pattern. Using 10 examples of each of the ten patterns, eigenvectors were calculated corresponding to the largest eigenvalues of the Grammian matrix for the data. Each of the 100 examples was then represented on the 8 eigenvectors corresponding to the 8 largest eigenvalues and thus the data became 8-dimensional. It was found by using the algorithm of Ho and Kashyap [4] that in 8 dimensions the 10 classes were linearly separable from each other. Two composite classes of 50 examples each were then formed by combining the examples of the digits 0,1,2,3,4, and 5,6,7,8,9. These classes were found to be not linearly separable and the transformation procedure was applied successfully to them. It is worth noting that not only was a linearly separable configuration found for the two composite classes, but the dimensionality was reduced from 8 to 3.

References

[1] R. S. Bennett, "The intrinsic dimensionality of signal collections," *IEEE Trans. Information Theory*, vol. IT-15, pp. 517–525, September 1969.

[2] R. N. Shepard and J. D. Carroll, "Parametric representation of nonlinear data structures," in *Proceedings of International Symposium on Multivariate Analysis*, P. R. Krishnaiah, Ed. New York: Academic Press, 1966.

[3] N. Nilsson, *Learning Machines*. New York: McGraw-Hill, 1965.

[4] Y. C. Ho and R. L. Kashyap, "An algorithm for linear inequalities and its applications," *IEEE Trans. Electronic Computers*, vol. EC-14, pp. 683–688, October 1965.

[5] D. F. Specht, "Generation of polynomial discriminant functions for pattern recognition," *IEEE Trans. Electronic Computers*, vol. EC-16, pp. 308–319, June 1967.

[6] G. S. Sebestyen, *Decision Making Processes in Pattern Recognition*. New York: Macmillan, 1962.

[7] H. Sugiyama and K. Joh, "A numerical procedure for conformal mapping in cases of simply, doubly, and multiply connected domains, from the viewpoint of the Monte-Carlo approach, pt. I," Osaka University, Osaka, Japan, Tech. Rept., vol. 12, pp. 1–9, 1962.

[8] R. N. Shepard, "The analysis of proximities: multi-dimensional scaling with an unknown function, pt. I," *Psychometrika*, vol. 27, pp. 125–140, June 1962.

[9] ——, "The analysis of proximities: multi-dimensional scaling with an unknown function, pt. II," *Psychometrika*, vol. 27, pp. 219–246, September 1962.

[10] T. W. G. Calvert, "Randomly generated nonlinear transformations for pattern recognition," Ph.D. dissertation, Carnegie Institute of Technology, Pittsburgh, Pa., June 1967.

[11] J. R. Rice, "The approximation of functions," *Linear Theory*, vol. 1. Reading, Mass.: Addison-Wesley, 1964.

[12] T. W. Calvert, "Projections of multidimensional data for use in man computer graphics," *1968 Fall Joint Computer Conf., AFIPS Proc.*, vol. 33. Washington, D. C.: Thompson, 1968.

[13] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "The probability problem of pattern recognition learning and the method of potential functions," *Avtomatika i Telemekhanika*, vol. 25, pp. 1307–1323, September 1964.

[14] D. B. Cooper, "Adaptive pattern recognition and signal detection using stochostic approximation," *IEEE Trans. Electronic Computers*, vol. EC-13, pp. 306–307, June 1964.

[15] W. H. Highleyman, "Linear decision functions with applications to pattern recognition," *Proc. IRE*, vol. 50, pp. 1501–1514, June 1962.

Feature Extraction on Binary Patterns

GEORGE NAGY, MEMBER, IEEE

***Abstract*—The objects and methods of automatic feature extraction on binary patterns are briefly reviewed. An intuitive interpretation for geometric features is suggested whereby such a feature is conceived of as a cluster of component vectors in pattern space. A modified version of the Isodata or K-means clustering algorithm is applied to a set of patterns originally proposed by Block, Nilsson, and Duda, and to another artificial alphabet. Results are given in terms of a figure-of-merit which measures the deviation between the original patterns and the patterns reconstructed from the automatically derived feature set.**

Introduction

THE CUSTOMARY object of feature extraction in pattern recognition is to secure a more consistent and lower dimensional representation of the differences between the pattern classes than is provided by the primary patterns. This allows the use of relatively simple decision schemes to assign unknown patterns, as characterized by the presence or absence of features, to predetermined classes.

The work reported here is confined to binary patterns and to features of the same species. Each feature is a binary vector of the same dimensionality as the patterns.

This restriction appears to eliminate many of the common pictorial, i.e., those usually defined in two spatial dimensions, measurements, such as height and width, central moments, medial axes or skeletons, lakes and bays, forks and intersections, curve following and line encoding, morpho- and quasi-topological invariants, extrema, local gradients, and the like. On the other hand, it will be seen that the vectorial features are consonant with n-tuples and, in noisy patterns, with threshold logic. The relationship between threshold logic units and the properties sought by more complex measurements has been extensively investigated by Minsky and Papert.

Manuscript received October 20, 1968. This paper was presented at the IEEE Systems Science and Cybernetics Conference, San Francisco, Calif., October 14–16, 1968.

The author is with T. J. Watson Research Center, IBM Corp., Yorktown Heights, N. Y.

Our aim is further restricted to finding a minimal set of features capable of reconstructing the input patterns to some desired degree of accuracy. The discriminating power of the features is not taken into account at all. Of course, if a feature set can be used for reconstruction of the patterns, then it is also sufficient to differentiate these patterns from one another, though it may not do this efficiently.

It should be noted that there are always at least two admissible sets of features available: the set of all the individual components, i.e., the unit vectors of the pattern space, and the set of all patterns. The object is to find a sufficient set smaller in number than the smaller of these sets. In some applications there may also be a premium on obtaining features each of which contains as few points as possible.

The problem may also be very simply stated as the decomposition of a binary $M \times N$ pattern matrix P into the Boolean product of an $M \times K$ feature matrix F and a $K \times N$ assignment matrix A in such a way that K is a minimum [7], [8]. The trivial admissible solutions then correspond to the choice of the identity matrix for either F or A.

Alternative Formulations

The problem of designing a highly nonlinear classification system in two stages has been approached in many ways. In terms of adaptive systems, Rosenblatt and Widrow schools, it is equivalent to training a two-layer machine. The two layers are not independent; the modifications in each layer must depend, through some back-propagating error correction or adaptation scheme, on the overall performance of the categorizer. Various probabilistic methods have been suggested to promote the exploration of the vast number of possible weight assignments, but success still seems around the corner.

Viewing the problem as one of dimensionality reduction, one could apply the statistical and algebraic techniques of factor and discriminant analysis, principal components, or Karhunen–Loeve expansions. Unfortunately, all of these depend on the generally nonbinary eigenvectors of the appropriate covariance matrices, and there seems to be no practical method of confining the resulting features to a binary space. There are, to be sure, orthonormal expansions defined on binary spaces, the Rademacher–Walsh functions, for example, which could be used as basis vectors, but here also the coefficients turn out to be real valued variables. In this direction, Chow's investigations of tree-dependence probably represent the best hope.

From the viewpoint of switching theory and Boolean algebra, the task is to reduce the number of conjunctive terms in a series of disjunctions. No systematic methods exist, to our knowledge, to accomplish this for expressions containing of the order of several hundred variables. This also applies to the related fixed word length coding problem.

It is sometimes sufficient merely to select a small subset of features from a pool of intuitively designed or computer generated candidates. Here the information theoretic and statistical distance criteria of Lewis, Allais, Kamentsky, Liu, and others, are available. The serious drawback of this procedure is that no matter how many features are tried, in any situation of practical interest they represent only an infinitesimal fraction of all possible combinations.

Block, Nilsson, and Duda (BND)

The point of departure for the experimental work reported here is a relatively simple set of patterns constructed by Block, Nilsson, and Duda, to evaluate their own feature extraction algorithms, as reported at the 1963 COINS symposium [1]. This excellent pilot study, described in the learning machine idiom, is based on the formation of successive intersections of the Boolean patterns, regarded as subsets of a retinal set, to obtain the core features of the pattern set.

Block *et al.* postulate the existence of a minimal set of features such that 1) each pattern can be synthesized from the features (as a union), and 2) the intersection of all the patterns containing a given feature is that feature. While the first condition is clearly essential, the second condition merely restricts the number of possible solutions to a given problem. In practice it is relatively easy to satisfy both conditions 1) and 2), the main difficulty being the requirement that the features constitute a minimal set.

The sequential algorithm proposed in [1] is shown to obtain features satisfying all the requirements provided that a certain threshold condition, guaranteeing that the size of the smallest feature is much greater than that of the intersection between any two features, is met. None of the examples successfully processed by Block, Nilsson, and Duda satisfy this condition. In fact, it is quite difficult to construct examples to meet the condition without resorting to nonoverlapping features.

The number of computations required by the sequential algorithm, for a known threshold, is proportional to $N \times M \times K^2$, where N is the number of patterns, M is the number of components in each pattern, and K is the number of features. BND also describe, without proof of convergence, a faster parallel algorithm. The features are found here in between $N \times M \times K$ and $N \times M \times K^2$ steps, but each step requires verifying whether the pattern set can be reconstructed with the already derived features.

The pattern set used by BND is shown, together with the unique optimal feature set, in Fig. 1(a) and (b). Notice that if the retina squares are renumbered as in Fig. 1(c), then the features can no longer be easily obtained by inspection.

Feature Determination as a Clustering Problem

In the scheme of [1], each associator or A unit becomes selectively attuned to a set of similar patterns, i.e., those sharing a common feature. Thus this method is in a sense equivalent to clustering the pattern vectors in the space defined by the components.

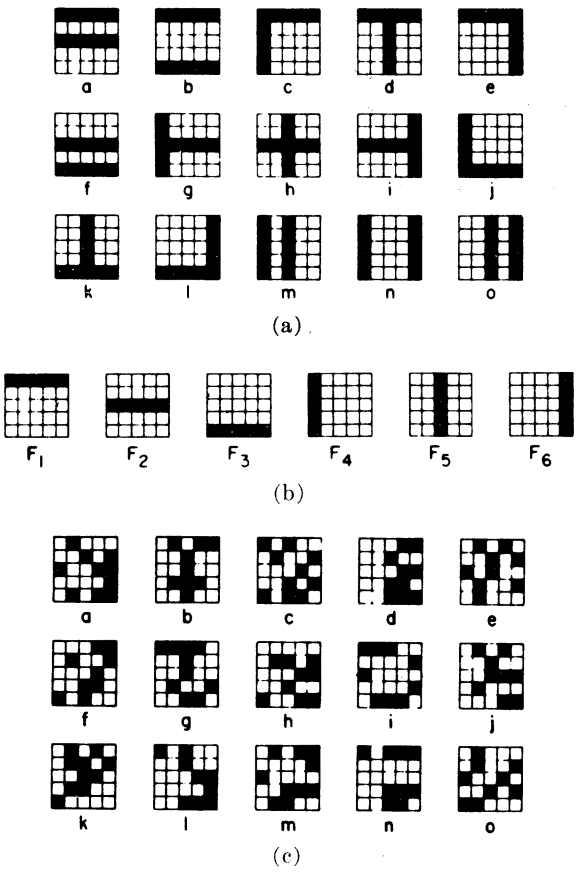

Fig. 1. (a) Pattern set proposed by Block *et al.* to test feature extraction scheme. (b) Smallest set of binary features sufficient to reconstruct the patterns of (a). (c) Scrambled version of pattern set produced by arbitrary renumbering of the retina squares.

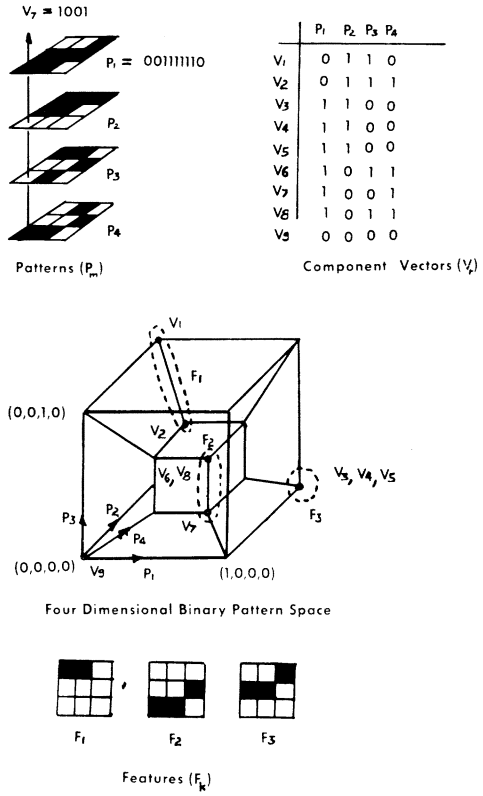

Fig. 2. Four-pattern example of disposition of the nine component vectors V_i in the four-dimensional pattern space, and the features obtained by clustering these vectors by inspection.

The novelty of the approach proposed in this paper resides in the determination of features by the disposition of the component vectors in pattern space rather than vice versa. Tight groups or clusters in this space contain elements which tend to be present or absent in the same patterns. Since this is precisely the intuitive meaning of "feature," success in our endeavour presumably depends only on the availability of adequate clustering techniques.

The idea is illustrated in Fig. 2 by means of a simple example consisting of four patterns on a 3 × 3 retina. The binary pattern space is schematically represented by a four-dimensional cube. Unfortunately an ordinary three cube could not be used because it is impossible to construct a non-trivial example with only three patterns, and thus, at most, two useful features.

By definition the ith element of the jth component vector is 1 if and only if the jth element is 1 in the ith pattern. Thus if the patterns are considered as the columns of a binary matrix, then the component vectors are the rows of this matrix, as shown at the top of Fig. 2.

The three clusters of component vectors obtained by inspection are circled in dotted lines. The collection of retina points or components in each cluster form the features shown at the bottom of Fig. 2. The reconstruction of the original patterns from these features, although imperfect, is as good as can be possibly achieved with only three features.

The principal clustering method used in the experiments described further on is the popular K-means or Isodata technique [2], [3] modified to allow overlapping clusters. Both randomly assigned and computed starting points were tried in clustering, with the latter proving vastly superior.

The computed starting vectors were derived from a similarity matrix obtained from the pair-wise distances among the component vectors as suggested by Abraham [4]. This procedure corresponds to the determination of the disconnected subgraphs of the undirected graph associated with the similarity matrix. The necessary ultimate connectivity matrix was obtained with Baker's algorithm [5] for Boolean matrix multiplication, although it has been kindly pointed out since, by Prof. P. Robert of the University of Montreal, that a procedure based on Warshall's algorithm would yield the same result more rapidly [6].

Another option allows the experimenter to introduce his own best guesses as starting features to see if the program can improve upon them.

A trivial sufficiency condition for the success of the algorithm is the existence of nonintersecting features cor-

responding to clusters concentrated at single points. As in Block, Nilsson, and Duda's method, small intersections between the features represent the most favorable conditions. The speed of operation of the least-squares clustering algorithm is proportional to $N \times M \times K$.

A Figure of Merit

Each feature set is automatically evaluated by the program by comparing the pattern set, as reconstructed from the features, to the original set. The figure of merit is the total number of bits, or Hamming distance, in which the synthetic pattern set differs from the original. In the example given earlier, for example, the figure of merit for the three features is two.

The synthetic pattern is the union of all the features assigned to a given pattern. Features are assigned in turn to every pattern, on the basis of closeness in the Hamming distance sense, until the appropriate term in the figure of merit begins to increase. At that point the synthetic pattern is deemed complete.

To render the system less sensitive to false starts and to avoid very similar features in the final set, two routines based on the pattern synthesis portion of the figure of merit computation were added to the program.

1) Whenever two features are similar, i.e., within a given Hamming distance of one another, one of them is replaced by the unreconstructed portion of the least successfully reconstructed pattern.

2) The least used feature, if used in fewer than a specified number of synthetic patterns, is replaced in the same way. In either case, the new set of features are used to form initial cluster centers for a renewed series of iterations.

Neither of these heuristic attempts proved signally successful; to cope with difficult practical problems, more profound changes will have to be introduced.

Results

The algorithm described in the preceding sections was programmed for an IBM 7094 computer, as shown in Fig. 3. The program requires about 1000 FORTRAN statements, with two short machine language subroutines for computing distances between binary vectors. The running time for each experiment is of the order of a few minutes.

For the BND patterns the optimal set of features, as shown in Fig. 1(b), were derived both with random and with computed initial cluster centers. In the latter case, the main clustering algorithm converges in three cycles at an overlap threshold of five. The program tries a full range of overlap thresholds, selecting the final feature set on the basis of the lowest figure of merit. For these patterns a figure of merit of 0 is reached at several values of the threshold.

For a more severe test of the method, the 36-character alphabet shown in Fig. 4(a) was constructed. This set was designed without any regard for reconstructibility in terms of features.

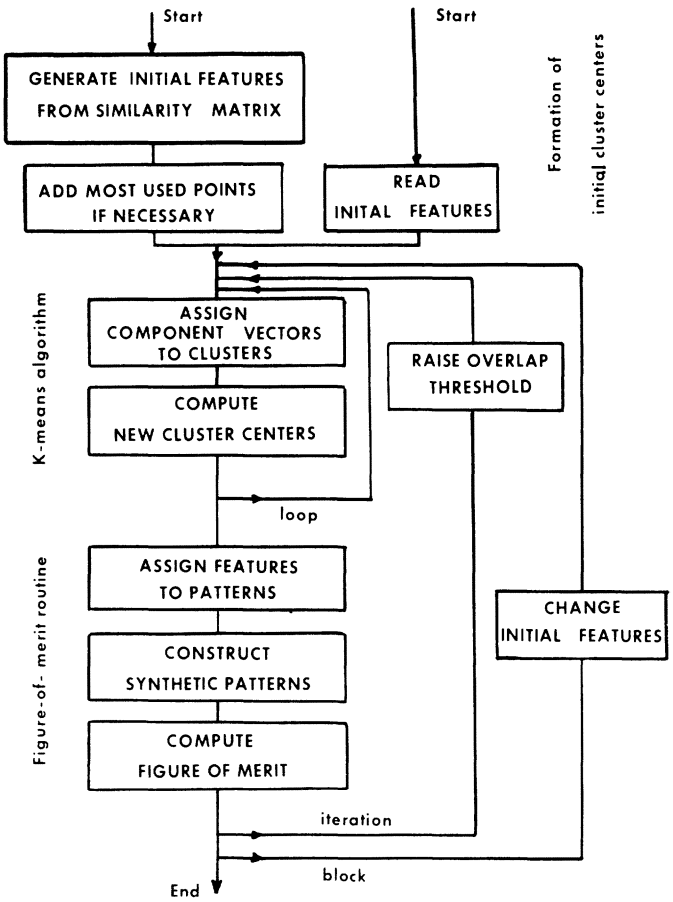

Fig. 3. Major portions of the feature extraction and evaluation program. Tests for leaving the nested loops have been omitted.

TABLE I
Number of Features

	10	15	18	20
Random initial features	140	63	—	—
Best guess features	112	45	—	—
Similarity matrix features	85	42	33	18

Table I shows the figures of merit F obtained with different methods of generating initial features and for varying numbers of features in the final set. By way of comparison, two sets of 10 features proposed by different persons came in at F of 140 and 147, respectively, although eventually the best features of these and the machine generated set were combined to produce an F of 82 with 10 features. Whether this figure is the lowest possible remains unknown.

The best sets of 10 and 15 features generated by the computer are shown in Fig. 4 along with the corresponding synthetic patterns. Of course, with 15 features the synthetic patterns resemble the originals more closely.

It is noteworthy that with the 36-character alphabet the sets of features consistently correspond to nonoverlapping clusters containing all the component vectors. The BND features have a 20-percent overlap.

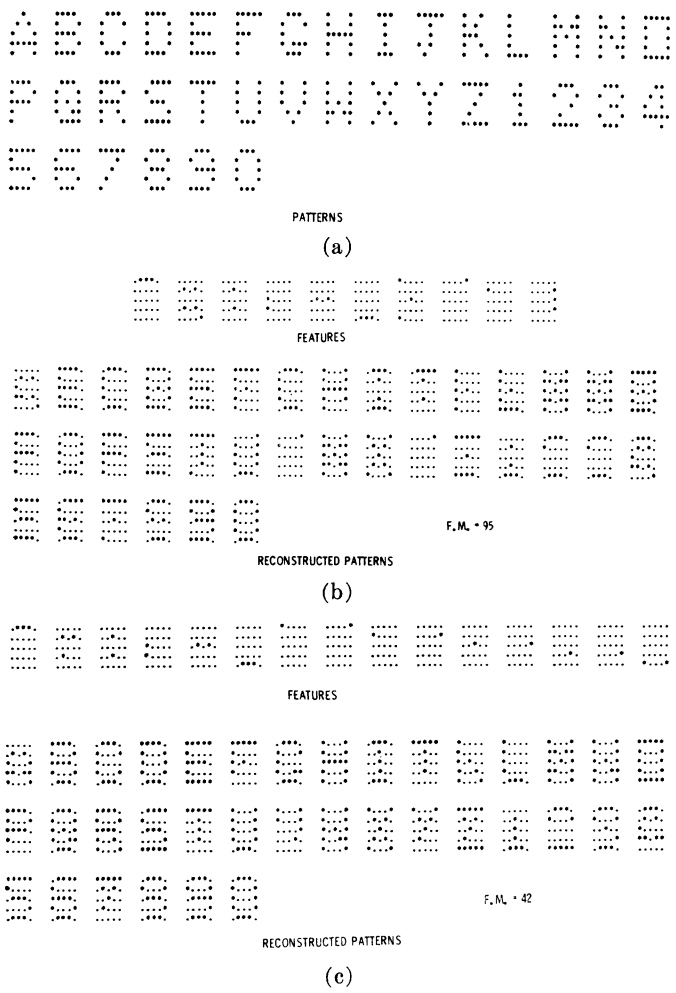

Fig. 4. (a) 36-character alphabet with significant overlap between patterns. (b) 10 automatically derived features and corresponding synthetic patterns. (c) 15 automatically derived features and corresponding synthetic patterns.

Ternary Features

While binary features are in principle adequate for full reconstruction of any binary pattern set, it is clear that for the classification of noisy patterns there is much to be gained by using multilevel features.

Consideration of black, white, or neutral features introduces very little complication in the feature generating portion of the program. The complement of each component vector, which has a 1 in every position corresponding to a pattern which is 0 or white, at that point, is included among the vectors to be clustered. Since the clustering algorithms do not distinguish between the black and the white component vectors, a given feature may include both black and white elements with all others neutral. Contradictions do not arise because a component vector and its complement are always sufficiently far from one another to be included in the same cluster.

The evaluation of a ternary feature set is, however, difficult. The union operation cannot be used for pattern reconstruction, and the figure of merit cannot be based on the Hamming distance. The arbitrary nature of the decisions necessary to circumvent these difficulties renders experimentation with ternary features quite unsatisfying. Much effort was spent on trying to find satisfactory ways of generalizing the reconstruction method and the figure of merit. It soon became clear, for example, that for patterns of interest it is unfructuous to try to treat the black and white points completely symmetrically.

Because in the figure of merit computation weights must be assigned to neutral elements in the synthetic pattern which correspond to black or white elements in the original patterns, the ternary feature sets could not be directly compared to the binary sets derived earlier. Without the weights an all neutral synthetic pattern could be misconstrued for perfect reconstruction.

In spite of our inability to formulate this portion of the problem clearly, it seems worthwhile to continue to try to close the gap between the rudimentary data reduction techniques available for binary channels and their more sophisticated continuous counterparts.

Conclusion

A new way of regarding features in binary patterns has been introduced. An algorithm based on this point of view was programmed for a digital computer, and the feature sets obtained by the algorithm were evaluated by means of a precise error criterion.

The new method seems faster than previous, algorithms. The transposition of the clustering idea from the pattern vectors to the component vectors results in a pleasing duality which is more fully developed in [7].

Although pictorial examples were used to test the method, it is clearly quite general and would encounter no more difficulty with the patterns of Fig. 1(c) than with those of Fig. 1(b). This lack of specialization can be a liability as well as an asset.

A more serious disadvantage is the absence of satisfactory sufficiency and necessity conditions for the correct convergence of the algorithm. This shows up strikingly in the four-pattern example, where there seems no way of extending the set of three plausible features to the four perfect features corresponding to the patterns themselves.

Recent work by Abdali shows that replacing the thresholded Hamming distance by a logical inclusion test leads to firmer theoretical foundations [7]. Since the inclusion method does guarantee complete reconstruction, albeit sometimes with an excessive number of features, perhaps it could advantageously replace the similarity matrix method as a preliminary step. This study also contains a first attempt to apply a mechanistic feature determination rule to automatically scanned handprinted characters.

The calculations in [1, Appendix I] show the amount of data reduction which could be realized by a workable binary feature extraction scheme. Aside from the somewhat questionable applications in pattern recognition, the development of improved feature synthesizing methods could lead to impressive economies in narrow-band image transmission and data compression, particularly where the

data exhibits consistency or repetitiousness but the volume transacted exceeds the capability of more conventional coding techniques.

Acknowledgment

The idea of sideways clustering of pattern vectors was first proposed by Dr. Richard Casey of IBM's Watson Research Center. Mr. Louis Loh, also of IBM, suggested numerous improvements in the course of seeing the program through nine successive versions.

References

[1] H. D. Block, N. J. Nilsson, and R. O. Duda, "Determination and detection of features in patterns," in *Computers and Information Sciences*, J. T. Tou and R. H. Wilcox, Eds. Washington, D.C.: Spartan, 1964, pp. 75–110.

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of the 5th Berkeley Symp. on Statistics and Probability*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297.

[3] G. H. Ball and D. J. Hall, "ISODATA, an iterative method of multivariate analysis and pattern classification," presented at the 1966 Internatl. Communications Conf., Philadelphia, Pa. (also Stanford Res. Inst. Rept., April 1965).

[4] C. Abraham, "Evaluation of clusters on the basis of random graph theory," IBM Corp., Yorktown Heights, N.Y., Res. Memo., 1962.

[5] J. J. Baker, "A note on multiplying Boolean matrices," *Commun. Assoc. for Computing Machinery*, vol. 5, no. 2, 1962.

[6] P. Robert and J. Ferland, "Généralisation de l'algorithme de Warshall," *Rev. Franç. Inform. Rech. Opér.*, vol. 2, no. 7, pp. 71–85, 1968.

[7] S. K. Abdali, "An algorithm for feature extraction and applications to hand-printed characters," Computer Sciences Dept., University of Montreal, P. Q., Canada, Rept., 1968.

[8] N. J. Nilsson, "Some experiments in reconstituting images from features," Stanford Res. Inst., Interim Rept. 1, Proj. 4621, October 1963.

Automatic Analysis of Sleep Electroencephalograms by Hybrid Computation

JACK R. SMITH, MEMBER, IEEE, MICHAEL NEGIN, MEMBER, IEEE, AND ARNOLD H. NEVIS

Abstract—**An automated sleep electroencephalogram (EEG) analyzer has been designed and tested in an effort to eliminate tedious and variable human interpretation of experimental EEG data. Data is presented to a hybrid computer from EEG tapes recorded during experimental studies in a human sleep laboratory. Special analog filters are used to identify specific transient waveforms in the EEG. Bandpass filters are used to detect the rhythmical waveforms. The outputs of these filters are then processed by digital logic circuitry, whose algorithms emulate the rules used by human readers quantitating the level of sleep each minute according to the EEG pattern. Preliminary results give 89-percent correlation with a minute-by-minute comparison to the human evaluation of the same test EEG.**

Introduction

RESEARCH investigators and electroencephalographers have been working on the automatic analysis of the electroencephalogram (EEG) ever since the pioneering work of Grass and Gibbs in 1938 [1]. They have anticipated that this automation would not only result in a large savings in manpower, but that the greater resolution of electronic devices would provide additional information beyond that which can be obtained from a visual reading of the EEG. This paper describes a procedure for automatically analyzing sleep EEGs by using hybrid computer techniques to process the data in a manner similar to that used by electroencephalographers.

Today the annual research budget for this automatic analysis is very large in terms of computer time, research efforts, and total costs. Almost all of the effort is concerned with a frequency decomposition of the EEG (i.e., an analysis of the inherent periodicities). Although this has provided considerable information, some of the significant EEG patterns are aperiodic, such as K-complexes and spike waves. Certainly these aperiodic events are critical to diagnostic electroencephalography and to sleep monitoring electroencephalography. To detect these aperiodic waveforms, different data-processing techniques must be employed.

The procedure described here utilizes matched filters to detect the aperiodic activity of short duration (1–2 seconds). This information with the information on periodic waveforms determines the state of sleep of the subject once each minute. It would be possible to carry out the same analysis on a digital computer, but the computation time would be excessive. This system combines the efficient data-processing capabilities of analog filters with the logic capabilities of digital circuits to analyze sleep waveforms in an efficient manner.

A similar approach was used in 1966 by Berger and Meier [2] to determine three stages of vigilance in animals.

Manuscript received December 20, 1968; revised May 8, 1969. This work was supported in part by NSF Grant GK 1563.

J. R. Smith and M. Negin are with the Department of Electrical Engineering, University of Florida, Gainesville, Fla.

A. H. Nevis is with the Department of Electrical Engineering and Medicine, University of Florida, Gainesville, Fla., and Veterans Administration Hospital, Gainesville, Fla.