Pattern Recognition; Academic Press, New York, 1972, 262 pages.	IEEE Systems Man Cybernetics, vol. SMC-3, no. 2, March, 1973, p. 209. IEEE Computer Transactions, vol. C-23, no. 1, January, 1974, p. 112. IEEE Computer Transactions, vol. C-22, no. 4, April, 1973, p. 429. IEEE Information Theory, vol. IT-19, no. 6, November, 1973, pp. 832–833.
Edward Patrick	Fundamentals of Pattern Recognition; Prentice Hall, New Jersey, 1972, 528 pages.	IEEE Systems Man Cybernetics, vol. SMC-3, no. 5, September, 1973, p. 528. IEEE Information Theory, vol. IT-19, no. 6, November, 1973, pp. 830–831.
Julius Tou and Rafael Gonzales	Pattern Recognition Principles, Addison-Wesley; Mass. 1974, 377 pages.	IEEE Systems Man Cybernetics, vol. SMC-6, no. 4, April, 1976, pp. 332–333. IEEE Information Theory, vol. IT-22, no. 5, September, 1976, pp. 632-633.
Jullian Ullmann	Pattern Recognition Techniques; Crane-Russak, New York, 1973, 412 pages.	IEEE Computer Transactions, vol. C-23, no. 2, February, 1974, pp. 220–222 IEEE Information Theory, vol. IT-20, no. 3, May, 1974, p. 400.
Satosi Watanabe (Ed.)	Methodologies of Pattern Recognition; Academic Press, New York, 1969, 579 pages.	IEEE Information Theory, vol. IT-17, no. 5, September, 1971, pp. 633–634.

which can measure the environment in which the objects occur; and we must provide a methodology permitting the recognition of an object in the class of objects of interest from those not in the class of objects of interest. Using this methodology we also must construct a decision rule which will decide what kind of object a particular object is, on the basis of the measurements made from the observed small-area ground patches.

Defining the class of objects of interest might seem to be easy since it is an intrinsic part of the automation need. We will see, however, that it is not so easy since the sensor may not gather sufficient information to allow the discrimination to take place. In these cases we may be forced to define our classes as the more discriminable ones even though they may be of less interest to us. To help us do this we need to employ a clustering process which tells us what the naturally distinguishable classes are given the sensor's data.

Choosing the measuring instruments or sensors and designing a way to preprocess—to standardize, to normalize, and to extract the relevant information in its simplest form from the measurements—so that objects of interest can be simply recognized from those of non-interest (and so that each class or category of objects of interest has a particularly simple description in terms of the preprocessed measurements) are among the most difficult problems in pattern recognition. These problems are called feature-extraction- or preprocessing-problems and are concerned with presenting in some standard form only the simplest and most important information to the decision rule.

Finally, the problem in constructing a decision rule we call the pattern-discrimination problem. It is based on a probability model and it allows us to estimate the error rates of the automatic decision process.

Most pattern recognition of remotely sensed image data is done by processing each pixel's information separately or independently. This means that a category assignment is made to each pixel purely on the basis of its own information. Processing proceeds on a pixel-by-pixel basis over the entire image.

When the pixel's information consists only of the sensor measurement-pattern obtained from one observation time, the measurement pattern is called a multispectral feature-vector and the kind of pattern recognition is called multispectral pattern-recognition. When items of spectral information from more than one observation time for the same ground area are stacked in the same measurement-pattern vector, this kind of pattern recognition is called multispectral-multitemporal pattern-recognition. When the measurement pattern for each pixel contains spectral information from its associated ground area as well as from neighboring ground areas, or when the decision rule which makes category assignments uses the information from a pixel and some of its neighboring pixels, the pattern recognition is called spatial pattern recognition, or spatial-spectral pattern recognition.

PRINCIPLES OF SPECTRAL DISCRIMINATION

In order to understand the pattern-discrimination methodology consider a simplified example. Suppose that there are three types of surface-cover material: vegetation, soil, and water. Suppose further that each of these has a unique spectral response which does not vary with season, atmospheric haze, sun-angle etc. Let these be the responses shown in Figure 18-1. Now select two wavelengths λ_1 and λ_2 for a remote sensor to make some measurements. Then, for each surface-cover category, use wavelengths λ_1 and λ_2 to determine its spectral measurement pattern. Plot these in measurement space as shown in Figure 18-2. Since they obviously plot in areas that are nicely separated from each other we would expect no difficulty in designing a decision rule to recognize these categories. Any time a new measurement pattern needs to be assigned to a category we see if it lies as the point in measurement space associated with vegetation, or soil, or water. If it does, we assign it to the corresponding category. If it doesn't we assign it to an unknown category.

In reality, the spectral response patterns from these surface categories as well as others vary due to natural random variations, systematic seasonal causes, atmospheric haze, etc. There is not a unique measurement pattern associated with each category. Rather, associated with each category is a probability distribution indicating, for any measurement pattern, the relative frequency of occurrence that may arise from a ground area of the given category.

If, using some training data, we plotted five observations of each of three vegetation categories, viz. soybean, corn and wheat, we might obtain the measurement-space plot of Figure 18-3. To assign a new measurement pattern, v, to one of the classes is now not such an easy problem. In essence we must use our training observations to estimate for each new measurement pattern v, the probability that soybeans, corn or wheat is its true category. If we can do that we can associate with each measurement that category having the highest conditional probability given the measurement. In effect, this association partitions measurement space as shown in Figure 18-4. Since our new measurement pattern is in the part of measurement space associated with soybeans the decision rule assigns it to the soybean class.

The procedure by which the measurement space of Figure 18-3 was partitioned is simple. Use the training data for each class to determine the class sample-mean. Then partition the measurement space so that each class has associated with it all the measurement patterns closest to its

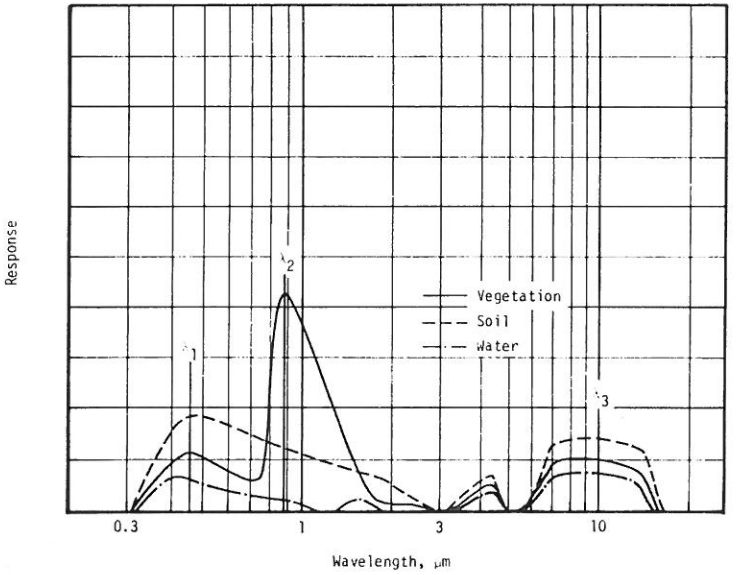

Fig. 18-1. Typical relative response curves for different materials, illustrating the possibility of discrimination by comparison of the curves at different wavelengths. Source: Landgrebe (1972b).

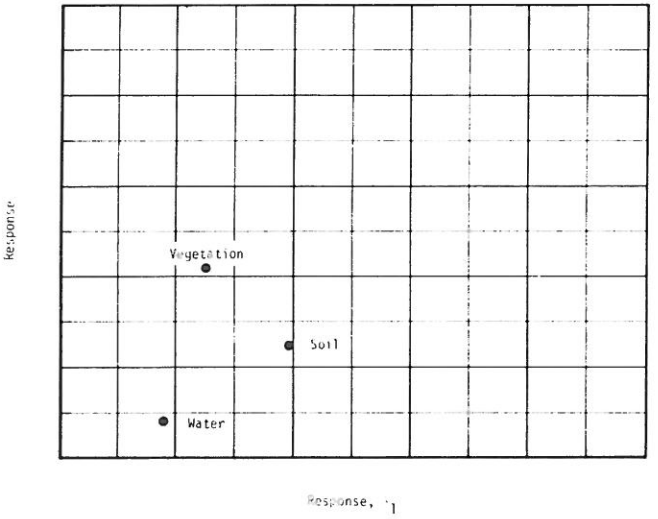

Fig. 18-2. The categories vegetation, soil, and water have distinct responses on wavelengths λ1 and λ2. Shown in this figure are these categories plotted in a measurement space whose axes are their λ1 and λ2 responses.

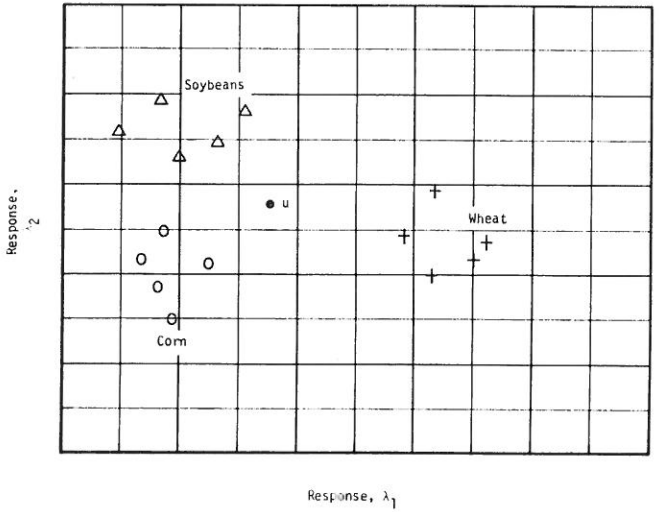

Fig. 18-3. A given material will not always have the exactly same response in a group of samples, but each material tends to cluster together. A typical two-dimensional sampling of three materials is shown. Source: Landgrebe (1972).

sample mean. Unfortunately, without a probability model we cannot say that this procedure is the one that yields the lowest error rate or maximizes any utility function. However, there is a probability model under which this is the appropriate thing to do.

It is the purpose, therefore, of the next sections to develop a probabilistic-decision theoretic-model for pattern discrimination which suggests techniques for decision-rule construction having certain optimal properties which we can measure in terms of utility or economic consequences.

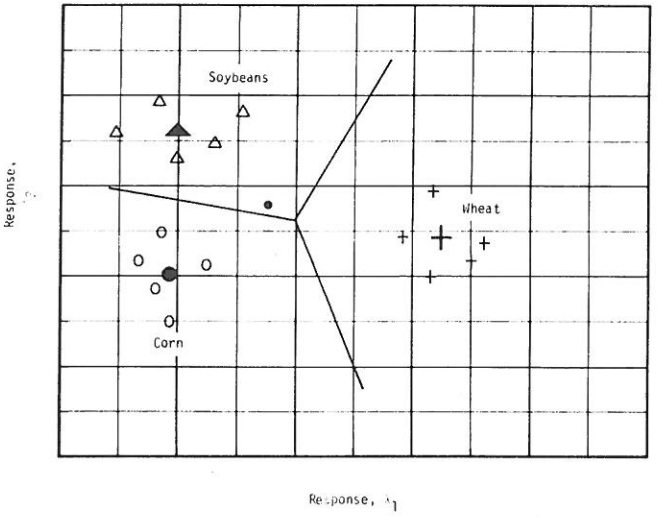

Fig. 18-4. Division of two-dimensional sampling space into domains assigned to different materials. In this case the unknown point u is considered to be soybean because of its location in the sampling space. Source: Landgrebe (1972).

ECONOMIC CONSEQUENCES OF DECISIONS

For each pattern d belonging to D, $d \epsilon D$, a decision rule f assigns a category alternative c^k from the set of category alternatives $C = \{c^1, \ldots, c^K\}$. The assignment may be deterministic or probabilistic. In any case, we assume that the assignment by the decision rule of category alternative c to a pattern d measured from a unit u carries economic consequences. These economic consequences are determined by the people who need to automate the discrimination ability of the trained human observer. The consequences are generally good when the chosen category alternative c is, in fact, the true category identification of the unit u. The consequences are generally bad when the category alternative c is not the true category identification of the unit u. Because such identification decisions must be made, and because they cause consequences when they are made, we may view the goal of decision-rule construction as the construction of a decision rule which in some sense maximizes the good consequences.

To speak of maximizing good consequences implies that we must have some numerical measure indicating the economic gain or loss of the consequence when the decision rule assigns category c^i to a unit u with measurement d when the true category identification of unit u is category c^j. Let $e(c^j,c^i)$ be the net worth or economic gain of such a consequence. In general, $e(c^i,c^i)$ will be positive signifying a gain for a correct identification, and $e(c^j,c^i)$, for $i \neq j$, will be negative, signifying a loss for an incorrect identification.

In determining a decision rule, we must choose a criterion of optimality by which we can judge the worth of the decision rule on the basis of the various economic gains or losses of the consequences (c^i,c^j). The optimality criterion defines how to judge how well the decision rule balances, in terms of these gains and losses, the possible consequences of its decision. The most often-used criterion is one which defines the best decision rule to be one which maximizes the expected gain under certain given conditions. Such a rule is called a Bayes decision rule.

Let us consider the economic gains of the possible consequences given that a unit u has measurements d. These gains are illustrated simply in Figure 18-5. Suppose the decision rule assigns a unit u having measurements d to category c^i. This assignment, at best, however, is only an educated guess; the true category identification for unit u can actually be any one of $c^1, c^2, \ldots, c^K$. In Figure 18-5 the decision-rule assignment of c^i corresponds to a selection of the i^{th} column. The true category identification of unit u corresponds to a selection of some row. This row, intersected with the i^{th} column, yields an entry which is the economic gain consequence.

The question of concern is how often will the true category identification of a unit u be category c^j when the unit u has measurement d. We denote by $P_d(c^j)$ the probability of the true category identification of a unit u being in category c^j given that the unit u has measurements d. It is these conditional probabilities which can be estimated from the training data or ground-observation data.

	c^1	c^2		c^K
$P_d(c^1)$	$e(c^1, c^1)$			$e(c^1, c^K)$
$P_d(c^2)$				
NATURE				
$P_d(c^K)$	$e(c^K,c^1)$			$e(c^K, c^K)$

Fig. 18-5. This shows the economic gains obtained under various alternatives conditioned on the measurement d being made of a unit K. Given that the observed measurement is d, the probability that nature chooses category c^j, corresponding to the j^{th} row, is $P_d(c^j)$. The decision rule will choose some category c^K, corresponding to the K^{th} column. The result of nature choosing category c^j and the decision rule choosing category c^K is the economic consequence (c^j, c^K).

The decision rule has no information regarding the true category identification of any unit. It only knows that the unit gives rise to a pattern d and that it has available estimates of the conditional probabilities $P_d(c^k)$, $k = 1, 2, \ldots, K$. The decision rule must assign the unit to a category, say c^i. This corresponds to a section of the i^{th} column. For this course of action a number of different consequences can occur. If the true category identification is c^1, then the gain of the consequence (c^1,c^i) is $e(c^1,c^i)$. If the true category identification is c^2, then the gain of the consequence (c^2,c^i) is $e(c^2,c^i)$. In general, if the true category identification is c^j, then the gain of the consequence (c^j,c^i) is $e(c^j,c^i)$. The next section discusses a decision-rule construction-procedure which maximizes the expected gain.

THE BAYES DECISION RULE MAXIMIZES EXPECTED GAIN

Let $f_d(c^i)$ denote the probability that the decision rule assigns the category c^i to the unit, given that the unit has pattern measurement d. Since, for any pattern d, there is no reason to suppose any interaction or collaboration between nature, (which may be thought of as choosing the true category identification) and the pattern discriminator, (which may be thought of as employing the decision rule to assign categories) we may assume that nature and the pattern discrimination are statistically independent. Thus, the probability that the unit has measurements d and the deci-

sion rule assigned the category c^k to the unit and the true category identification for the unit is c^j may be written as $f_d(c^k)P_d(c^j)P(d)$. Therefore, the expected gain for the decision rule f may be expressed by

$$E[e;f] = \sum_{d\epsilon D} \sum_{j=1}^{K} \sum_{k=1}^{K} e(c^j,c^k) f_d(c^k) P_d(c^j) P(d). \tag{18-1}$$

To see how to find the decision rule which maximizes the expected gain, we rewrite the expression for $E[e;f]$ as

$$E[e;f] = \sum_{d\epsilon D} P(d) \sum_{k=1}^{K} f_d(c^k) \sum_{j=1}^{K} e(c^j,c^k) P_d(c^j). \tag{18-2}$$

$P(d)$, being the probability of measuring pattern d for a unit, is non-negative. Hence $E[e;f]$ will be maximized (maximum taken over all f) if and only if for each $d\epsilon D$ the expected gain given d using f is maximized; that is,

$$E[e|d;f] = \sum_{k=1}^{K} f_d(c^k) \sum_{j=1}^{K} e(c^j,c^k) P_d(c^j) \text{ is maximized.} \tag{18-3}$$

Since $\sum_{k=1}^{K} f_d(c^k) = 1$ and $f_d(c^k) \geq 0$, $k = 1,2,\ldots,K$, it is easy to see that the maximum of the above expression is

$$i = 1,2,\ldots,K \sum_{j=1}^{K} e(c^j,c^i) P_d(c^j) \tag{18-4}$$

and the decision rule f will certainly achieve this maximum if

$$f_d(c^i) = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \text{ where } k \text{ is any index such that}$$

$$\sum_{j=1}^{K} e(c^j,c^k) P_d(c^j) \geq \sum_{j=1}^{K} e(c^j,c^i) P_d(c^j), \; i = 1,2,\ldots,K. \tag{18-5}$$

In this case the optimal decision rule can be deterministic if the index k is unique or it can be either deterministic or probabilistic if k is not unique. Any optimal decision rule is called a Bayes rule.

For example, suppose there are three categories c^1, c^2, and c^3 with conditional probabilities and economic gains for the various alternatives and consequences as shown in Figure 18-6. The optimal decision rule will assign the unit u to category c^3 since the average gain for row 3 is 5/6 which is larger than the average gain for row 1 which is $-1/3$ or for row 2 which is $1/2$.

	c	c^1	c^2
$P_u(c^1) = 1/6$ c^1	4	2	0
$P_u(c^2) = 1/2$ c^2	0	1	0
$P_u(c^3) = 1/3$ c^3	−1	0	3

Fig. 18-6. Illustrates the economic gains for an example problem where the pattern measurements d are made on a unit and there are three possible categories.

BAYES DECISION RULES AND CATEGORY PRIOR PROBABILITIES

It is often the case that the conditional probabilities $P(c/d)$ are not known but that the conditional probabilities $P(d/c)$ of the measurements, given the categories, are known. Fortunately, there is a well known relationship between $P(c/d)$ and $P(d/c)$ which involves the prior probabilities of $P(d)$ and $P(c)$ of the measurements and categories, respectively.

By the definition of conditional probability, we may express $P_d(c)$ by

$$P_d(c) = \frac{P_c(d)P(c)}{P(d)} \tag{18-6}$$

so that the average gain obtained by the use of decision rule f may be rewritten as

$$E[e;f] = \sum_{d\epsilon D} \sum_{k=1}^{K} \sum_{j=1}^{K} f_d(c^k) e(c^j,c^k) P_{c^j}(d) P(c^j). \tag{18-7}$$

$E[e;f]$ is maximized if and only if for each $d\epsilon D$, the gain conditioned on d,

$$E[e|d;f] = \sum_{k=1}^{K} f_d(c^k) \sum_{j=1}^{K} e(c^j,c^k) P_{c^j}(d) P(c^j) \tag{18-8}$$

is maximized. The maximum value of $E[e/d;f]$ is

$$\sum_{j=1}^{K} e(c^j,c^k) P_{c^j}(d) P(c^j) \tag{18-9}$$

where k is some index for which

$$\sum_{j=1}^{K} e(c^j,c^k) P_{c^j}(d) P(c^j) \geq \sum_{j=1}^{K} e(c^j,c^i) P_{c^j}(d) P(c^j), \; i = 1,2,\ldots,K. \tag{18-10}$$

An optimal deterministic decision rule f may therefore be defined by

$f_d(c^i) = 1\ i = k$
$\quad\quad = 0\ i \neq k$ } where k is any index such that

$$\sum_{j=1}^{K} e(c^j, c^k) P_{c^j}(d) P(c^j) \geq \sum_{j=1}^{K} e(c^j, c^i) P_{c^j}(d) P(c^j), \quad i = 1, 2, \ldots, K. \qquad (18\text{-}11)$$

Note the strong dependence which f has on the category probability $P(c)$. Because of this, any time we define an optimal Bayes decision rule, we must state that it is optimal only relative to the category prior-probability function $P(c)$.

MAXIMIN DECISION RULE

Figure 18-7 illustrates the expected gain of a Bayes decision rule in a two-category classification problem with the identity gain function. Selecting a value of prior probability, the corresponding value of expected gain is the highest expected gain achievable by any decision rule. Therefore, use of any decision rule which is not a Bayes rule is guaranteed to perform below the curve. In particular, if a Bayes rule is used in a new situation where the encountered prior probability function differs from the one employed in the design, then the Bayes rule is not optimal in the new situation.

Recognizing this, a conservative decision-rule designer will attempt to construct a decision rule which maximizes the smallest gain achieved by the decision rule under some encountered prior probability function. It turns out that the expected gain for a decision rule which maximizes the smallest gain as the encountered prior probability function varies has a value equal to the smallest possible Bayes gain (the lowest point on the curve of Figure 18-7). This value is the same regardless of the actually encountered prior probability functions. This kind of decision rule is called a maximin decision rule since it maximizes the minimum expected gain. In general, it is not a deterministic decision rule and designing a maximin decision rule is equivalent to solving a large linear-programming problem.

THE GAUSSIAN ASSUMPTION

The conditional probability $P_c(d)$ of the data-measurement vector d given the category c plays an essential role in decision rule determination. $P_c(d)$ could be stored as a table. However, because of the large number of possible data vectors, $P_c(d)$ is often represented as a parametric function, the parameters being the category mean measurement μ_c and its covariance matrix Σ_c. The simplest probability density-function having these parameters is the Gaussian one which is defined by

$$P_c(d) = \frac{1}{(2\pi)^{N/2} |\Sigma_c|^{1/2}} e^{-1/2 (d - \mu_c)' \Sigma_c^{-1} (d - \mu_c)} \qquad (18\text{-}12)$$

where N is the dimension of the measurement vector d.

In the case of identity-gain function e and equal probabilities for the prior $P(c)$ for each category c, the Bayes decision rule assigns measurement d to that category c minimizing

$$\log |\Sigma_c| + (d - \mu_c)' \Sigma_c^{-1} (d - \mu_c) \qquad 18\text{-}13$$

This kind of decision rule is sometimes called quadratic or piecewise quadratic because the de-

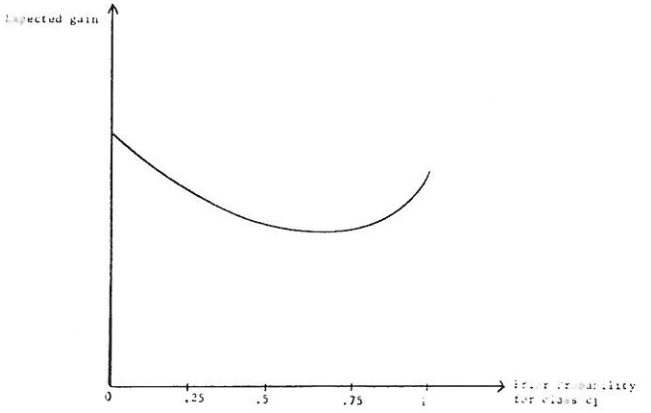

Fig. 18-7. Illustrates how the expected gain of a Bayes decision rule can vary with a change in prior probability for class c_1 in some two class example problem. Notice that as the prior probability for class c_1 becomes 1, the prior certainty reflects itself in an a posteriori certainty which makes the expected gain high. When the prior probability for class c_1 becomes 0, the prior probability for class c_2 becomes 1 and the situation is similar. For class c_1 prior probabilities between 1 and 0, the prior situation is less certain and the expected gain must be less than the end cases. The shape of the function is guaranteed to be convex.

cision boundaries they form in measurement space are piecewise hyperquadratic boundaries.

In case the covariance matrices for all categories are equal, the Bayes decision rule reduces to assigning the measurement d to that category c minimizing the Mahalanobis distance

$$(d - \mu_c)' \sum^{-1} (d - \mu_c) \qquad 18\text{-}14$$

between the measurement vector d and the category mean vector μ_c for category c. This decision does simplify to a linear decision rule. Assign measurement d to that category c maximizing

$$(\mu' \Sigma^{1}) d - \left(\frac{\mu_c' \Sigma^{-1} \mu_c}{2} \right) \qquad (18\text{-}15)$$

By precomputing the terms in parentheses, the number of multiply and add operations for this decision rule is only $N + 1$ per category, a significant saving over the quadratic rule, especially when the dimensionality N is large.

FEATURE SELECTION

Multitemporal multispectral remotely sensed imagery can produce a ten- or twenty-dimensional data vector for each pixel. The data have inherent redundancies and processing all of the data or storing all of them may not be cost effective. Feature-selection procedures are used to select those dimensions most suitable for processing.

There are two kinds of feature selections depending on whether the classes and their statistics are known or not known. If they are not known, the best feature-selection procedure is called principal components. If they are known, the easiest-to-use feature selection is based on Bhattacharyya distance.

PRINCIPAL COMPONENTS

Principal components is a standard statistical technique for selecting that subspace of given dimension in which the most data variance lies. If $x_1, \ldots, x_N$ are the sample data vectors, μ the sample-mean vector, and $\sum$ the sample covariance-matrix, the best K dimensions in which to project the data would be that K-dimensional subspace spanned by the K eigenvectors of Σ having the largest eigenvalue. Thus if T is a matrix whose K rows are these eigenvectors, the K principal components of $x_1, \ldots, x_N$ is $Tx_1, \ldots Tx_N$, each Tx_i being a K-dimensional vector.

BHATTACHARYYA DISTANCE

The Bhattacharyya distance is a measure of the separability between two classes. For two Gaussian classes having means and covariances μ_1, Σ_1 and μ_2, Σ_2 respectively, the Bhattacharyya distance is given by

$$\frac{1}{8} (\mu_1 - \mu_2)' \left(\frac{\Sigma_2 + \Sigma_2}{2} \right) (\mu_1 - \mu_2) + \frac{1}{2} \ln \frac{\left| \frac{\Sigma_1 - \Sigma_2}{2} \right|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}} \qquad (18\text{-}16)$$

To use this distance measure for selecting the best K features from the original N dimensions on an L-class problem, the Bhattacharyya distance needs to be calculated between each of the $L(L-1)/2$ pairs of classes for each of the $\binom{N}{K}$ possible ways of choosing K features from N dimensions. The K dimensions which are best are those K dimensions whose sum of the Bhattacharyya distances between the $L(L-1)/2$ pairs of classes is highest. The Bhattacharyya distance between a pair of classes for a selection of K dimensions out of N dimensions is calculated using the mean and covariance matrix in the selected K dimensions.

SYNTACTIC PATTERN RECOGNITION APPLIED TO REMOTE SENSING PROBLEMS

GENERAL APPROACH

The approach of using hierarchical structures and grammar rules to describe the structures of pattern has recently received increasing attention (Fu, 1974). This approach is often called the structural or syntactic approach to distinguish it from the decision-theoretic or statistical approach. Practical applications include the description of chromosome images, the recognition of characters, spoken digits, electrocardiograms, and two-dimensional matehmatical expressions, the identification of bubble chamber- and spark chamber-events, and the recognition of fingerprint patterns (Fu, 1978). In the syntactic approach, each pattern is described in terms of its parts, i.e., subpatterns. Each subpattern can again be described in terms of its parts. The simplest subpatterns are called the pattern primitives, and they constitute the basic symbols (the set of terminals) of the pattern language. The description of each primitive can be either deterministic or statistical and the recognition of primitives is often based on the decision-theoretic approach. Each class of patterns is now described by a set of sentences consisting of the primitives, and it can be generated by a pattern grammar. With the above description, it might be said that in the syntactic approach we often use the decision-theoretic approach for primitive recognition; however, the emphasis will be on the use of syntactic rules to describe the structure of patterns (the compositions rules of the primitives and subpatterns).

Multispectral signals measured by Landsat over Marion County (Indianapolis), Indiana were analyzed using clustering analysis. Fourteen clusters were found and the data from the urban

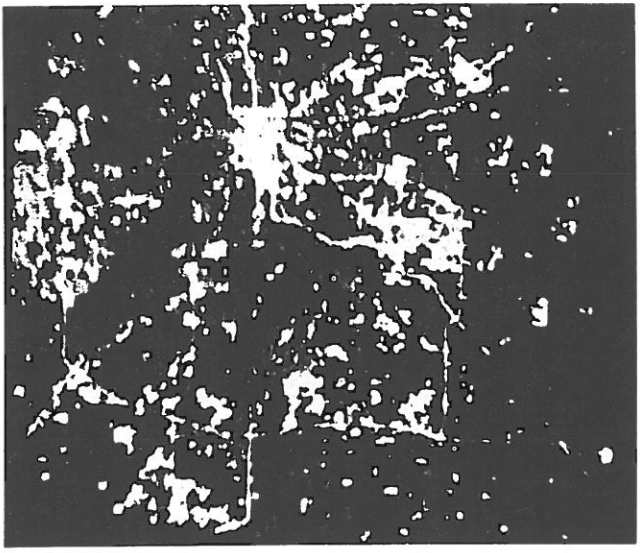

Fig. 18-8. Photograph of Marion County imagery from digital display.

area within the scene were accordingly classified using a Bayes classifier. The result of the Bayes classifier (Figure 18-8) provides the basic pattern primitives. Some manual preprocessing based on these dependencies was used to improve the accuracy of the classification. A hierarchical graph model for these relationships can be constructed as shown in Figure 18-9. Obviously, there are spatial dependencies among the various classes.

The hierarchical graph model shown in Figure 18-9, was constructed directly from observations of the classified data and of aerial photographs of the corresponding region. For simplicity, there are some relationships between the entities in the figure which have not been included; for example, the fact that the SCENE is made up of the EARTH and CLOUD PAIRs (i.e., clouds and shadows). The CLOUD PAIRs obscure the EARTH, so a relation *O* for "obscures" could be shown linking CLOUD PAIRs to EARTH. Also, if a pair of entities are related, then their descendants are also related. However, these relations are shown only at the level at which they first occur. The form of this diagram is the same as the derivation diagram for a web grammar.

This scene consists, at the highest level, of the EARTH obscured by CLOUD PAIRs. Each CLOUD PAIR consists of a CLOUD and a SHADOW, related by a distance-and-angle *R*. A CLOUD consists mostly of points classified as clouds (blank) but also points classified as concrete (X) and as suburban (S). This confusion seems to arise because both concrete and clouds are highly reflective. The suburban class is a mixture of concrete and grass. A SHADOW tends to consist mostly of points classified as shadows (★) but also points classified as commercial (C) and inner city (I). The confusion here seems to occur because the commercial and inner-city classes consist largely of asphalt rooftops with low reflectance.

The EARTH consists of URBAN and RURAL areas. The RURAL area consists of open grassy (O) and wooded (W) areas. The URBAN area consists of the DOWNTOWN area, surrounded by the INNER CITY area, with nearby SUBURBAN areas and a system of HIGHWAYS. The DOWNTOWN area is characterized by the fact it contains the largest concentration of commercial land use. The INNER CITY area surrounds the DOWNTOWN and contains a high concentration of inner city points. The SUBURBAN and HIGHWAY areas are both near the DOWNTOWN and contain mostly suburban-classified points. They are distinguished by the fact that HIGHWAYS occur in linear patterns. This model is now used to guide the analysis of the picture. In essence this analysis is an attempt to verify the model and to make it more specific. Each subentity of the picture and each relationship can now be elaborated and tested separately.

RECOGNITION OF CLOUDS AND SHADOWS

As pointed out in the previous section, clouds and shadows are characterized by the fact that a cloud is a bright area which has associated with it

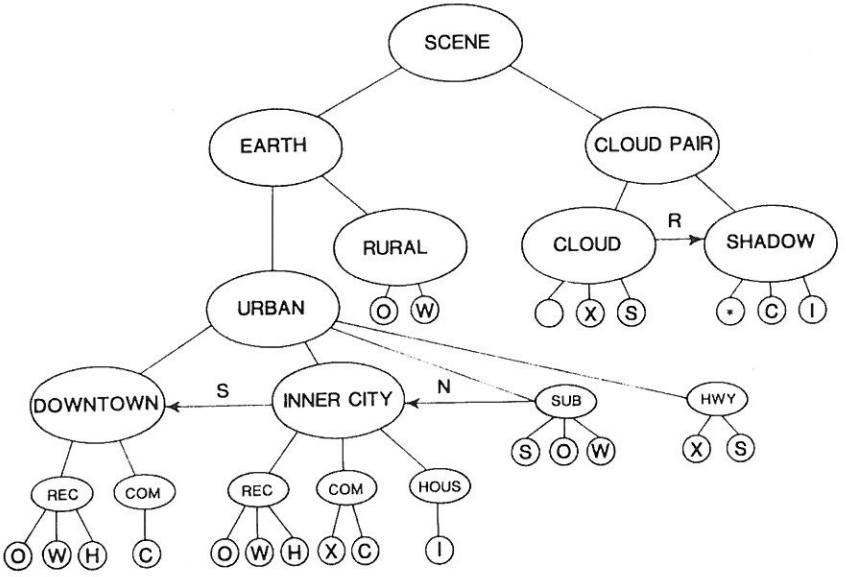

Fig. 18-9. A hierarchical graph model of the scene in Fig. 18-8.

a congruent dark area (the shadow) at a certain specific distance and orientation. It is found that the bright areas are generally classified by the pointwise classifier as clouds, but sometimes they are mistakenly classified as concrete or suburban. The dark areas are usually classified as shadows but sometimes as commercial or inner city. Examples can be observed in Figure 18-8.

If the relationship R between clouds and shadows is not known from some other source such as sun angle and height of clouds, the analysis process must determine R by some type of cross-correlation operation. For one-dimensional patterns, this process of finding similar patterns that are an arbitrary number of symbols away in the pattern can be modelled by a context-sensitive string operation. This is almost certainly also a context-sensitivie problem in a web system for two-dimensional patterns such as clouds and shadows. If the relation R is known from other sources the processing is simplified. Shape matching of a cloud and its partner shadow can be performed to confirm that they actually are pairs. This process is similar to the one-dimensional problem of being able to recognize all strings of the form ww^R (where w^R is w reversed). This is known to be a context-free operation in a string system.

Finally, an even simpler type of recognition would be to check a finite radius around a given cloud point. If there are more cloud points, this verifies the classification of the original point. Then an equivalent area at a distance given by the relation R can be searched. This process is illustrated in Figure 18-10. Since the radius r of the area searched is finite, the process is essentially finite state.

In the case of the clouds and shadows the simplest possible algorithm was tested first. The picture is scanned left-to-right and top-to-bottom. Whenever a shadow point is encountered, the translater searches a finite window at a distance and angle away given by relation R. If a cloud point is found, the pair qualifies as a cloud-shadow pair and neighboring ★, X, and S points within the window around the cloud are also interpreted as cloud points. Likewise if the pattern qualifies as a pair, neighboring C and I points

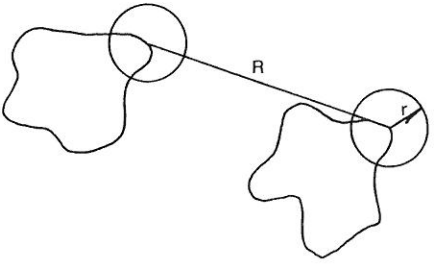

Fig. 18-10. Finite State Recognition of Clouds.

within a similar window centered at the original shadow point are classified as shadow points.

A grammar depicting this analysis is shown in Figure 18-11. Rules (1)–(4) generate cloud and shadow points that are not part of a pair. These points will be regarded as noise, not as true clouds and shadows. Rule (5) shows that a cloud-shadow pair (CP) can occur anywhere in the picture (the relation "a" represents "arbitrary" relationships). Rule (6) shows that a cloud-shadow pair consists of a cloud and a shadow separated by the relation R. Rules (7)–(10) show how a cloud can occur. The corresponding rules for a shadow are similar and are not shown. Rules (7) and (8) show that any cloud must contain at least one ★ point. Rule (9) shows that once one ★ point is detected, then any point labelled ★, X or S occurring within the window (i.e., within the relation w) is classified as a cloud point. Rule (10) terminates the search when the entire window has been scanned.

This grammar models the essential features of the recognition algorithm and therefore is good for illustration, but it is not complete in detail. Some complexities are buried in the relation w which models the window. A grammar which modelled this window in detail would contain a sequence of counting states that stimulate the scanning of the window. This more detailed grammar would contain a few more nonterminals and rules but would not better illustrate the complexity of the recognition process. As long as the window is of finite size, a regular-linear grammar can be found.

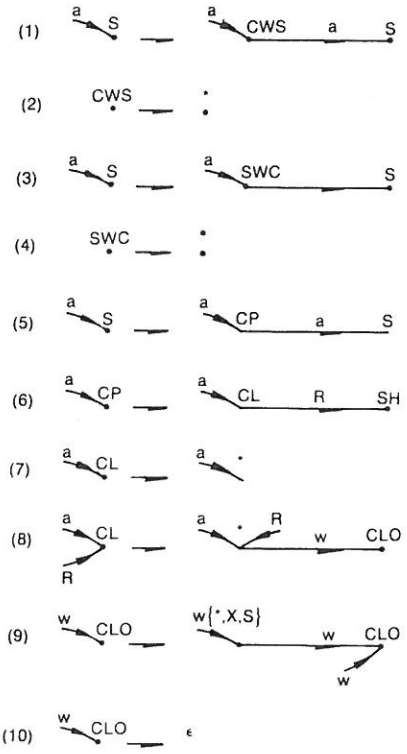

Fig. 18-11. The Cloud/Shadow Grammar.

REFERENCES

Aggrawala, A. K., (Ed.), 1977, Machine Recognition of Patterns: IEEE Press, New York.

Andrews, Harry, 1972, Introduction to Mathematical Techniques in Pattern Recognition: Prentice Hall, New Jersey.

Arkadev, A. G., and E. M. Braveman, 1966, Computer and Pattern Recognition: Thompson Book Company Inc., Washington, D.C.

Batchelor, B. G., 1974, Practical Approaches to Pattern Classification: Plenum Press, New York.

Batchelor, B. G., (Ed.), 1978, Pattern Recognition Ideas in Practice: Plenum Press, New York.

Casey, Richard A., and George Nagy, 1971, Advances in pattern recognition: Scientific American, vol. 224, no. 4, April, pp. 56–71.

Chen, C. H., 1973, Statistical Pattern Recognition: Hayden, Washington, D.C.

Duda, Richard, and Peter Hart, 1973, Pattern Classification and Scene Analysis: Wiley, New York.

Fu, K. S., 1968, Sequential Methods in Pattern Recognition and Machine Learning: Academic Press, New York.

Fu, K. S., D. A. Landgrebe, and T. L. Phillips, 1969, Information processing of remotely sensed agricultural data: Proceedings of the IEEE, vol. 57, no. 4, April, pp. 639–653.

Fu, K. S., (Ed.), 1971, Pattern Recognition and Machine Learning: Plenum Press, New York.

Fu, K. S., 1974, Syntactic Methods in Pattern Recognition: Academic Press, New York.

Fu, K. S., and A. B. Whinston, 1977, Pattern Recognition Theory and Application: Noordhoff-Leyden, Netherlands.

Fukunaga, Keinosuke, 1972, Introduction to Statistical Pattern Recognition: Academic Press, New York.

Ho, Y. C., and A. K. Aggrawala, 1968, On pattern classification algorithms introduction and survey: Proceedings of the IEEE, vol. 56, no. 12, December, pp. 2101–2114.

Kanal, L. N. (Ed.), 1968, Pattern Recognition: Thompson Book Company, Washington, D.C.

Kanal, L. N., 1972, Interactive pattern analysis and classification system: a survey and commentary: Proceedings of the IEEE, vol. 60, no. 11, October, pp. 1200–1215.

Kanal, L. N., 1974, Patterns in pattern recognition: 1968–1974: IEEE Transactions on Information Theory, vol. IT-20, no. 6, November, pp. 697–722.

Meisel, William, 1972, Computer-Oriented Approaches to Pattern Recognition: Academic Press, New York.

Mendel, J. M., and K. S. Fu, (Eds.), 1970, Adaptive, Learning and Pattern Recognition Systems: Theory and Applications: Academic Press, New York.

Nagy, George, 1968, State of the art in pattern recognition; Proceedings of the IEEE, vol. 56, no. 5, May, pp. 836–862.

Nagy, George, 1972, Digital image-processing activities in remote sensing for earth resources; Proceedings of the IEEE, vol. 60, no. 10, October, pp. 1177–1199.

Nilsson, N. J., 1965, Learning Machines; McGraw-Hill, New York.

Patrick, Edward, 1972, Fundamentals of Pattern Recognition; Prentice Hall, New Jersey.

Sebestyen, G. S., 1962, Decision Making Processes in Pattern Recognition; The Macmillan Company, New York.

Sklansky, J., (Ed.), 1973, Pattern Recognition, Introduction and Foundation; Dowden, Hutchinson and Rose Inc., Pennsylvania.

Tou, Julius, and Rafael Gonzalez, 1974, Pattern Recognition Principles; Addison-Wesley, Mass.

Ullman, Jullian, 1973, Pattern Recognition Techniques; Crane-Russak, New York.

Watanabe, Satosi, (Ed.), 1969, Methodologies of Pattern Recognition; Academic Press, New York.

Watanabe, Satosi, 1972, Frontiers of Pattern Recognition; Academic Press, New York.

Young, T. Y., and T. W. Calvert, 1974, Classification, Estimation, and Pattern Recognition; American Elsevier, New York.