CR 85004

JUNE 1967
REPORT SM-48464-TPR-5

RESEARCH ON THE UTILIZATION OF PATTERN RECOGNITION TECHNIQUES TO IDENTIFY AND CLASSIFY OBJECTS IN VIDEO DATA

TECHNICAL PROGRESS REPORT NO. 5
COVERING PERIOD FROM
31 JANUARY THROUGH 31 MAY 1967

PREPARED UNDER CONTRACT NAS 12-30
FOR THE

NATIONAL AERONAUTICS & SPACE ADMINISTRATION
ELECTRONICS RESEARCH CENTER
CAMBRIDGE, MASSACHUSETTS

ASTROPOWER LABORATORY
2121 CAMPUS DRIVE • NEWPORT BEACH, CALIFORNIA

MISSILE & SPACE SYSTEMS DIVISION
DOUGLAS AIRCRAFT COMPANY, INC.
SANTA MONICA CALIFORNIA

DOUGLAS

Report SM-48464-TPR-5

RESEARCH ON THE UTILIZATION OF PATTERN RECOGNITION TECHNIQUES TO IDENTIFY AND CLASSIFY OBJECTS IN VIDEO DATA

Technical Progress Report No. 5
Covering Period From
31 January Through 31 May 1967

Prepared Under Contract NAS 12-30
for the
National Aeronautics and Space Administration
Electronics Research Center
Cambridge, Massachusetts

MISSILE & SPACE SYSTEMS DIVISION
ASTROPOWER LABORATORY
Douglas Aircraft Company
Newport Beach, California

FOREWORD

This report was prepared by Astropower Laboratory of Douglas Aircraft Company, Newport Beach, California, in fulfillment of NASA Contract NAS 12-30. The work was sponsored by the NASA Electronics Research Center, Cambridge, Massachusetts. Mr. Eugene M. Darling, Jr. was the Technical Officer for this office.

The studies began on June 18, 1965. This document is the fifth technical report.

Work on the contract was performed by the Electronics Department of the Astropower Laboratory. Dr. R. D. Joseph is the Principal Investigator. Contributors were G. E. Axtelle, D. B. Drane, C. C. Kesler, W. B. Martin, J. N. Medick, D. J. Nikodym, A. G. Ostensoe and S. S. Viglione.

SUMMARY

This report covers Items 4, 7, and 8 under contract NAS 12-30. These items are concerned with augmenting known properties with statistically derived properties, and with studies directed toward improving the design of decision mechanisms and statistically derived property filters.

The augmentation of known properties with statistically derived properties produced substantial performance gains in almost half of the cases studied. At least small improvements were registered in all but one case. The smallest gains were achieved on tasks where performance with the known properties alone was already very high.

Tests on both new and modified decision algorithms revealed that, although computational aspects were improved, there were no significant gains in performance. However, one technique (only partially tested) gives initial promise of real improvement in performance.

A method for designing property filters which use deterministic selection of subspaces as well as nonparametric distribution estimation to determine the switching surfaces was tested on two tasks. This approach yielded results which sometimes exceeded and other times fell short of previous performance.

TABLE OF CONTENTS

1.0	INTRODUCTION	1
2.0	AUGMENTATION OF KNOWN PROPERTIES	2
	2.1 Introduction	2
	2.2 Methodology	2
	2.3 Results and Conclusions	6
3.0	DECISION SYSTEMS	10
	3.1 Introduction	10
	3.2 Distribution Estimation	10
	3.3 Piecewise-Linear	11
	3.4 Results and Conclusions	13
4.0	PROPERTY FILTERS	17
	4.1 Introduction	17
	4.2 Subspace Selection by Mutual Information	18
	4.3 Distribution Estimation	20
	4.4 Optical Processing	24
	4.5 Results and Conclusions	27
5.0	PROGRAM FOR THE NEXT PERIOD	29
REFERENCES		30

LIST OF ILLUSTRATIONS

1	Pattern Recognition Mechanism	3
2	Performance on Task CVC (15 x 15 Aperture)	14
3	Optical Matched Filter	25

LIST OF TABLES

I	Known Properties	4
II	Generalization Performances	7
III	Generalization Performance With New Algorithms	15

1.0 INTRODUCTION

The purpose of this program is to examine the feasibility of using pattern recognition techniques for both extracting significant information from pictorial data, and for reducing the total amount of data that must be transmitted to convey this information.

Research during this period was accomplished in three study areas which were concerned with deriving more effective and more complete adaptive design techniques. One area, the design of decision mechanisms, has already been heavily researched. The other two areas represent virtually untouched problems.

Section 2 of this report describes the continuation of the studies on augmenting a set of known property filters with statistically derived properties. The statistical properties are designed to complement the known properties, rather than to provide an independent, parallel system.

Section 3 describes efforts to improve the design of decision mechanisms, given a set of property filters. These studies yielded improved algorithms, but not improved decision mechanisms.

Section 4 reports studies aimed at improving the statistical design of property filters. Two approaches were tried: (1) the deterministic selection of subspaces, and (2) the nonparametric design of nonlinear switching surfaces within the subspaces.

2.0 AUGMENTATION OF KNOWN PROPERTIES

2.1 Introduction

Figure 1 shows a design philosophy for recognition systems. "Signal Conditioning" provides both formatting of the input patterns and (where possible) invariance to irrelevant pattern differences such as gray-scale changes, translations, etc. The "Known Properties" are included to exploit the designer's knowledge of the recognition task. They measure pattern characteristics which the designer feels are useful for classifying the patterns. If the Known Properties are sufficient for the Decision Mechanism being used, then the design effort is readily completed. If they are not, however, more property filters must be obtained. One way in which the additional property filters may be designed is by the statistical analysis of sample patterns (i.e. Statistical Property extraction).

The augmentation of a set of Known Property filters with a set of Statistical Property filters specifically designed to complement the Known Properties is a virtually untouched problem. Technical Progress Report Number 4[1] reported results achieved by such augmentation when applied to cloud recognition tasks, namely the separation of noncumulus from cumulus clouds (Task NVC), and within the cumulus class, the separation of polygonal cells from solid cells (Task PVS). These results are repeated here for completeness (Table II). Six additional augmentations are also reported. Experiments were performed for three lunar terrain recognition tasks, namely the separation of: (1) craters from wrinkle ridges/rima (Task CVR), (2) wrinkle ridges from rima (Task RVR), and (3) craters with flat floors from craters with central peaks (Task CVC). Aperture sizes of 25 by 25 and 15 by 15 elements were used.

2.2 Methodology

A set of 29 Known Properties was designed by the NASA Technical Officer for this program, Mr. Eugene M. Darling, Jr. These properties, listed in Table I, were specifically designed for recognition of cloud patterns. Systems using these Known Properties achieved 90.0 and 97.0 percent generalization performances on Tasks NVC and PVS, respectively. Since these Known Properties alone achieved excellent performance, there is little room left for improvement by augmentation.

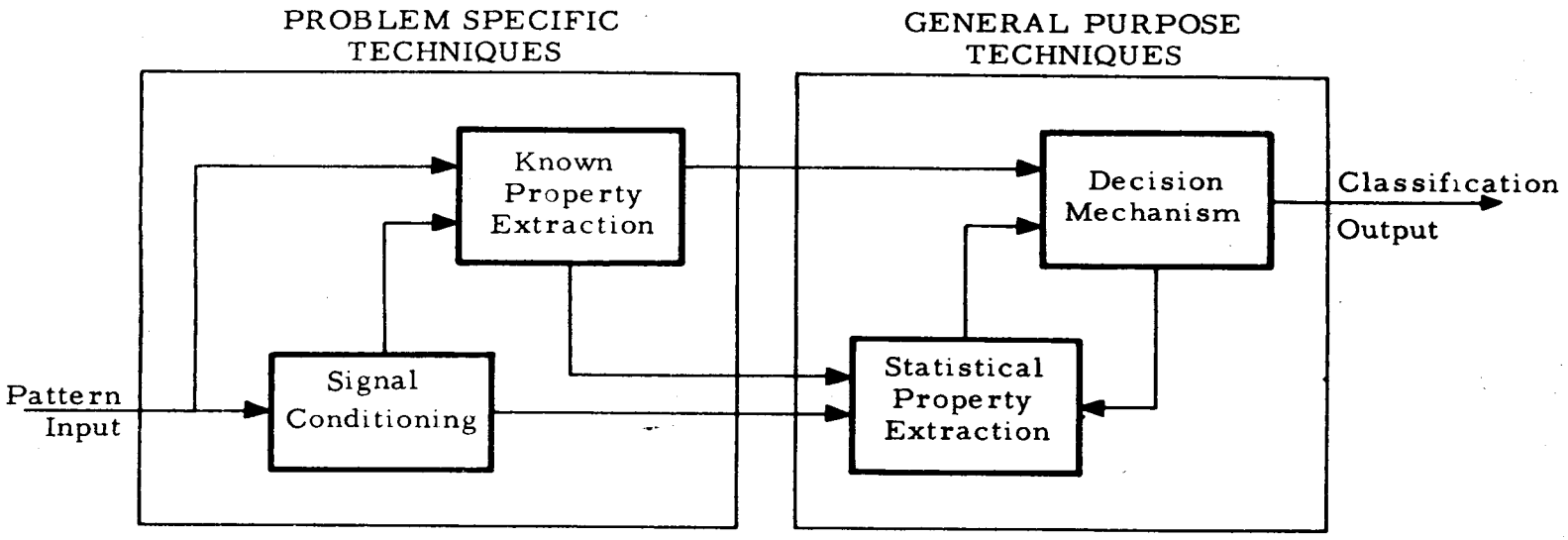

Figure 1. Pattern Recognition Mechanism

TABLE I

KNOWN PROPERTIES

1.	Mean Brightness		15.	Mean Cloud Segment	
2.	Brightness Variance		16.	Number of Clouds	
3.		0	17.	Mean Cloud Size	
4.		1	18.	Variance Cloud Size	
5.	Relative	2	19.		1-25
6.	Frequency	3	20.		26-100
7.	of Each	4	21.		101-225
8.	Gray Level	5	22.	Relative	226-400
9.		6	23.	Frequency	401-900
10.		7	24.	of Cloud	901-1600
11.	Information in the X-Direction (adjacent points)		25.	Size	1601-2500
12.	Information in the Y-Direction		26.	(i.e. number of elements	2501-3600
13.	Mean Gray Level Area (connected regions of constant gray level)		27.	in a 75 x 75 element	3601-4900
14.	Variance, Gray Level Area		28.	aperture)	4901-5625
			29.	.80 contour area (auto-correlation function)	

Fifteen of these 29 Known Properties relate to general characteristics of a two dimensional brightness field. The remaining properties are specifically tailored to the cloud recognition problem. The general properties are numbers 1 through 14, and 29. These 15 properties were used in the aforementioned lunar terrain recognition experiments. It was anticipated that the use of a small set of Known Properties, none of which was designed for the lunar problem, would result in lower performance levels more amenable to improvement by augmentation than was the case with the previously described cloud experiments.

The method of augmentation was as follows. A linear decision mechanism for the Known Properties was first designed. For this purpose the Iterative Design algorithm (modified to process the continuous outputs of the Known Property filters) was used. This approach was used for three reasons: (1) design times are short; (2) the designs are optimized in terms of the pattern losses used in the rest of the process; and (3) compensation for variation of means and variances between property filters is automatic.*

After this design was complete, each sample pattern in the training set was assigned a loss value. If D_i was the value of the linear decision function for the i-th pattern, and δ_i plus or minus one, depending on the true classification of that pattern, the loss for that pattern was given by

$$L_i = \exp\{-\delta_i D_i\}$$

This portion of the system design was held fixed for the rest of the process.

The losses for the training samples were used as initial values for the sample pattern losses in the QSID program. This program selects statistically designed property filters to minimize the system loss (sum of the losses of all training patterns). Thus the Statistical Property filters are selected to complement the Known Property filters.

---

*In some other experiments using continuous data an Error Correction process was used. The data had to be normalized prior to the application of the algorithm. If this normalization had not been made, four million cycles through the patterns would have been necessary for the threshold to compensate for the means of the property filter values.

After the set of augmenting Statistical Property filters was selected, the decision mechanism generated by the QSID program was discarded. That portion of the decision mechanism was redesigned by applying the Iterative Design algorithm, again using initial values for losses of the training patterns. Performance on the sample patterns was determined by adding the decision functions for the Known and Statistical Property sets. Again, the use of the Iterative Design algorithm made the equal weighting of these two decision functions automatic.

2.3 Results and Conclusions

The process for augmenting Known Properties was applied to eight recognition tasks. Table II presents the results obtained both with the Known Properties alone, and with Known Properties augmented by Statistical Properties.

Although results for networks using Statistical Properties only (i. e., no Known Properties) were available for most of these recognition tasks, these results were not included in Table II. A comparison between the performances of a Statistical Property system and an augmented Known Property system does not evaluate the augmentation process. The computer program which derived the augmenting properties is the same one that derived the original statistical properties. The only difference consisted of providing initial values for the pattern losses to insure that the augmenting statistical properties would not try to duplicate the efforts of the Known Properties. Thus, if the performance of the augmented systems is higher (or equal to) that of the Statistical Property systems, one concludes that the Known Properties are (or are not) measuring significant pattern features not accessible to the quadratic Statistical Property filters. Thus, this comparison evaluates the utility of the Known Properties in the augmented system. As the purpose of these experiments was to evaluate the augmentation process, statistical property systems should not be compared to augmented systems.

In assessing the performance differences which are shown in Table II, two factors should be given consideration: (1) an objective evaluation that such a difference might occur by chance, due to the random selection of the sample test patterns, and (2) a subjective evaluation of whether or not the difference matters. To aid in this latter consideration, the column "Percentage

TABLE II

GENERALIZATION PERFORMANCES

Task	Aperture	Known Properties	Known Properties plus Augmenting Statistical Properties	Percentage Decrease in Error Rate	Probability of Chance Occurrence
NVC	75 x 75	90.00	92.75	27.5	.083
PVS	75 x 75	97.00	94.25	-91.7	.029
CVR	25 x 25	92.00	97.00	62.5	.0009
	15 x 15	81.50	96.00	78.4	$<10^{-10}$
RVR	25 x 25	78.50	80.00	7.0	.300
	15 x 15	81.00	82.75	9.2	.260
CVC	25 x 25	71.50	85.75	50.0	.00003
	15 x 15	96.00	96.25	6.2	.426

Decrease in Error Rate" is given in Table II. The column headed "Probability of Chance Occurrence" provides an indication of the probability that the difference occurred by chance. The method used in its computation is given below.

It is desired to test the hypothesis that the systems (Known Property system and augmented Known Property system) are equal against the (one sided) alternative that the augmented system is better. Assume that the pattern set (consisting of 400 patterns) used to test one system was selected independently of the pattern set used to test the other system, and that each system has the same error rate, r. Let r_K and r_A be the observed error rates for the Known Property system and the augmented system, respectively. Then the variance of $r_K - r_A$ is $r(1 - r)/200$. Under the hypothesis,

$$\varphi = \frac{\sqrt{200}\,(r_K - r_A)}{\sqrt{r(1 - r)}}$$

will have zero mean and unit variance; and will be very nearly normally distributed. To obtain the "Significance Level," r was estimated by $\frac{1}{2}(r_K + r_A)$, and a table of the cumulative normal distribution was consulted to determine the probability that a difference greater than or equal to observed difference would occur by chance. The normal approximation should be very accurate for moderate probabilities, and indicative of very small probabilities for the extreme values. For task PvS, the one case in which r_K was less than r_A, the "Probability of Chance Occurrence" is the probability that a difference greater than the observed difference $r_A - r_K$ would occur by chance.

This analysis is conservative. The test patterns actually used were the same for each system. One would therefore actually expect a smaller performance difference than with two independent sets of test patterns. To see the extent of this, suppose that associated with the i-th test pattern was a parameter, p_i. Let p_i be the probability that each system misclassifies the i-th pattern (i.e., the systems make independent classifications, each having a probability of $1 - p_i$ of being correct). Let p_i have a Beta distribution with parameters $\frac{r}{1 - r}$ and 1, over the possible set of test patterns. Then the expected error rate is r, and the variance of $r_K - r_A$ is $r(1 - r)^2/(200(2 - r))$.

For task NVC, this analysis would give a "Probability" of .022, compared to the value of .083 given by the conservative analysis. The choice of parameters for the Beta distribution is arbitrary, and critically affects the results. For example, parameters 1 and $\frac{1-r}{r}$ also give an expected error rate of r, but yield .00003 as the Probability of Chance Occurrence.

Summarizing the results of the augmentation studies, eight tests were made. On task PvS a significant performance decrease was encountered, and on task CVC at the 15 by 15 aperture an insignificant performance change occurred. These failures are not disappointing, as performances with the Known Properties alone are 97 and 96 percent respectively. On task CVR at both apertures, and on task CVC at the 25 by 25, very substantial reductions in the error rates accompanied the augmentation, and the likelihood that these improvements occurred by chance is quite remote. Task NVC provides a borderline case. The reduction in the error rate is only one-third to one-half as large as those achieved in the better cases, and there is one chance in twelve that the improvement might occur by chance. Task RvR at both apertures provided the most disappoining results. The reductions in the error rates were relatively small, and quite likely to have occurred by chance.

In at least half of the test cases on which augmentation would be important to a designer, substantial performance increases were achieved. Thus, the augmentation process appears to be a valuable one, particularly when it is most appropriate — when the Known Properties alone do not produce very high performance levels.

3.0 DECISION SYSTEMS

3.1 Introduction

Earlier results obtained by using a variety of algorithms to design decision networks for a given set of property detectors did not show a clear superiority for any technique. In particular, the recursive techniques appeared to be equally effective.

MADALINE was the only technique tested earlier which yielded a nonlinear decision on the property profiles. Consequently, it has an inherently greater capability than the other techniques, gained at the expense of greater system complexity. Its failure to exploit this added capacity might be attributed to the fact that the particular property filters used were designed to achieve linear separability of the training patterns. Two new nonlinear decision techniques were included in this study, and provide a partial confirmation of this hypothesis. They also failed to outperform the linear techniques when applied to the same sets of property filters.

The two new techniques are attractive in their own right, and their inclusion in this study appears worthwhile. The piecewise-linear technique has been favorably reported in the literature.[2] The Distribution Estimation method provided good results as reported in Section 4 of this report.

The error Correction and MADALINE techniques required relatively long design times. Furthermore, these two designs exhibited strong oscillations in performance as a function of the number of training cycles. This behavior made it difficult to select a proper stopping point for the design process. In an attempt to overcome these difficulties one modified Error Correction algorithm, and two modified MADALINE algorithms were tested.

3.2 Distribution Estimation

A distribution estimation technique was applied to the QSID and DAID property profiles derived earlier for sample patterns. This method is nonparametric, and is intended to recover from sample vectors an estimate of the (continuous) underlying distribution. Because this technique was intended for continuous distributions, it is perhaps not at its best when applied to the binary property filter outputs. The program was written for use in

designing property filters (Section 4), and was limited to the processing of 35-dimensional distributions. In this study it was applied to the first 5, 25, or 35 property filters generated by the QSID or DAID program. Most experiments were performed on task CVC (15 x 15 aperture) where the total property filter set consisted of only 65 units. The small set processed by distribution estimation thus represented a larger fraction of the total than would have been the case for any other task.

Let there be M (training) sample vectors, each being n-dimensional. Denote the j-th coordinate of the i-th sample by y_{ij}. Let the j-th coordinate of a test pattern be denoted by x_j. The likelihood estimate for the test vector Y is given by:

$$f(X) = \frac{1}{M} \left\{ \prod_{k=1}^{n} \sqrt{b_k/\pi} \right\} \left\{ \sum_{i=1}^{M} \exp\left(-\sum_{j=1}^{n} b_j (x_j - y_{ij})^2 \right) \right\}$$

A more complete description of this estimate, and the means for estimating the parameters b_j is given in Section 4.3.

A decision is derived for a test pattern by estimating the likelihood functions (i.e., density estimates for the test pattern vector) for each class, using the sample patterns and b_j values appropriate to that class, and then selecting the largest likelihood value. The resulting decision surface is highly nonlinear.

The technique was applied to task CVC (15 x 15 aperture) for 5, 25, and 35 property filters, to task RvR (15 x 15 aperture) for 25 property filters, and to task NVC for 35 property filters.

3.3 Piecewise-Linear

In the piecewise-linear algorithm used in this study, an even number of response units was selected. Half of the response units were assigned to each class. A decision is achieved by determining the response unit with the highest input sum. The decision surface is piecewise-linear, and the response unit input weight vector hopefully represents a cluster point of the training patterns of the appropriate class.

The algorithm is as follows. If a correct decision on a training pattern is obtained, no changes are made in the response units' input weights. If an incorrect decision is reached, the response unit for the correct class with the largest input sum is determined. For this response unit, each input weight from an active property filter is incremented by $\frac{\delta_i}{m}$, where m is the total number of active property filters. On odd-numbered cycles (through the training patterns) only, the response unit which caused the incorrect decision is also modified, the connections from active property filters being incremented by $-\frac{\delta_i}{m}$. δ_i here is 1 for a positive class training pattern, and -1 for a negative class training pattern.

For training purposes, the weights were allowed to vary as freely as with the earlier Error Correction and MADALINE systems, but these weights are not considered to be the actual system weights. The actual system weights are taken to be the average values of these freely varying weights over the last full cycle through the training patterns. This change, which was also used on the modified Error Correction and one modified MADALINE,* had the desired effect. Performance did not oscillate strongly with training cycles, and the best performance point came much earlier in the design run. The technique was applied to the DAID property filters for tasks CVN, PVS and RVR (50 x 50 aperture), with 4, 6, and 8 response units.

Another modified MADALINE will be called Sequential MADALINE. This technique does not use the averaging process. The original MADALINE algorithm appeared to use a large initial segment of the design run in allocating portions of the recognition task to the various ADALINES. In sequential MADALINE, the algorithm is run with 1 ADALINE for N cycles through the training patterns. After each cycle, the performance level is compared with previous performance level, and if it is better, the machine state is recorded. After N cycles, the machine state corresponding to the best performance level is used as initial weights for one ADALINE of a three ADALINE system. After N cycles, the best three unit system is used as a starting state

*These two algorithms are called Averaging Error Correction and Averaging MADALINE, respectively.

for a five unit system, and so on. In the experimental work, N was taken to be 10. The purpose of this change is, of course, to address the first ADALINE to the largest part of the problem, and subsequent ADALINES to increasingly finer parts of the discrimination. This technique was applied to tasks CVN and PVS, with seven ADALINES as a limit.

3.4 Results and Conclusions

The distribution estimation program was limited to processing 35 property filters or less. For task CVC (15 x 15 aperture), the program was run for 5, 25, and 35 property filters. The resulting generalization error rates are plotted in Figure 2, together with results obtained with the Iterative Design technique for various numbers of property filters. The distribution estimation method appears to be better, but its performance with 35 property filters does not equal that of Iterative Design with the full set of 65 property filters. The program will be modified to process all 65 property filters, to determine whether or not the indicated advantage can be maintained.

Distribution estimation was also applied to tasks NVC and RVR (15 x 15 aperture) with 35 and 25 property filters respectively. Performances achieved were 76.50 and 77.00 percent respectively. The unmodified Error Correction algorithm applied to the same tasks with the same number of property filters yielded 73.75 and 79.25 percent respectively. Best performances obtained on these tasks were 86.25 and 84.50 percent using 400 and 235 property filters respectively. Thus the distribution estimation technique achieves performances which are about average on these tasks. As the technique is much more complex than the earlier ones, this is not encouraging. Further tests will be performed.

The results obtained with the Piecewise-Linear technique and the three modified algorithms are summarized in Table IV. For comparison purposes, the results achieved earlier with the unmodified three unit MADALINE are included. This MADALINE had been the best technique on task RVR; it and Iterative Design were best on task PVS; it was second to Iterative Design by a quarter of a percent on task NVC. On task NVC, all of the systems exhibit a small but consistent improvement over earlier results. Elsewhere, only the Sequential MADALINE registered any gains, and those were small.

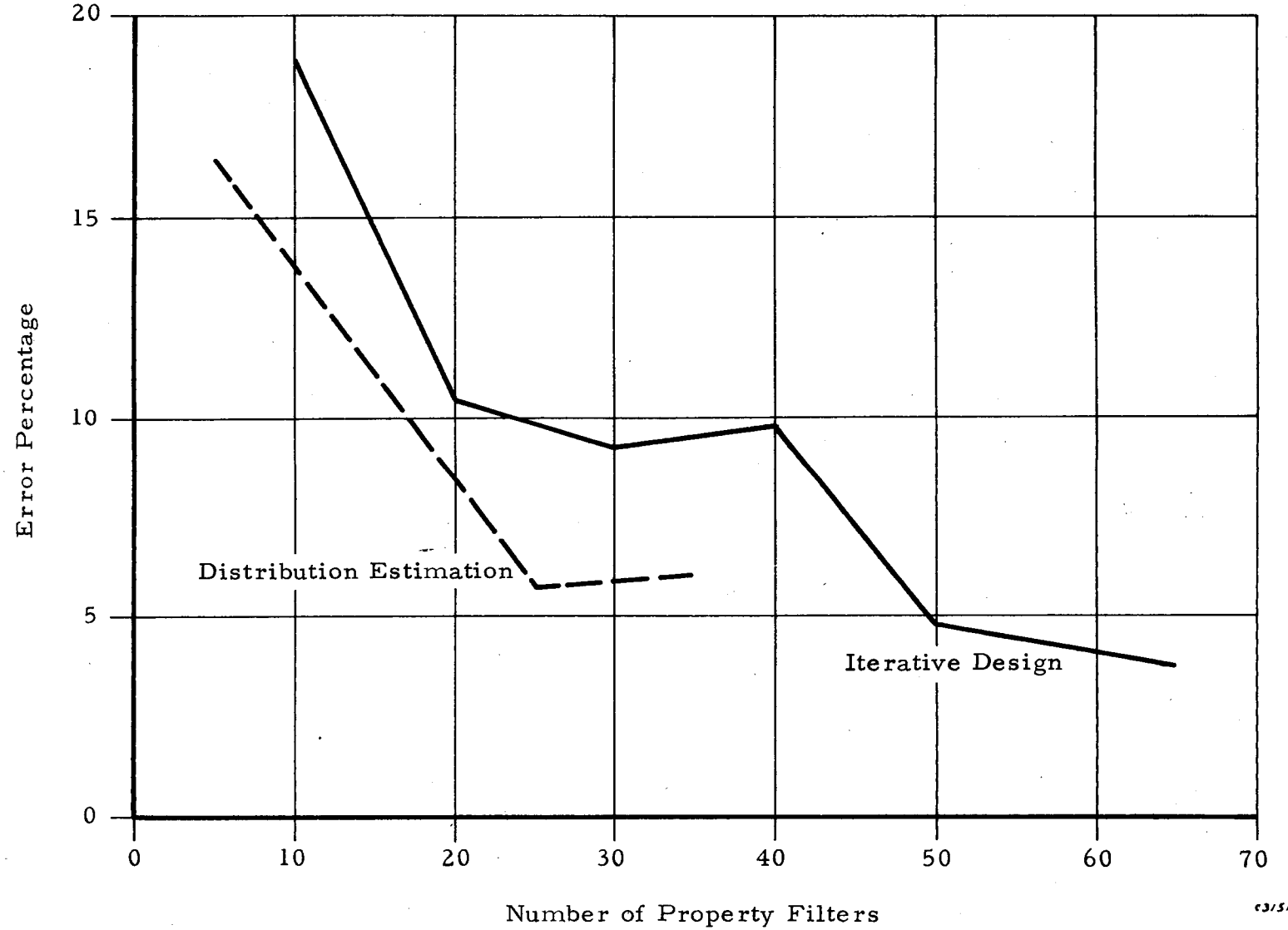

Figure 2. Performance on Task CVC (15 x 15 Aperture)

TABLE III

GENERALIZATION PERFORMANCE WITH NEW ALGORITHMS

Technique	Linear Decision Elements	Task CVN	Task PVS	Task RVR (50 x 50)
Averaging Error Correction	1	87.75	83.75	74.00
Averaging MADALINE	3	88.00	85.00	75.00
	5	88.00	85.00	73.75
	7	87.50	85.50	75.25
Sequential MADALINE	1	86.00	83.50	–
	3	87.75	85.75	–
	5	87.00	85.25	–
	7	87.25	86.50	–
Piecewise-Linear	4	87.75	84.00	73.75
	6	87.25	85.00	74.00
	8	87.50	85.75	73.50
Comparison: Unmodified MADALINE	3	86.00	85.50	75.75

In summary, a number of new and modified algorithms for the design of the decision functions were tested on the same property filter sets as were the earlier algorithms. The primary objective of these experiments was to improve performance levels, and in this they failed to yield the hoped for gains. Distribution Estimation, a difficult technique to implement, was incompletely tested due to computer program limitations (which are being remedied). This technique showed promising signs in one of the three tests performed. The remaining algorithms failed to yield any significant performance increases.

The modified algorithms did display some operational improvements. Design times were shortened. For example, the Sequential MADALINE design runs used 7 hours of SDS 930 time, compared with the 12 to 16 hours used by the earlier MADALINE, with no sacrifice in performance. The Averaging MADALINE and Error Correction did not display the large performance fluctuations during the design run which were characteristic of the unmodified algorithms. The smoother performance curve gives the designer greater confidence in selecting a stopping point for the design process.

4.0 PROPERTY FILTERS

4.1 Introduction

The results presented in Section 3 indicate that the decision mechanism algorithms are probably as effective as the set of Statistical Property filters permits. An improved means for designing the set of Statistical Properties thus seems desirable.

In the preceding work, the Statistical Property filters were designed using the DAID algorithm or its slight modification, the QSID algorithm. In these, the set of property filters is selected sequentially. Each time the existing set is to be augmented, a pool of candidate filters is generated. This is accomplished by randomly selecting a number of subspaces. For each subspace, a switching surface is defined by the vectors X satisfying the quadratic equation

$$0 = X'(S_1^{-1} - S_2^{-1})X - 2X'(S_1^{-1}M_1 - S_2^{-1}M_2) + M_1'S_1^{-1}M_1 - M_2'S_2M_2 + \ln\frac{|S_1|}{|S_2|}$$

M_1 and M_2 are the mean vectors, and S_1 and S_2 the covariance matrices of the two pattern class. If each pattern class had a Gaussian distribution in the subspace, and if M_1, M_2, S_1, and S_2 were the true parameters of these distributions, this equation would provide an optimum switching surface. When M_1, M_2, S_1, and S_2 are unknown, they are customarily estimated from sample patterns (as in the DAID and QSID algorithms). This quadratic analysis is widely used, even when it is known that the underlying distributions are not Gaussian (for example, see Reference 3). The random selection of subspaces and the assumption of Gaussian statistics can lead to property filters of low efficiency. To mitigate this, many more property filters are generated than are included in the property filter set. Selections are made on pragmatic grounds, based on how well the candidates augment the existing system in separating the sample patterns.

A new technique for generating Statistical Property filters was tested. This method avoids the two least desirable aspects of the earlier techniques — the random selection of subspaces, and the assumption of Gaussian statistics. Selection of coordinates for a subspace was accomplished sequentially, using a mutual information value as a selection criterion. Within

a subspace, a switching surface was defined using a nonparametric distribution estimation technique.

Tests were performed on the cloud pattern tasks. The textural nature of these patterns permitted a number of subspaces to be formed by translations of a subspace generated by the mutual information process. Sets of 9 and 25 subspaces, translations of a single select subspace, were used. The decision mechanism used in these systems was quite crude — consisting of an unweighted majority vote of the property filters.

4.2 Subspace Selection by Mutual Information

A program was developed to sequentially select the coordinates of a subspace. Mutual information values were used to control the selection process. To establish the mutual information, two partitions of the sample patterns in the training set were required. Let $p(i, j)$ denote the fraction of patterns in the i-th cell of the first partition and the j-th cell of the second partition, $p_1(i)$ denote the fraction in the i-th cell of the first partition, and $p_2(j)$ denote the fraction of patterns in the j-th cell of the second partition. Then the mutual information is given by:

$$\sum_i \sum_j p(i, j) \log_2 \frac{p(i, j)}{p_1(i) p_2(j)}$$

One of these partitions was provided by the actual classification of the sample patterns. The other partition chosen was an attempt to reflect the suitability of the subspace already selected, if augmented by a candidate coordinate. In this application, the value of the mutual information was between "zero" and "one." If $p(i, j) = p_1(i)\, p_2(j)$ for all i and j, that is, the actual pattern classifications were statistically independent of the second partition, the value would be "zero." If $p(i/j)$ were always "zero" or "one," that is, the second partition gives complete information about the actual pattern classification, the value of the mutual information is given by $-\Sigma\, p_1(i) \log_2 p_1(i)$. Since two classes containing equal numbers of sample patterns were used, this value is "one."

The selection of the second partition was difficult, and the final choice represented several compromises with computational reasonableness.

Ideally, a property filter would be designed using distribution estimation for each subspace formed by augmenting the subspace already chosen by a candidate coordinate. The decisions of these property filters would then form the second partitions. To avoid the excessive computer time required to accomplish this, it was decided to design one property filter on the subspace already selected, and to refine the property filter's partition using the discrete gray scales of the candidate coordinates. As a second compromise with memory size, only the most significant two bits of the three bit gray scale were used. Thus the second partition has eight cells, formed from the one bit decision of the logic unit and the two bit gray scale of the candidate coordinate. As a final simplification, and perhaps the most weakening compromise of all, quadratic property filters were used to minimize computer time. The derivation of the quadratic unit is described in Section 4.1.

Coordinates for the subspace are selected from a 25 by 25 subaperture of the 75 by 75 NIMBUS patterns. It was felt that the subaperture would be sufficiently large to contain the significant pattern features. The NIMBUS patterns were divided into nine subapertures. Due to the textural form of these cloud structures, the patterns within the nine subapertures should be of the same nature. All nine subapertures were used, so that the sets of training patterns contained 9000 examples of each class.

Once a subspace was selected, it was expanded to a set of 25 subspaces by translation. Consider the subspace to be in the central subaperture. Horizontal and vertical translations of 0, plus and minus 10, and plus and minus 25 raster elements were used. The subspace positioned within each of the original nine subapertures is included in the resulting set of 25 subspaces. A set of nine subspaces was also formed by using only the 0 and plus and minus 25 translations (corresponding to the original subapertures).

Results on task NVC were sufficiently poor with the nine subspace case, that the 25 subspace case was not tested. As a control for task PVS, a random subspace was selected in the 25 by 25 subaperture, and a set of 25 translations formed as above. The performance achieved by applying distribution estimation to this random subspace was virtually identical to the performance of the same technique on the mutual information subspace, the

error rates being 9.75 and 9.25 percent, respectively. Thus these initial tests have not demonstrated that these mutual information subspaces offer any advantage over randomly selected subspaces.

4.3 Distribution Estimation

A nonparametric distribution estimation technique was used to establish the switching surfaces for the subspaces. In particular, the binary property filter is defined as on for a test pattern if the density estimate (likelihood function) for the positive class is larger than that for the negative class at the test pattern vector. The problem of accurately recovering the unknown underlying distribution (or density) associated with a collection of samples has received considerable attention. E. Parzen[4] has treated a general class of consistent estimators for the one-dimensional case. Most of his results have been extended to the multi-dimensional case by V. K. Murthy.[5] The mathematical results can be described very briefly. Let $\{x^k\}_{k=1}^{M}$ be a set of M identically distributed N-dimensional random vectors. The empirical distribution function F_M is defined by the expression

$$F_M(x_1, x_2, \ldots, x_N) \equiv \frac{1}{M} \left\{ \begin{array}{l} \text{number of observations } x^k \\ \text{such that } x_j^k \leq x_j,\ j=1, 2, \ldots, N \end{array} \right\} \qquad (1)$$

An estimator f_M for the N-variate density f is given by

$$f_M(x) \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{n=1}^{N} b_n \cdot H\{b_1(x_1 - y_1), \ldots, b_N(x_N - y_N)\}\, dF_M(y_1, \ldots y_N) \qquad (2)$$

The function H satisfies the conditions

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(x_1, x_2, \ldots x_N)\, dx_1 \ldots dx_N = 1,$$

$$H(x_1, x_2, \ldots x_N) = H(\pm x_1, \pm x_2, \ldots, \pm x_N) \geq 0,$$

$$\lim_{\|x\| \to \infty} \{\|X\|\, H(x_1, x_2, \ldots, x_N)\} = 0, \text{ where } \|X\| \equiv \left\{ \sum_{n=1}^{N} x_n^2 \right\}^{1/2}$$

and $b_n > 0$ for $n = 1, 2, \ldots, N$. For the applications of interest it is also assumed that the density f of the underlying distribution F is everywhere continuous.

It can easily be shown that Equation (2) is equivalent to

$$f_M(x) = \frac{1}{M}\left\{\prod_{n=1}^{N} b_n\right\} \sum_{k=1}^{M} H\{b_i(x_i - x_1^k), \ldots, b_N(x_N - x_N^k)\} \qquad (3)$$

V. K. Murthy, in extending E. Parzen's results for the one-dimensional case, has shown that if the b_n are functions of the sample size M and satisfy the conditions

(a) $\lim_{M\to\infty} b_n = \lim_{M\to\infty} b_n(M) = \infty$ for $n = 1, 2, \ldots, N$

(b) $\lim_{M\to\infty} \frac{M}{\prod_{n=1}^{N} b_n} = \infty$

then f_M is a consistent estimate of f at every point **X**. That is, if conditions (a) and (b) are satisfied then at every point **X**

$$\lim_{M\to\infty} E\{f_M(x)\} = f(x)$$

and

$$\lim_{M\to\infty} \text{var}\{f_M(x)\} = 0$$

The specific case under investigation involves

$$H(x) = \frac{1}{\pi^{N/2}} \prod_{n=1}^{N} \exp\{-x_n^2\}$$

so that Equation (3) becomes

$$f_M(x) = \frac{1}{M\pi^{N/2}}\left\{\prod_{n=1}^{N} b_n^{1/2}\right\} \sum_{k=1}^{M} \prod_{n=1}^{N} \exp\left\{-b_n(x_n - x_n^k)^2\right\} \qquad (4)$$

The notation used in Equation (4) is slightly different from that used in Reference 5 and entails substituting $\sqrt{b_n}$ for b_n in Equations (2), (3) and condition (b). Since in every practical case one deals with a finite number of samples, M, the problem is to obtain for a fixed M an estimate of the values for b_n. It is possible to introduce a property of the unknown underlying distribution (via the sample set $\{x^k\}_{k=1}^{M}$) by investigating an expression of the form

$$\frac{1}{b_n} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ijn} \exp(-\rho_n a_{ijn}) \tag{5}$$

where

$$\rho_n > 0 \text{ for } n = 1, 2, \ldots, N$$

$$\lim_{M \to \infty} \rho_n = \infty$$

and

$$a_{ijn} \equiv (x_n^i - x_n^j)^2, \quad \text{i and j refer to the i-th and j-th sample vector.}$$

The problem of determining the b_n is thus traded for the problem of determining ρ_n. It can be shown that in order for b_n to satisfy the consistency conditions (a) and (b), ρ_n is of the order of log M (i.e., $\rho_n = 0(\log M)$). In obtaining this result the inequality

$$\frac{M}{\prod_{n=1}^{N} b_n^{1/2}} \geq M \prod_{n=1}^{N} \left[\exp\{-\rho_n \mu_n(M) + \log \mu_n(M) \right]^{1/2}$$

is derived, where

$$\mu_n(M) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ijn}.$$

Choosing

$$\rho_n = \frac{\delta_n}{\mu_n(M)} \left[\log M + \log \{\mu_n(M)\}^{1/\delta_n} \right]$$

yields

$$\frac{M}{\prod_{n=1}^{N} b_n^{1/2}} \geq M \prod_{n=1}^{N} \exp\{-\tfrac{1}{2}\delta_n \log M\} = M^{1-\frac{1}{2}\sum_{n=1}^{N}\delta_n}$$

so that the consistency conditions are satisfied with ρ_n as given above if

$$\sum_{n=1}^{N} \delta_n < 2 .$$

Furthermore, it can easily be shown that if ρ_n is of the form

$$\rho_n = g_n M^{\beta_n} + d_m$$

where $\beta_n > 0$ and g_n, d_n are constants such that $\rho_n > 0$ for $M \geq 1$, then the consistency conditions are violated. Thus, in calculating the values of b_n the expression

$$\rho_n = \frac{\delta_n}{\mu_n(M)} \log (M \cdot \mu_n(M)^{\frac{1}{\delta_n}})$$

is used. The behavior of ρ_n (and hence of b_n) is influenced by the quantities $\mu_n(M)$ which yield an indication of the "spread" of the unknown distribution along each of the N axes. Thus for a fixed sample size M a distribution with large second central moments requires smaller values for b_n (and hence smaller values of ρ_n) than a distribution possessing small second central moments. This behavior is exhibited by ρ_n as given above.

Distribution estimates were made separately for each subspace. Thus, although the coordinate configurations of the subspaces in a system are simple translations of one another, each subspace has its own unique switching surface.

Attempts were made to simplify the distribution estimates by finding the local maxima of the density functions. It was hoped that the number of sample points used in the distribution estimation could be reduced

by replacing those clustered around a mode by the modal point itself. Even with variations in clustering program parameters, only two modes could be found for the Polygonal Cell distribution, which would be insufficient to provide a real simplification.

It was hoped that 25 to 100 cluster points would be found, with nearly all of the sample points belonging to some cluster. Having found such a set of cluster points, one may proceed to simplify the system in one of two ways. One may use only the cluster points to estimate the distributions, or one may use the cluster points plus the sample points not belonging to a cluster to estimate the distribution. When 1000 samples are being used to estimate the distribution, and when one finds two cluster points which represent perhaps a dozen sample points, the second alternative offers a one percent reduction in system complexity. This hardly seems worth the effort. The use of the first alternative with two cluster points implies that the distributions are simple and highly separable. A single quadratic property filter should do nearly as well, since four points are separable by a quadratic surface. Earlier experiments with DAID units indicate that this is not the case.

4.4 Optical Processing

Optical processing[6] offers some attractive features for pattern recognition. Great speed is possible due to the highly parallel nature of the processing. Processing in the frequency plane offers freedom from translational variations. In addition, the textural nature of the cloud patterns suggests that a frequency plane analysis might be fruitful.

Interesting potential uses for an optical system are as independent attention centering mechanisms for a recognition system, and as a property extraction device. Such devices would be based on matched filters placed in the frequency plane. The theory of such filters is well known.[6]

The two-dimensional Fourier transform of an input image is present in the frequency plane (Figure 3). By placing a suitably designed transparency or mask in the frequency plane, it is possible to operate on the Fourier transform of the input image with the results of the operation displayed in the output plane. The optimum optical matched filter is defined to

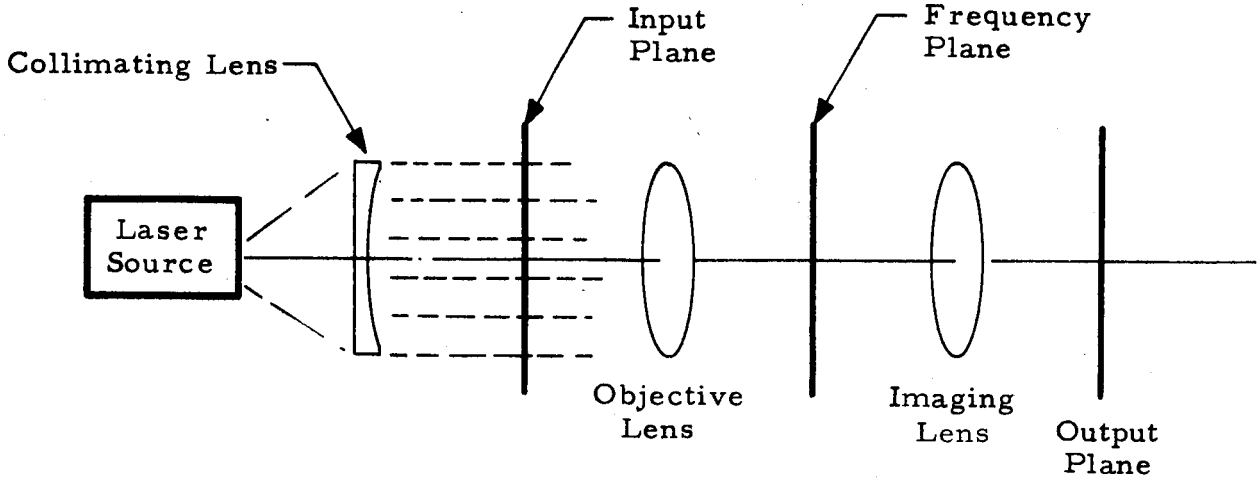

Figure 3. Optical Matched Filter

be one whose modulation transfer function is proportional to the complex conjugate of the spectrum $S(p,q)$ of an image of a pattern of interest divided by the background noise spectral density function $N(p,q)$ where

$$s(\xi,\eta) \equiv \text{recorded image of a pattern of interest against a noiseless background}$$

$$S(p,q) \equiv \int_{-\infty}^{\infty}\int s(\xi,\eta)\, e^{i(p\xi+q\eta)}\, d\xi\, d\eta$$

$$R(\tau_1,\tau_2) \equiv \text{autocorrelation function characterizing the noise}$$

$$N(p,q) \equiv \int_{-\infty}^{\infty}\int R(\tau_1,\tau_2)\, e^{i(p\tau_1+q\tau_2)}\, d\tau_1\, d\tau_2$$

$$\equiv F\left[R(\tau_1,\tau_2)\right];$$

the optimum filter mask is represented by

$$H(p,q) = \frac{KS^*(p,q)}{N(p,q)} .$$

The spectra of available patterns may be sampled at N points. Using the technique described in Section 4.3, estimates of the density functions $f_M(x_1,\ldots,x_N)$ of each pattern class may be obtained. The cluster points (or local maxima) of these distributions might be assumed to be "pure signals" associated with the given class of patterns, while neighboring points might be assumed to be the pure signal corrupted by additive noise.

Property filters for a two class problem might be designed by contrasting each cluster point of the density function of one class with each cluster point of the density function of the other class. One-class filters might be designed from the appropriate density function by contrasting each cluster point with the autocorrelation function of the "local noise" around that cluster point.

Matched filtering is a linear operation and, unless the patterns are nearly linearly separable in their frequency plane representations, one would not expect a single matched filter to give high levels of performance. Indeed, only limited success has been obtained by researchers, thus far, in using an optical matched filter as a recognition device. The use of a number

of matched filters (as property detectors) separating the clusters of the competing distributions would seem to offer more chance of success.

4.5 Results and Conclusions

The subspace selection algorithm was applied to task PVS. The first eight coordinates were selected with no difficulty. The coordinates which were selected were widely separated in the subaperture. For coordinates nine and ten, the program tried to select coordinates already chosen. When this happens, the algorithm chooses the second best candidate, and so on. With these alternate selections, coordinates nine and ten of the subspace followed the trend of widely separated points. For the eleventh and subsequent selections, the program tried to repeat the fourth coordinate, and on being denied this, chose an adjacent point. Performance of the quadratic property filter in classifying the training patterns varied smoothly, reaching a maximum with ten coordinates. At that point it made 71.65 percent correct decisions on the 18,000 subaperture patterns (nine subapertures from each of 2,000 training patterns).

The first ten coordinates selected were used as the subspace. Nine translations were taken and switching surfaces assigned using distribution estimation. Using a majority vote of the nine property filters to determine the system decision, 86.75 percent of the test patterns were correctly identified. This is a quarter of a percent better than the Sequential MADALINE accomplished with 400 DAID property filters, and at least one percent better than any other system obtained with statistically derived properties.

The same subspace was then used in 25 translations. The system yielded 90.75 percent correct decisions, a substantial increase. The property filters themselves were 73.59 percent correct in 10,000 generalization decisions (25 subapertures from each of 400 generalization patterns). Using the randomly generated subspaces, the 25 property filters were 73.17 percent correct, and the majority vote system achieved 90.25 percent. Either this was a fortunate choice of random subspace, or the subspace selection process did not add much to the system.

The subspace selection program was then run to choose a ten coordinate subspace for task NVC. Repetition of previous selections began

with the fifth coordinate and continued thereafter. The best quadratic property filter classification of the sample patterns occurred with five coordinates, and was 60.67 percent. The ten coordinate quadratic unit gave 59.71 percent. The subspace selected was a rather strange one. The first nine choices came from the left half of the top row of the 25 by 25 raster. The tenth coordinate was about halfway down the last column.

Nine translations were taken, and a distribution estimation system tested. It achieved 66.5 percent on the test patterns — 84 percent of the noncumulus patterns, 30 percent of the solid cell patterns, and 68 percent of the polygonal cell patterns were correctly identified. Contributing to this uneveness might be the fact that equal numbers of cumulus and noncumulus patterns were used to estimate the distributions. There were, therefore, twice as many noncumulus patterns as there were solid cell or polygonal cell patterns. Solid cell patterns appear more similar to noncumulus than to polygonal cell patterns.

Due to the poor overall results, the 25 translation experiment was not run. The unusual subspace may have contributed to these results. The subspace selection program was continued to 23 coordinates. Except for two coordinates near the center of the subaperture, the remaining coordinates formed a loose cluster near the lower right corner of the aperture. Best classification with a quadratic property was 67.86 percent with 17 coordinates, a figure much more in line with that achieved on task PVS. Future plans call for testing a subspace consisting of selections 8 through 17, and a random subspace.

Summarizing these results, a technique for property filter generation utilizing sequential subspace selection and nonparametric distribution estimation was tested. On task PVS a substantial improvement over previous results with statistical systems resulted; on task NVC an even more substantial performance decline was experienced. In both cases, the subspace selection process appears weak. On task PVS, the selected subspace proved to be only slightly better than a randomly selected subspace, and the odd subspace selected for task NVC more than likely was responsible for the poor results on that task.

5.0 PROGRAM FOR THE NEXT PERIOD

Primary emphasis in the final reporting period will be on Items 5 and 6 of the contract. These items are concerned with the definition of pattern area boundaries and the investigation of hardware feasibility. Some additional investigation of distribution estimation to design decision mechanisms will be performed. A final report will be drafted.

The determination of pattern area boundaries will be investigated on the cloud pattern tasks using existing network designs. Test patterns in which a portion of the aperture contains one pattern type, while the remainder of the aperture contains a second type of cloud pattern, will be synthesized. The proportion of each pattern type will be varied. System performance will be tested on these mixed patterns.

Preliminary design of a recognition system will be performed to establish hardware feasibility. The system design will contain an optical input system, as briefly outlined in TPR-3,[1] and a sequential digital implementation of 400 six input hyperquadratic first layer property filters followed by one or more linear threshold response units. The system will contain a memory for storing the results obtained from training on the general purpose computer (weights, thresholds, and coordinate of the data points which input to first layer property filters). Control for generating the property filters as well as accessing a new subsection will be discussed. Hybrid analog and digital techniques will be employed where weight and power savings result. Estimates of the size, weight, power and component cost of the resulting recognition device will be tabulated.

REFERENCES

1. "Research on the Utilization of Pattern Recognition Techniques to Identify and Classify Objects in Video Data," Douglas Aircraft Company Reports No. SM-48464-TPR-2, SM-48464-TPR-3, and SM-48464-TPR-4, Santa Monica, California, June and August 1966, and February 1967.

2. Nilsson, N. J., Learning Machines: Foundations of Trainable Pattern Classifying Systems, McGraw-Hill Systems Science Series, New York, 1965.

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

4. Parzen, E., "On Estimation of a Probability Density Function and Mode," Ann. Math. Statist. 33, 1067-1076, 1962.

5. Murthy, V. K., "Nonparametric Estimation of Multivariate Densities with Applications," Douglas Aircraft Company Paper No. 3490, 1965.

6. Tippet, et al. (eds.), Optical and Electro-Optical Processing, MIT Press, 1965.