

THE UNIVERSITY OF MICHIGAN
OFFICE OF RESEARCH ADMINISTRATION
ANN ARBOR

AN ABSTRACT MODEL FOR THE PATTERN RECOGNITION PROCESS

Technical Report No. 121

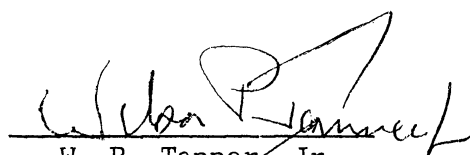
3642 - 1 - T

AFOSR 488

Cooley Electronics Laboratory
Department of Electrical Engineering

By: Donald L. Richards

Approved by


W. P. Tanner, Jr.

Project 3642

Contract No. AF 49(638)-884
Air Force Office of Scientific Research
Air Research and Development Command

April 1961

TABLE OF CONTENTS

	Page
ABSTRACT	iv
1. INTRODUCTION	1
2. THE FORMAL MODEL	2
2.1 Notation and Organization	2
2.2 Patterns and Inputs	4
2.3 Initial Knowledge	5
2.4 Learning	9
2.5 Decision	10
3. DIAGRAM OF THE SYSTEM	10
4. CONCLUDING DISCUSSION	12

ABSTRACT

The goal of this work, unlike that of other work in the field of pattern recognition, is the construction of a general recognition model which can apply to any group of patterns and to any amount of initial knowledge about them. In the preliminary model which has been developed, the initial knowledge is presented to the recognizer as a list of conditional probabilities; then, given certain characteristics of the sequence of items which has been seen, the most probable correct answer can be selected. Where these conditional probabilities are not known exactly, they are estimated by a primitive "learning" process. Decisions are based on a posteriori probabilities and on other criteria studied in decision theory. The model is described in precise detail, and some possible goals of further work are mentioned.

1. INTRODUCTION

Many attempts have been made to recognize patterns by machine, especially such patterns as: geometric figures, machine-printed alphanumeric characters, hand-written letters, speech, and enemy fire-power in reconnaissance photos. Seldom has a study of one of these sets of items provided results which were applicable to any of the others, and progress in each case has depended largely on the researcher's familiarity with the particular set chosen. Yet clearly these problems are worthy of study and clearly there are many similarities among them. In this work we shall identify and concentrate upon the elements of similarity, attempting to construct a formal (almost axiomatic) system which applies to all the above patterns. This process of abstraction, though common in mathematics, has never, as far as we know, been applied to pattern recognition. In applying it, we gain simplicity, generality, and exactness. We lose, initially, the insight that familiarity with one particular set of patterns may bring; but when our model is completed, it should apply directly to each such set and should provide a precise framework with which such insights can be fruitfully combined.

What elements, then, are common to the various pattern recognition situations? We shall identify them as follows: there is a set of items which the recognizer observes, one at a time; there is a set of possible responses the recognizer can make; and there is later information as to which response was correct in connection with each item observed. Usually the recognizer's goal is to identify correctly as many items as possible, but generality is increased if the more complicated goals of decision theory (most of which depend on correct identification as a preliminary stage) are also allowed. In any case, when the goal has been defined, the recognizer's performance can (and must) be evaluated in terms of that goal.

Note that, in order to perform in a better-than-random way, a recognition system must initially be provided with some overall information about the universe of patterns. True, a human being can correctly identify an item which he has never seen before, but this abstraction process is possible only because of long experience with the "real" world and because of evolutionary adaptive mechanisms. To

duplicate this ability in our model we must provide conceptual analogues both of: 1) information about (or "experience with") individual items, and 2) initial information which is relevant to the environment of patterns. The information in (1) will be gained in the course of a recognition experiment, through the presentation of items and of the correct response associated with these items. Without information of this type a system possesses no flexibility or ability to learn. The information in (2) will be made explicit at the beginning of each experiment. Without information of this type a system possesses no ability to abstract or to respond to new situations. Together, these two types of information largely determine the performance of any system and the nature of any abstract model of recognition systems. With this in mind, we now present our model formally.

2. THE FORMAL MODEL

2.1 Notation and Organization

Let there be a finite set P of elements called patterns and denoted by $X_1, X_2, X_3, \dots, X_m$ or by $Y_1, Y_2, Y_3, \dots, Y_m$. In addition let there be a (possibly infinite) set I of inputs, denoted $\{I_n\}$. Let time be indexed by the variable t , so that $t = 1, 2, 3, \dots$. An experiment will consist of a sequence of inputs I^1, I^2, I^3, \dots , the superscript denoting the time of presentation. At each time $t = 1, 2, 3, \dots$, a pattern X^t from P may or may not also be presented. Let there be a system S which for each time, $t (= 1, 2, 3, \dots)$, takes I^t as input, takes X^t as input when it is available, and produces as output a pattern Y^t from P . Y^t may be a function of I^1, I^2, \dots, I^t and of $X^1, X^2, X^3, \dots, X^{t-1}$ (where present), as well as of any prior knowledge available to the system. It may not be a function of X^t or of I^u or X^u where $u > t$.

(Intuitively, we identify the sequence, I^1, I^2, \dots , with items seen and each X^t with the "correct" pattern which corresponds to I^t . Similarly, Y^t represents the system's guess upon seeing I^t . We may regard X^t as being presented "after" Y^t is produced, but within the same time-step. If Y^t were permitted to be a function of X^t , recognition would be trivial.)

Let the prior knowledge available to S be precisely specified as follows:

1) S has available a decision function, D, which specifies S's goal and can be used to measure its performance (e.g., the decision function may require that Y^t correspond to X^t as often as possible).

2) S is supplied with a finite list of characteristics, C_1, C_2, \dots, C_n . A characteristic is formally defined as an effectively calculable function from the set of all double sequences $(I^1, I^2, \dots, I^t; X^1, X^2, \dots, X^{t-1})$ into a finite subset of the integers. Less formally, we ask that after any input I^t is presented it is possible to determine from observed I's and X's what the value C_j^t of the characteristic C_j is. We denote the k^{th} value of C_j by c_{jk} .

(In the "pure pattern recognition" case, the characteristics will apply to I^t only, and will be "characteristics of I^t " in the ordinary sense. For another more complex example, consider the characteristic which has the value $i = 1, 2, \dots, m$ if and only if X^{t-1} was X_i . A machine using this characteristic alone would ignore the I's and would be a type of sequence predictor.)

3) For any or all values of any or all of the above characteristics, S may have available a list of probabilities. Suppose that a characteristic C_j has c_{jk} as one of its possible values. Then S will have available either no probabilities associated with $C_j = c_{jk}$, or it will have available a complete list of the probabilities, $P_{C_j=c_{jk}}^t(X^t=X_1), P_{C_j=c_{jk}}^t(X^t=X_2), \dots, P_{C_j=c_{jk}}^t(X^t=X_n)$.

These may be written as " $P_{c_{jk}}(X_1), \dots, P_{c_{jk}}(X_n)$ "; the whole list (or, alternatively, the i^{th} item) being denoted " $P_{c_{jk}}(X_i)$." Note that they are conditional probabilities.

(The characteristics are simply the factors which are relevant in deciding which patterns the various inputs represent. There are no theoretical limits on their number or complexity; however, the system will ignore any connections between inputs and patterns which are not available in the form of characteristics,

no matter how obvious they may be. When the list of probabilities, $P_{c_{jk}}(X_i)$, is supplied, S will use these exact values in making its decision and producing Y^t ; otherwise, it will estimate these same probabilities from the observed sequences.)

4) S also has available a set of functions, f_i , which express the legitimate a posteriori probability of each X_i in terms of probabilities like those above. Frequently the f_i will be trivial and obvious functions. An explanation of their importance in other cases is best postponed.

A more detailed discussion of the model and of the operation of S follows.

2.2 Patterns and Inputs

The word "pattern" has deliberately been left undefined, a "pattern" being merely an element of the set P. X^t and Y^t can be considered as names: X^t as the name associated with I^t ; Y^t as the name chosen at time t to optimize the decision function. (Alternatively, each X_i might be regarded as a set of inputs, so that each input I_h belongs to at least one pattern. However, this alternative concept is inelegant with respect to the X's and inapplicable to the Y's. The current formulation, which does not require any specific connection between I^t and X^t , is simpler and at the same time more general.)

The inputs, I^t , are usually the objects under study: in speech recognition, the speech sounds; in character recognition, the letters, etc. The following formal definition seems to apply to all such situations and will be adopted as a means of clarifying and limiting the model's scope:

DEFINITION: An n-dimensional input is a mapping from the Cartesian product, $\prod_{i=1}^n D_i$, of n ordered sets to an ordered set, R.

Thus a 0-dimensional input is a mapping with no domain, i.e., is nonexistent. A one-dimensional input is a mapping from D_1 to R and corresponds to a voltage waveform or acoustic signal (or speech sound) if we let D_1 be the real line representing time and R be the set of values representing voltage. A two-dimensional input is a

mapping from $D_1 \times D_2$ to R and corresponds to a printed letter or handwritten character (or photographic image) if we let D_1 and D_2 represent the horizontal and vertical directions, respectively, and we let R represent a set of gradations from white to black. A position in a tic-tac-toe game is a two-dimensional input where $D_1 = D_2 = \{1, 2, 3\}$ and $R = \{X, \text{blank}, O\}$. These examples should indicate how the definition is to be interpreted; examples of higher dimensions can, of course, be constructed.

For any one experiment a single dimensionality and a single domain ID_1 will presumably be selected and a subset of the possible inputs with this domain will be used as $\{I_h\}$.

The set I of inputs may be infinite; but at any one time, t , S will have observed only a finite number of them, namely, I^1, \dots, I^t . Likewise any single input may have an infinite domain (such as an interval or the real line), but S need not observe or store this infinite item directly. The only information S must handle is that which is used in computing the values of the characteristics. We demand that the characteristics be effectively calculable functions, i.e., that for any double sequence, $I^1, \dots, I^t; X^1, \dots, X^{t-1}$, S can compute the value of any characteristic in a finite number of steps. For each input, I^t , S then handles whatever information about I^t is needed to compute characteristic values at time t or at any later time. The amount of this information will be finite.

2.3 Initial Knowledge

We have specified that initially S will have available a decision function, a list of characteristics, and, perhaps, lists, $P_{c_{jk}}(X_i)$, for some characteristic values, c_{jk} . Discussion of decision functions and of the process of decision can be postponed, but lists of characteristics and probabilities merit discussion here. We shall call the probabilities, $P_{c_{jk}}(X_i)$, "post-probabilities." Their use as a decision statistic is justified in decision theory, where they are usually called a posteriori probabilities. In order to produce an output Y^t which optimizes its decision function, S must have available (except in trivial cases) estimates of the likelihood that X^t is X_1 , that X^t is X_2 , etc. The more relevant information that is used to

construct these estimates, the more accurate they will be, and the better the decision will be (on the average). The ideal likelihood estimate for X_i is thus the post-probability, $P_{Rt}(X_i)$, where R^t denotes "all relevant information which is available for decision at time t ;" but, as we have seen, all relevant available information is to be represented by the characteristics. Where we denote the value of C_j at time t by c_{jt} , we then have $P_{Rt}(X_i) = P_{c_{1t} \cdot c_{2t} \cdot \dots \cdot c_{nt}}(X_i)$, and we call the expression on the right, "the post-probability of X_i ."

Suppose first that the post-probability for each X_i is always available at any time t . Then it can be used directly as a likelihood estimate for X_i . However, we have specified that S has available lists of the form $P_{c_{jk}}(X_i)$ rather than of the form $P_{c_{1k} \cdot \dots \cdot c_{nk}}(X_i)$. To fit this special case into our general framework, it is only necessary to treat the combined characteristics C_1, \dots, C_n as a single one, C_w . Thus for each distinct combination of values which C_1, C_2 , etc., can have, we let C_w have a single distinct value. The likelihood estimate for X_i at time t is then $P_{c_{wt}}(X_i)$, the value c_{wt} being determined collectively by $c_{1t}, c_{2t}, \dots, c_{nt}$. Accordingly, the list of characteristics available to S reduces to one element, C_w , and the lists of post-probabilities will be of the form $P_{c_{wt}}(X_i)$ for all i and t .

Suppose next that the post-probabilities above are not directly available for all X_i and all possible values of C_1, C_2 , etc. They can then be estimated from the sequence of I's and X's presented to the system by a process which will be described in the next section. However, the estimation process may well require a very large number of time intervals to produce good estimates. Thus, whenever possible, it is desirable to compute the post-probability as a function of several, more readily estimatable, parts. A natural way of doing this is to separate the characteristics which are involved into groups and to express the post-probability as a (calculable) function of the post-probabilities of the groups. Thus, where the characteristics, C_1, \dots, C_n , are grouped to produce D_1, \dots, D_p , we will have:

$$P_{c_{1t} \cdot c_{2t} \cdot \dots \cdot c_{nt}}(X_i) = f_i \left[P_{d_{1t}}(X_i), P_{d_{2t}}(X_i), \dots, P_{d_{pt}}(X_i) \right]$$

(for $0 < p \leq n$, for all X_i .)

In the special case where $p = n$ and $D_j = C_j$, the probabilities $P_{C_j}(X_i)$ are clearly sufficient to compute the overall post-probability for X_i . In the other extreme case where $p = 1$, the post-probability is irreducible, f_i is the identity function, and there is effectively only one characteristic, as we have already seen. To make the intermediate cases fit into the framework also, we simply let D_1, D_2, \dots, D_p be the characteristics. There are p of these, each representing a combination of one or more C 's in exactly the same way that C_w represented n C 's in the irreducible case.

More needs to be said about the purpose and form of the functions f_i . If the post-probability of X_i were: 1) mathematically equivalent to some function of the separate $P_{C_j}(X_i)$, or 2) readily available, there would be no need for the f_i . Unfortunately things do not always work out this way, and we must have a way of computing post-probabilities from other statistics which are readily available. A conceptually simple and convenient set for these purposes is the set $P_{C_j}(X_i)$ for each C_j . In some practical cases, the functions, f_i , can be chosen according to the experimenter's estimate of the dependence of the post-probability on the various $P_{C_j}(X_i)$. In others no particular basis for the choice of the f_i will exist. Then certain assumptions of homogeneity can be made and the functions which result from these assumptions can be chosen.

Let us illustrate the choice of the f_i by a detailed example. Suppose there are two patterns, P_1 and $P_2(X_1 \text{ and } X_2)$; three characteristics, C_1, C_2 , and C_3 ; and that C_1 and C_2 have the possible values 0 and 1 while C_3 has the possible values 0, 1, and 2. Suppose first that no estimates of f_1 and f_2 are known, so that assumptions of homogeneity must be made. Let us consider as a specific case the estimation of f_1 , and, in particular, the estimation of $P_{C_1=C_2=C_3=0}(X_1)$ from $P_{C_1=0}(X_1)$, $P_{C_2=0}(X_1)$ and $P_{C_3=0}(X_1)$. We note that when $C_1=0$, six separate situations can occur, i.e., C_2 may have either of two values and C_3 may have any of three. Lacking other information about f_1 , we then make the crucial assumption of homogeneity; we assume that each

of these six situations is equally possible. It follows that $1/6 P_{c_1=0}(X_1)$ is an estimate of $P_{c_1=c_2=c_3=0}(X_1)$. Similarly, when $c_2=0$, six separate situations can occur, and, when $c_3=0$, four separate situations can occur. Making a similar assumption of homogeneity in each case, we find two more estimates of $P_{c_1=c_2=c_3=0}(X_1)$, respectively:

$$1/6 P_{c_2=0}(X_1) \quad \text{and} \quad 1/4 P_{c_3=0}(X_1).$$

As our final estimate, we then take the average of these three:

$$1/3 [1/6 P_{c_1=0}(X_1) + 1/6 P_{c_2=0}(X_1) + 1/4 P_{c_3=0}(X_1)]$$

or

$$1/36 [2P_{c_1=0}(X_1) + 2P_{c_2=0}(X_1) + 3P_{c_3=0}(X_1)].$$

Every step used in deriving this estimate can be carried through in the same way for each of the eight combinations of values of the characteristics and for X_2 as well as X_1 . Thus, we estimate

$$P_{c_1=k_1 \cdot c_2=k_2 \cdot c_3=k_3}(X_i) \text{ by } 1/36 [2P_{c_1=k_1}(X_i) + 2P_{c_2=k_2}(X_i) + 3P_{c_3=k_3}(X_i)]$$

in this example, and we have

$$f_1(a,b,c) = f_2(a,b,c) = 1/36(2a + 2b + 3c).$$

This example may help to explain the homogeneity assumptions and the estimates which result. A similar explanation could be given for the general case, but to avoid excessive notation, we shall merely state the result. Where there are characteristics, C_1, C_2, \dots, C_n ; where the number of values taken by C_j is V_j ; and where the homogeneity assumptions for X_i are made, $P_{c_{1t} \cdot c_{2t} \cdot \dots \cdot c_{nt}}(X_i)$ is estimated by:

$$M \sum_{j=1}^n V_j P_{c_{jt}}(X_i)$$

and $f_i(a_1, a_2, \dots, a_n) = M \sum_{j=1}^n V_j a_j$, where M is the constant $\frac{1}{n \prod V_j}$.

Note that this estimation is performed by the experimenter, and is not part of the system's operation; S begins each experiment with f_1, f_2, \dots, f_m directly available to it, whether computed in this way or not.

2.4 Learning

In the previous section, we have shown how the post-probability for X_i can be obtained from the list, $P_{c_{jk}}(X_i)$. We have specified that the list, $P_{c_{jk}}(X_i)$ may be available to S at the beginning of its operation. However, if some of these lists are not available, they too must be estimated.

The estimation of $P_{c_{jk}}(X_i)$ for a particular characteristic value, c_{jk} , and pattern X_i , is a simple one. We merely note that

$$P_{c_{jk}}(X_i) = \frac{P(c_{jk} \cdot X_i)}{P(c_{jk})}$$

by definition of conditional probability. Then, interpreting these probabilities as relative frequencies, we consider all previously identified inputs in the experiment for which C_j had the value c_{jk} . The number of these which were identified as pattern P_i , when divided by the total number, serves as an estimate for $P_{c_{jk}}(X_i)$. Naturally this proportion may change in the course of an experiment. But, by the law of large numbers, it will tend to become a more and more exact estimate of the correct probability. Thus, in a rudimentary sense, S will gradually "learn" to perform correct identifications as time goes on.

To compute these estimates, S must have in storage some numbers which indicate how often each characteristic value, c_{jk} , has occurred in conjunction with each pattern, X_i . At time t in an experiment, consider the number of previous time-steps at which characteristic C_j had value c_{jk} and X_i was presented as the correct pattern. We denote this number by $\#(c_{jk} \cdot X_i)$. Let S store the list $\#(c_{jk} \cdot X_1), \#(c_{jk} \cdot X_2), \dots, \#(c_{jk} \cdot X_m)$ for each c_{jk} . Then at time t , $P_{c_{jk}}(X_i)$ can be estimated by:

$$\frac{\#(c_{jk} \cdot X_i)}{\sum_{r=1}^m \#(c_{jk} \cdot X_r)} .$$

If no X^t is presented, the same number will remain in storage for computations at time $t + 1$. However, if X^t is presented, the stored values must be changed to keep the estimates correct for time $t + 1$.

Suppose that $X^t = X_1$ and the characteristics at time t have the respective values, $c_{1t}, c_{2t}, \dots, c_{nt}$. Then S merely adds one to each of the stored values $\#(c_{1t} \cdot X_1), \#(c_{2t} \cdot X_1), \dots, \#(c_{nt} \cdot X_1)$, to keep the estimation correct. The same is true, of course, for X_2, X_3 , etc.

2.5 Decision

The decision processes involve, and need to involve, no more than the decision function, D , and the post-probabilities for each X_i . The former is available to S directly; the latter can be estimated for each time t , as we have seen.

The most common decision function (and the one used almost exclusively in earlier pattern recognition experiments) is that of "percent correct." To operate under this decision function, S need only choose the X_i which is most likely to be correct at each stage, i.e., that which has the highest post-probability.

A more general case is that in which the goal is to maximize expected value with respect to a table of values and costs. Formally such a table is a mapping from $\{X_i\} \times \{Y_s\} = PXP$ into the integers (or real numbers). Informally, it specifies how much value (positive or negative) S receives at time t when $X^t = X_i$ and $Y^t = Y_s$, for every possible i and s . If the value, $D(X_i, Y_s)$, for X_i and Y_s is 1, where $i=s$; and -1 otherwise, this reduces to the "percent correct" case.

The expected value of choosing Y_s is defined as

$$E(Y_s) = \sum_{i=1}^m P_{Rt}(X_i) D(X_i, Y_s),$$

where $P_{Rt}(X_i)$ is the post-probability of X_i as before. To maximize expected value at time t , S merely chooses the Y_s for which $E(Y_s)$ is highest.

Still other decision functions have been treated in the literature of decision theory and need not be discussed here.

3. DIAGRAM OF THE SYSTEM

The following is a list of the notation used in the diagram of the system in Fig. 1:

A. Prior Knowledge:

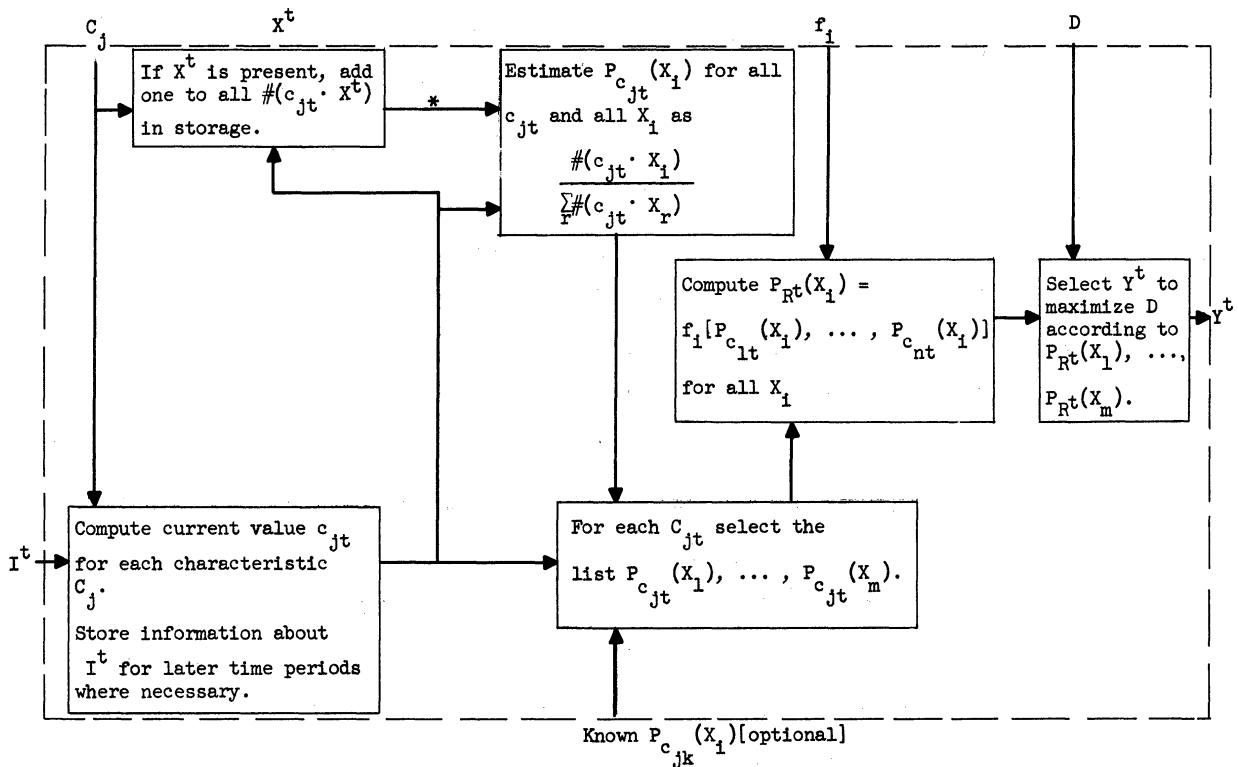
1. D: decision function
2. C_j : list of characteristics and their possible values.
3. Known $P_{c_{jk}}(X_i)$: for some characteristic values c_{jk} , the list $P_{c_{jk}}(X_1), \dots, P_{c_{jk}}(X_m)$. (May be omitted)
4. f_i : for each X_i , a function which estimates $P_{c_{1k}} \cdot \dots \cdot c_{nk}(X_i)$ from $P_{c_{1k}}(X_i), \dots, P_{c_{nk}}(X_i)$.

B. Knowledge Presented at Time t :

1. I^t : input at time t .
2. X^t : "correct" pattern for time t . (May be omitted)

C. Output at Time t :

1. Y^t : system's "guess" at time t .



*The $P_{c_{jt}}(X_i)$ in the box on the right are estimated before these values are transmitted.

Fig. 1. Diagram of System

4. CONCLUDING DISCUSSION

We have attempted to present a formalism of some of the elements and processes inherent in any pattern recognition situation. To do this, it has been necessary to abstract, and to omit discussion of the details of any specific application. At the same time, certain somewhat arbitrary choices have had to be made; for example, the "characteristics" discussed here provide only one of many ways of formalizing information and of interrelating initial knowledge to experience with individual items. We have made these choices in an attempt to: 1) limit generality as little as possible, 2) achieve simplicity of structure, and 3) permit ready application to particular situations. These three goals run through the work. Simplicity has obvious advantages and generality is sought, not for itself, but as a means to eventual widespread application.

The indispensable first step has been taken: a basic model has been made precise. At this point it is possible to ask many further questions in a meaningful way, and use the formal model in the search for answers. In conclusion, let us summarize some of the areas which may be investigated in the future.

1. Adaptation: Suppose the system S is considered as a part of a larger system, whose goal is to improve the performance of S over a series of experiments. How can this be done most effectively? Can successful adaptation be achieved by modifying the characteristics or by generating new ones? By systematic modification of the f_i functions? What further modes of learning can increase the power of S ?
2. Memory: A study of the memory capacity required by S for its various operations could be made. In particular, how much (on the average) will memory limitations impair successful recognition?
3. Computer Simulation: In what ways can the model be streamlined so as to permit more efficient computer simulation? Simulation is now possible, of course, but might well be postponed until it is of greater theoretical or practical value.
- 4) Theoretical Additions: A method of weighting the importance and the degree of independence of the characteristics might be

added. Where the $P_{c_{jk}}(X_i)$ are known as estimates (rather than being completely known or unknown), one might allow the "learning" mechanism to accept these estimates initially and to treat them as though they were based on an earlier sequence of inputs. Estimates might alternatively be supplied as a range of possible values, rather than as a single exact value.

UNIVERSITY OF MICHIGAN



3 9015 03695 5675