AN EVALUATION OF RECENT DEVELOPMENTS IN THE FIELD OF LEARNING MACHINES

Oliver G. Selfridge

Lincoln Laboratory*, Massachusetts Institute of Technology

When it was suggested that I contribute a paper to this session, I had in mind that I would discuss and try and put into some kind of technological context the other papers of the session. Much of my own work of recent years has been in the field of learning machines, and artificial intelligence. There are some of us who are interested in seeing machines behave intelligently, and some of us who are only interested in having the machine simulate theories about how real brains work. I suppose that the former must predominate here, and I belong to that class myself.

It is therefore a reasonable question to ask how we shall recognize intelligent behavior in a machine when we manage to find some. I'm not sure that I can answer that except by saying that I should try to use the same standards that I use in people; but I start out by being prejudiced that people, my friends at least, are intelligent and that machines are not, even the ones I'm friendly to. There are a very few computer programs that have behavior which, even if not bright, cannot be called stupid; the famous checkers program by Arthur Samuel of IBM is one. But I do not call them intelligent yet. The learning that they do, while interesting and flexible, does not seem to me to show ineluctable evidence of being of the caliber that can be called intelligent. However, it is a real question.

Unfortunately, in the present session, that question does not even arise. I shall discuss them from the point of view of artificial intelligence, but none of the three other papers purports to discuss or

*Operated with support from the U.S. Army, Navy and Air Force.

propose subjects that have very much to do with artificial intelligence, so perhaps the kindest thing to do for the sake of consistency would be to retitle the session post facto. It may be argued that I am being over strict here and applying high standards of relevance not usually applied. After all, it may be said, even if these works do not themselves discuss or propose techniques of artificial intelligence, perhaps the techniques they do discuss may be useful in artificial intelligence. That, of course, I do not doubt. Automatic Recognition Techniques[1], the first paper, may well have presented ideas useful in devices to classify and recognize pictorial inputs; that is, useful as part of pattern recognition schemes, which are almost certainly going to be necessary in any artificially intelligent machine. But when I attend a session on computer systems, I do not expect to find a paper on the manufacturing processes of ferrite cores, though they are certainly useful in most computer systems today.

Dr. Rosenfeld of course has no illusions and no pretensions on this score. In his abstract there was no suggestion that he purported that his work was connected with artificial intelligence except that he said, "This technique was suggested by an analysis of human recognition processes."

Dr. Rosenfeld's aim, as I understand from his abstract, is to deal chiefly with aerial photographs, but I intend to discuss his work as it applies to more general kinds of visual input.

For the kind of problems Dr. Rosenfeld is talking about, his techniques are almost certainly going to be useful. But I think that to some slight extent he overstates the generality of that applicability.

His technique is based, as he says, on "two fundamental operations, each of which serves to reduce the irrelevant information content of the input while preserving its major informational features." Here he is making a very arbitrary judgment about what is and what is not relevant. In some cases he is right, but often he will be wrong. Of the two operations, the first measures texture and classifies different regions according to it, and the second extracts simple geometric figures - "those involving simple straight line combinations (parallels, perpendiculars, etc.), circles, and so forth."

There seems to me to be no doubt he is describing important features of pictorial inputs, but they are not universally important, nor always even applicable. Dr. Rosenfeld defines texture as depending on the local first and second order distribution functions of the input amplitude; that is, on the mean and variance of the input amplitude locally. This will take care of many interesting cases. Fig. 1 is from a paper by Bela Julesz[2] of the Bell Telephone Laboratories in the recent special issue of the Proceedings of the Professional Group on Information Theory. The difference between the left half and the right half of the picture is that the mean amplitudes are different. (One must compute means and variances over not too small pieces of the picture.) In Fig. 2 each part of the figure has the same mean, but the variances are different. And we can tell that there is a distinction between the two parts, also, although it is not as clear as in the first case.

The notion of visual texture is certainly applicable to more cases than merely aerial photographs. Fig. 3 shows an example which is probably susceptible. The complicated crowd at the top of the picture would have,

I presume, more or less uniform texture. Fig. 4 from the latest issue of The Farm can also be classified on the basis of Dr. Rosenfeld's texture. But such photographs seem to me to be rare on the basis of a few minutes hunting through magazines.

The same general comment ought to be made from the extraction of the simple geometric figures. Fig. 5 from the current issue of Datamation shows circles that can be extracted, but only because the photograph was taken full face; from any other points of view circles look like ellipses. The figure of the bullfighter and the crowd (Fig. 3) was taken from a recent issue of the magazine Playboy. That magazine has many examples, which I am not going to show as slides, of photographs with abundant texture and a happy absence of straight lines and simple geometric figures. It is on the basis of this evidence that I dispute Dr. Rosenfeld's assertion that "most important among these figures will be the simplest ones, e.g., those involving straight line combinations... circles and so forth."

But one should realize that there are other obvious limitations, as Dr. Rosenfeld well acknowledges. Fig. 6 has a consistent first and second order distribution all over the picture, though it is conceivable that there are some edge effects that may be detectable. But not as boundaries between areas distinguishable by his techniques. And if we add a noise level of a curious kind, as in Fig. 7, the square still stands out, though not so clearly as before.

Thus I am suggesting that the techniques of Dr. Rosenfeld are only a couple of extra tools for our work bench. And all the tools we have cannot approach our own visual processing system. Fig. 8 is blown up from

an advertisement in Aviation Week, I believe. It is immediately obvious what the larger letters are, although most of them are absolutely unreliable. But look at the smaller ones! Not only can we read the town, but the state as well. (It may help to defocus slightly.)

It ought to be said more often that real life neurons are, as far as we can discover, almost fantastically reliable in the ordinary sense. I don't know where they got their reputation for unreliability, but it is unfounded. And there are many creatures who thrive happily with single neurons handling many of the important functions of their bodies.

This theme, reliability through redundancy, is in great favor these days, and the paper by Professor Pierce[3] follows well in the footsteps of Von Neumann. The trend these days is somewhat away from the majority-of-three technique which others have analyzed in detail.

It is interesting in science and technology how the rise in popularity of some descriptive term causes it to be applied very widely and where it would not have been applied before. Some years ago in Lincoln Laboratory was developed a communication system called Rake[4] which handled the reception of a radio signal which had been smeared out in time by ionospheric multipath. The contributions from the various relative delays had to be undelayed, so to speak, and recombined to form the true signal. Fig. 9 shows the general scheme of the Rake receiver. But before the contributions are recombined they first have to be weighted by, guess what!, their reliability. Exactly as with Professor Pierce.

The reliability of the signal at each of the delays is measured by the factors a_i, and these factors are computed from the average correlation of the signal with both reference signals at that delay. This procedure

is obviously identical with the procedure that Professor Pierce numbers IIB. (I might add that there are other communication problems that have been solved and analyzed in the same way.)

I mention this both because it is interestingly relevant and because my friend, Dr. Robert Price, who is the co-inventor with Dr. Paul Green, of Rake, tells me that he had not realized that Rake was an adaptive system till many years after it had been built. And are not automatic gain circuits adaptive, too?

I think I must disagree when Dr. Pierce says that "An adaptive vote-taker is a simple type of self-organizing system", as he does in his summary. An adaptive vote-taker is surely a simple type of adaptive system, but I don't see where the self-organization comes in. The Rake system is a maximum likelihood detector, or perhaps I should say decision system, and it can be called adaptive if you like, though it may be stretching things a little. But its organization is always absolutely unchanged. Even so with the adaptive vote-taker; its organization and processing are always absolutely unchanging.

What such systems can do is the following: given a set of more or less independent variables each contributing to some decision, they can assign stable weights to the variables according to a maximum likelihood rule.

Dr. Pierce claims that "the use of statistical decision theory for building reliability is new..." Of course it is very far from new - Rake is an example among many others. His Bayes analysis has been carried out in a large number of other places, and is a standard part of elementary courses in communication theory. His Fig. 4 (shown in Fig. 10), which

does not, by the way, look like an electrical circuit at all, goes back not merely to 1959, but is substantially the same as the neurons in the diagrams of McCulloch & Pitts[5] nearly 20 years ago.

I find myself slightly disturbed by the positive nature of Dr. Pierce's conclusion that his analysis has proved that "vote-takers are excellent devices for use in redundant systems." Since he has not in fact tried them, why is he so sure that actual systems will have the requisite statistical properties? And later, he says: "The use of even simple binary adaption can greatly increase the reliability... Consider a system with 100 stages... Errorless decision elements are used after each step." Well, if I were building the system, I would build it entirely of those errorless decision elements.

The third paper, by Dr. Malcolm Uffelman[6] of Scope Incorporated, starts out with "The general problem considered is the design of a system for the classification of minimally constrained stimuli." Most of the paper is in fact about that general problem. But I wonder why he calls it a "conditioned reflex" system. Let me first of all take a look at conditioned reflexes as they are known to the people who deal with them, who are called psychologists.

The conditioned reflex was of course first intensely studied by Pavlov[7], who set the level of Russian psychology ever since. You are no doubt all aware of the primitive experiments. A dog is shown a piece of meat and at the sight his mouth fills with saliva in anticipation. If he is shown a piece of meat and simultaneously a bell rings or a metronome ticks while he is salivating, then, after several such trials, the sound of the bell is sufficient by itself to cause salivation. The sight of the

meat we call the unconditioned stimulus, and the sound of the bell the conditioned stimulus, and the salivation the response.

I hope that my description of it made it sound like a fairly simple phenomenon, because it is indeed a simple and very reliable experiment as I described it. However, if we examine the circumstances surrounding it, and wonder about its reasonable and logical extensions, it becomes very obvious that it is a simple expression of a very complicated and difficult underlying mechanism. For example, Pavlov showed that the conditioned stimulus must precede the unconditioned stimulus in order to establish a stable conditioned response, and in general the order of the events seems very important to the organism. Now why should that be?

There is another large question which has bothered psychologists and still does. Suppose that we have conditioned the dog to salivate to the meat on hearing a note on the piano, say Middle C. To what other stimuli will the dog salivate? Will the dog salivate to the next note D? Suppose we play Middle C on a bassoon? and so on. Suppose we choose a note three octaves away? Suppose we give him some stimulus which isn't even an auditory one? In general the responses vary, and we are more likely to see responses with the first examples I gave than with the later. The phenomenon here is called stimulus generalization. The opposite phenomenon also occurs, called stimulus differentiation. If the dog salivates both to Middle C and Middle D, say, we can train the dog to differentiate between them by reinforcing one of them (by actually giving the dog the meat) and never reinforcing the other.

Now in what respects can we say that Conflex I exhibits anything like a conditioned reflex? Not in very many, I am afraid. Nothing depends on

the order of presentation. Nothing corresponds to extinction, for instance. Extinction is the name given to the phenomenon where the response to the conditioned stimulus gradually dies down when it never receives the meat.

Dr. Uffelman has made here a very curious choice to use Random Stimuli in random classes. Why on earth random? Are they supposed to be more useful because they are random? At least none of the enormous body of work or conditioned reflexes has ever used random stimuli. We may not easily understand or analyze the binding rules for classes of similar stimuli for man and beast, but we may all be very sure that they are not random rules.

Conflex I exhibits only generalization and differentiation. From the paper Dr. Uffelman presented, he claimed no more. He says also that generalization is based purely on stimulus overlap. That is, Conflex I averages all the signals in one class so that those units always present have a large vote in favor of that class. One question here is what are the units of the signals. They must be the cells in the D-field, which correspond to some combination of the actual receptor units.

Let us re-phrase that statement while treating Dr. Uffelman's diagram (Fig. 1) as an unreliable binary process. The inputs to the summing elements are given heavier or lighter vote weights by the weighting factors m_{ij} here according as the inputs are more or less reliable. In other words Conflex I is really an adaptive system of the kind discussed by Dr. Pierce.

Naturally there are differences. I should like to be able to compare their results directly, but I cannot. For one thing, Dr. Uffelman's unreliability comes from and with the signal, rather than from malfunctioning of circuit elements.

In Dr. Uffelman's conclusions he states "The theory developed...has shown that a relatively powerful, adaptive pattern recognition system can be built." Relative to what? Certainly not relative to anything that shows conditioned reflexes. And certainly not relative to other pattern recognition techniques which go beyond mere overlap.

It is true that his scheme permits "a simple method for expanding the class size of a basic system", by means of the "M-fields". Now in the ordinary run of the mill random net study the M-fields themselves do not form a separate layer at all but are merely the weighting of the connections. An analysis Marvin Minsky and I[8] did two years ago at the London International Information Theory Symposium applies to Dr. Uffelman's scheme.

The other comments we made in that paper hold, too, but they reflect more widely on whether one can extend the techniques (that can make a machine like Conflex I work) to perform tasks that are harder.

For there are tasks, even in this context, that are harder. Instead of assigning the stimuli to the classes <u>at random</u>, supposing they matched to some real property but a fairly sophisticated one. For example, suppose that one class consists of pictures of fraternal twins and another of identical twins. Where are the random fields sensitive to this difference?

There is thus an enormous gradation of classification or pattern recognition problems. Conflex I can learn to recognize and respond to a black blob on a white background. It can certainly not separate fraternal from identical twins. I believe that the threshold of its capabilities is not very far along the road from the first to the second. Sadly, I

have only vague ideas about how to proceed faster or much further down the road.

As I said, the implications for artificial intelligence of Dr. Uffelman's paper and Dr. Pierce's paper are roughly the same. We can optimize a system with many variables, so that each pulls his appropriate weight. But only if their effect is, say, additive or independent.

If they have to be combined in some complicated non-linear way, we have very little to go on. The trouble is that most of the interesting problems seem to involve just such cases. Although many of the people with random nets will tell you their nets can generalize beyond simple overlap technique or simple linear maximization, they have not yet shown any examples.

Dr. Rosenfeld has described his work as a tool which may be useful for artificial intelligence. Both the other papers may be useful tools, too. One thing I miss very much is the experiments which can tell us whether those tools are in fact useful; and to what extent and in what way. Perhaps the authors can soon evaluate this work in the best way, by seeing how valuable they are to the advancing front of science and technology.

REFERENCES

1. Rosenfeld, A., "Automatic Recognition Techniques Applicable to High-Information Pictorial Inputs," Proc. IRE, March 1962, to be published.

2. Julesz, B., "Visual Pattern Discrimination," IRE Trans. on Information Theory, Vol. IT-8, No. 2, p. 84.

3. Pierce, W. H., "Adaptive Decision Elements to Improve the Reliability of Redundant Systems," Proc. IRE, March 1962, to be published.

4. Price, R. and Green, P. E., "A Communication Technique for Multipath Channels," Proc. IRE 46, p. 555.

5. McCulloch, W. S. and Pitts, W., "A Logical Calculus of Ideas Imminent in Nervous Activity," Bull. Math. Biophys., Vol. 5, pp. 115-137, 1943.

6. Uffelman, M. R., "Conflex I - A Conditioned Reflex System," Proc. IRE, March 1962, to be published.

7. Pavlov, I. V., CONDITIONED REFLEXES, 1927.

8. Minsky, M. L. and Selfridge, O. G., "Learning in Random Nets," Reprint from INFORMATION THEORY, Fourth London Symposium published by Butterworths, 88 Kingsway, London, W.C.2.

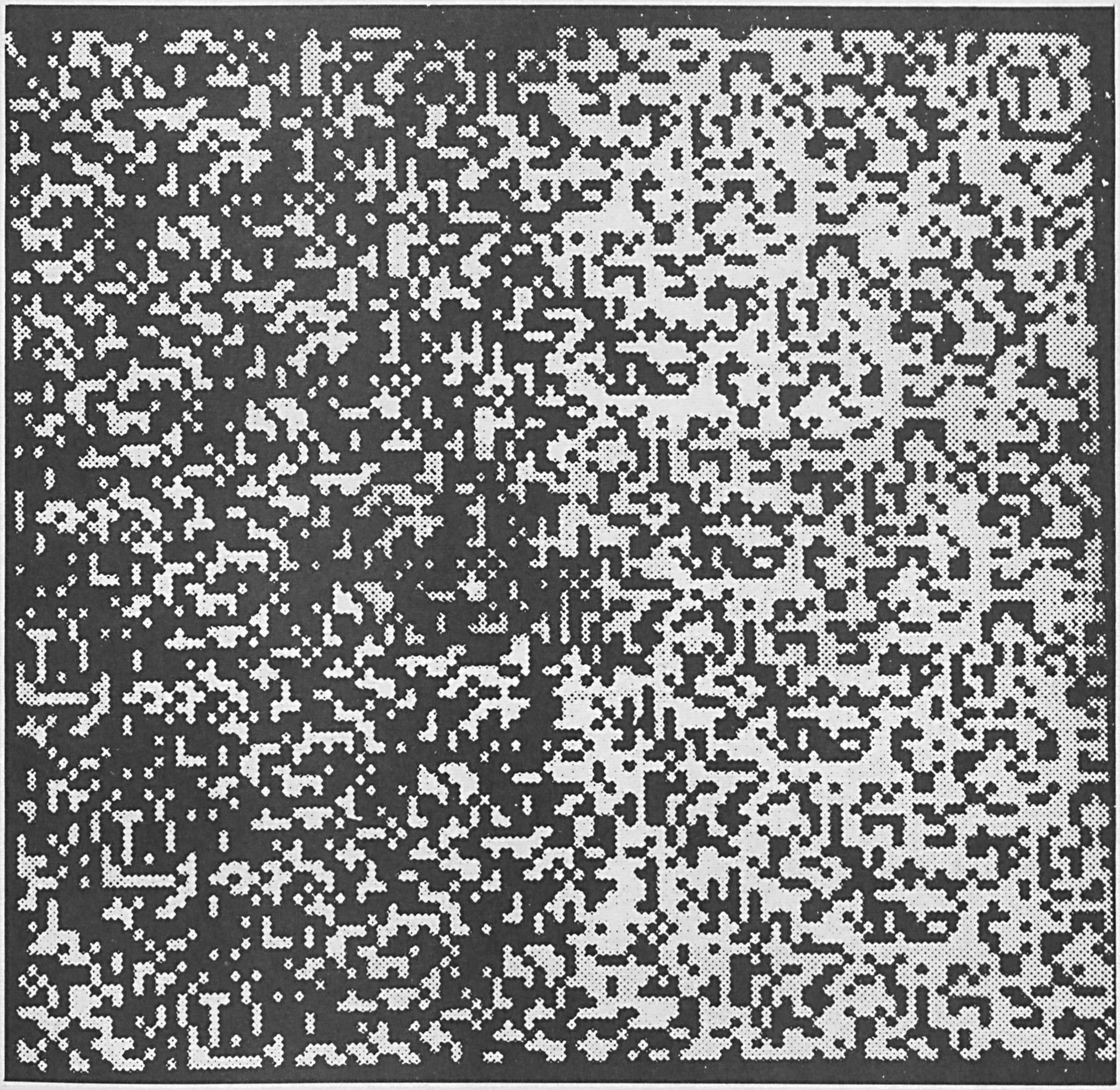


Figure 1

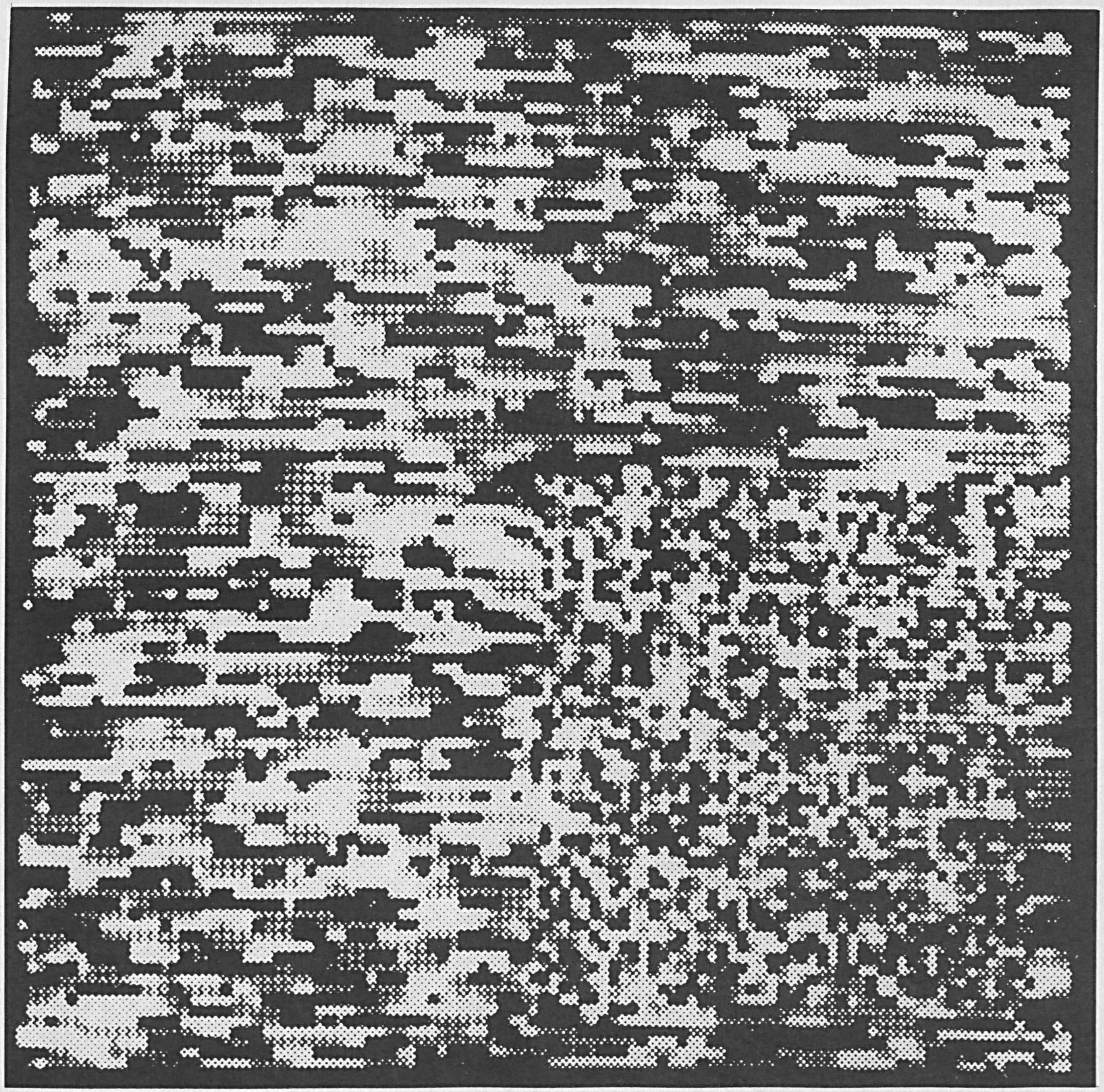


Figure 2

Figure 3

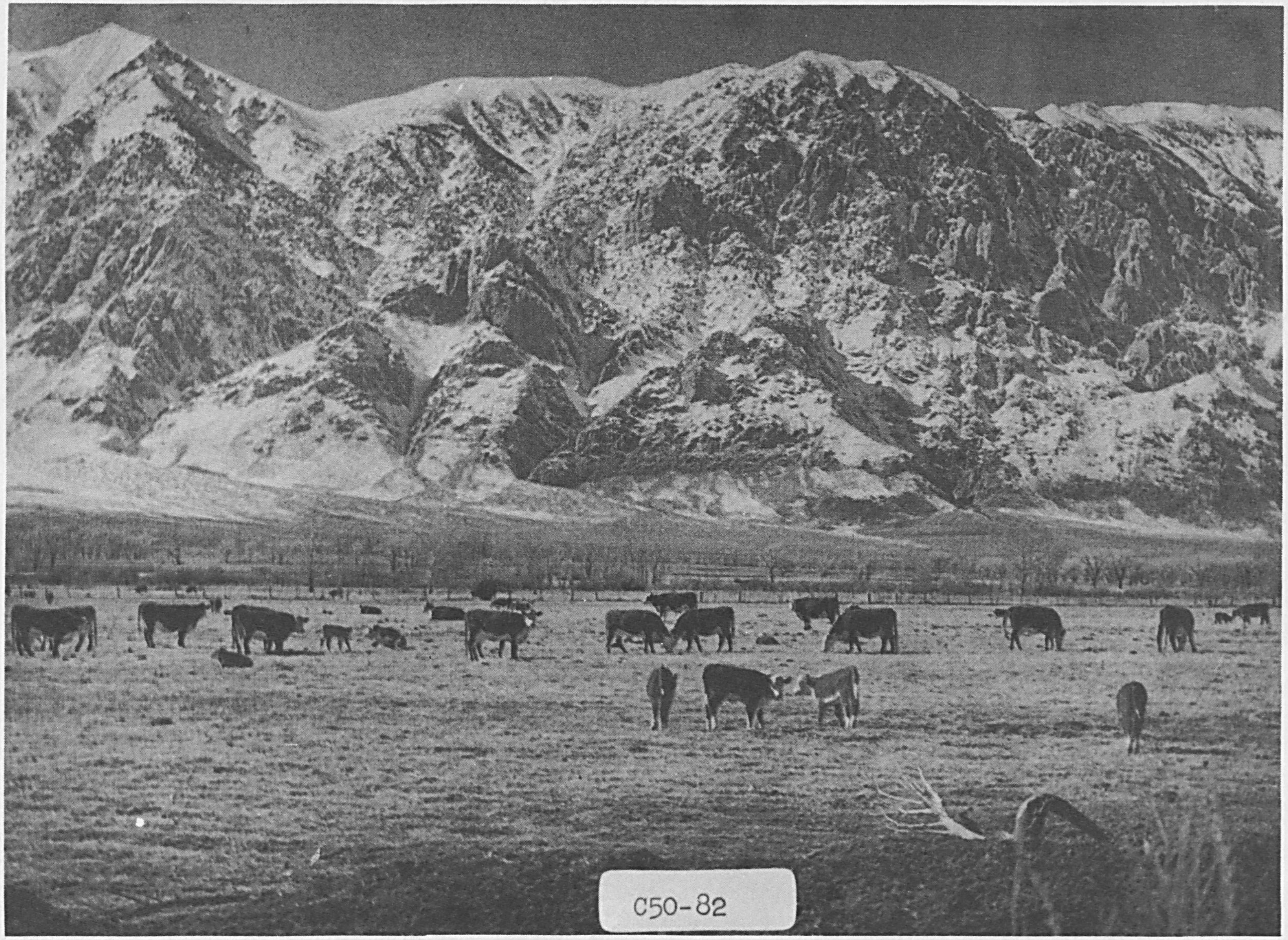
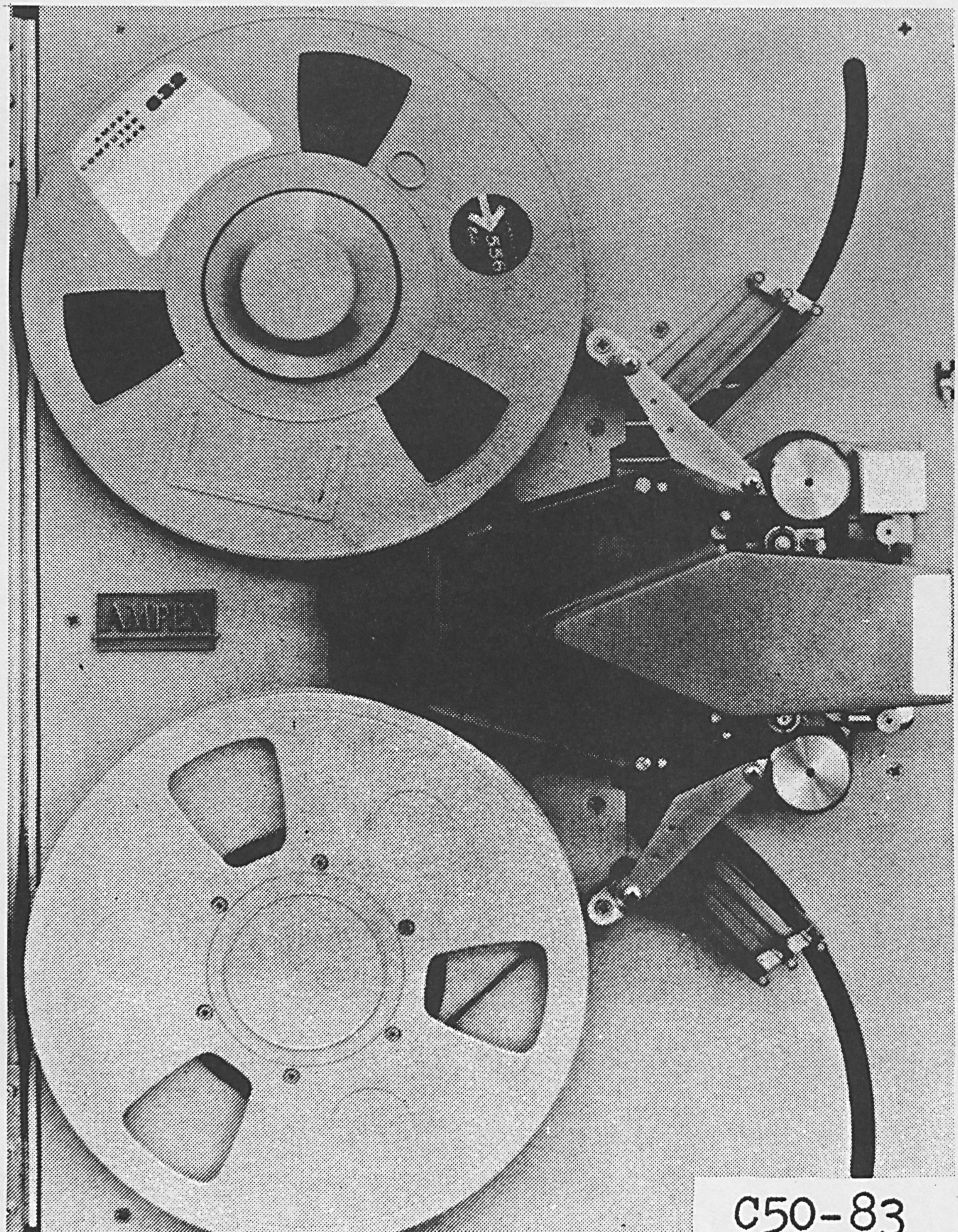


Figure 4



C50-83

Figure 5

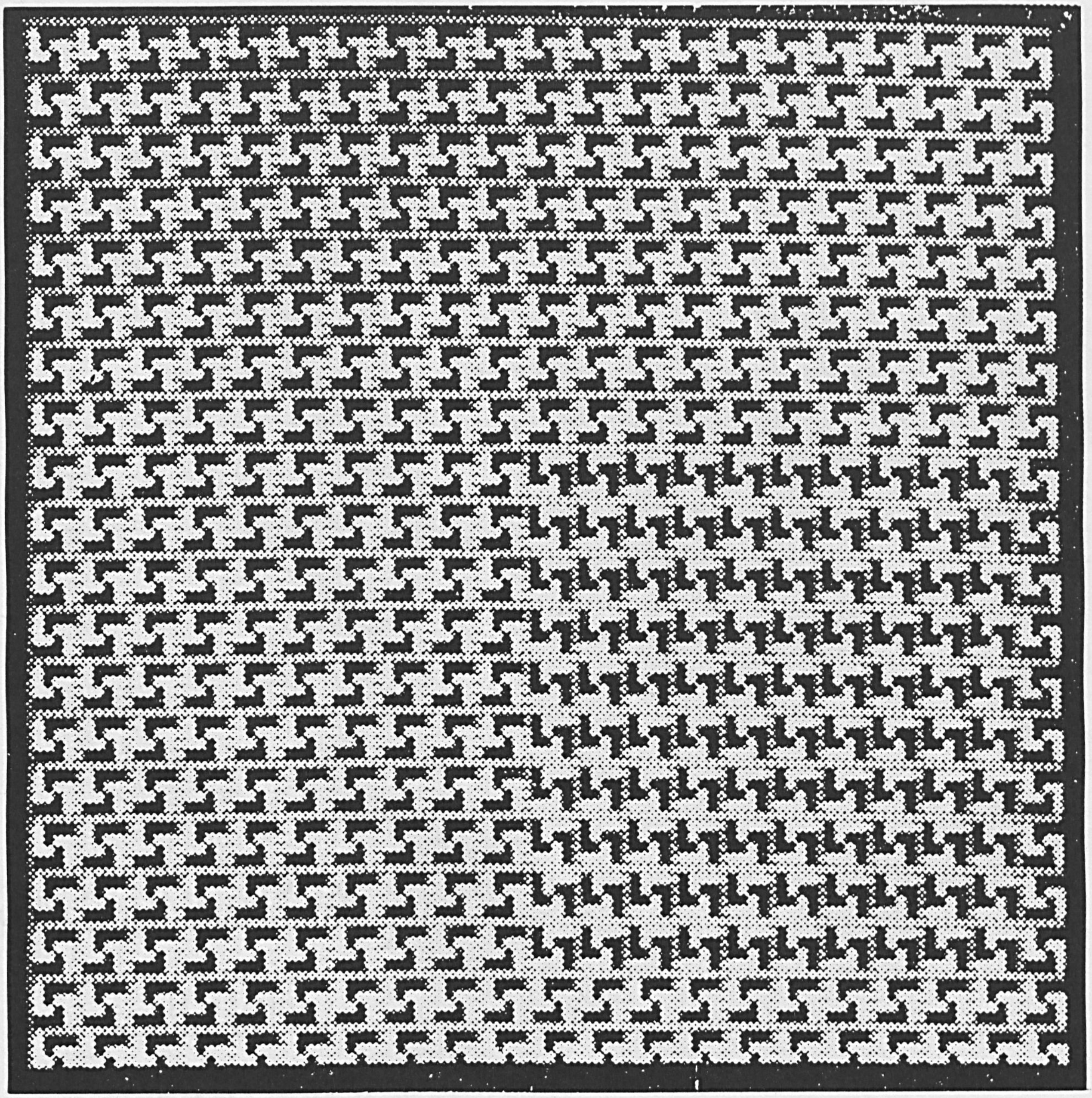


Figure 6

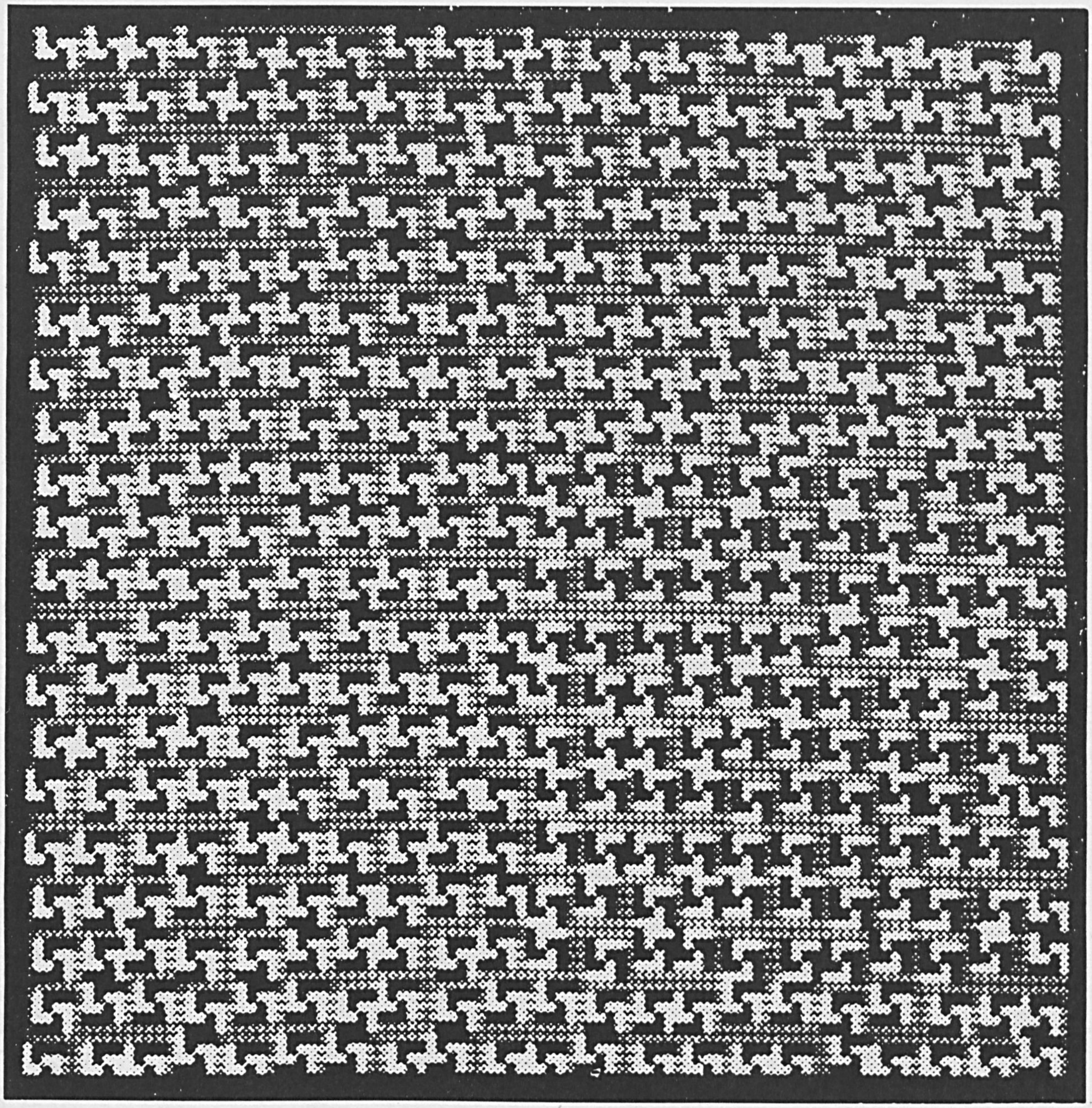


Figure 7

Figure 8

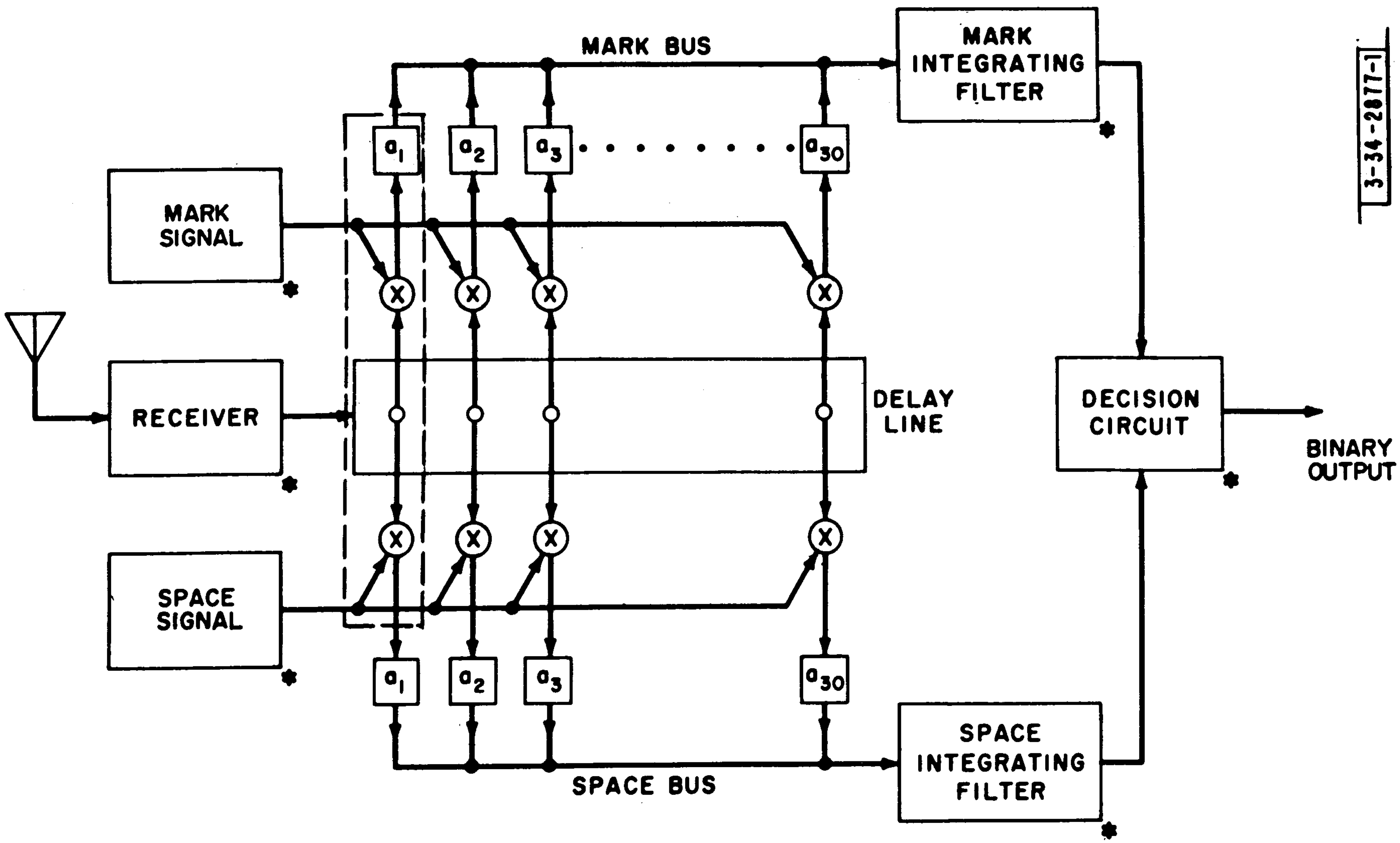

MARK BUS
MARK INTEGRATING FILTER
3-34-2877-1
MARK SIGNAL
a_1
a_2
a_3
a_{30}
RECEIVER
DELAY LINE
DECISION CIRCUIT
BINARY OUTPUT
SPACE SIGNAL
SPACE BUS
SPACE INTEGRATING FILTER

Figure 9

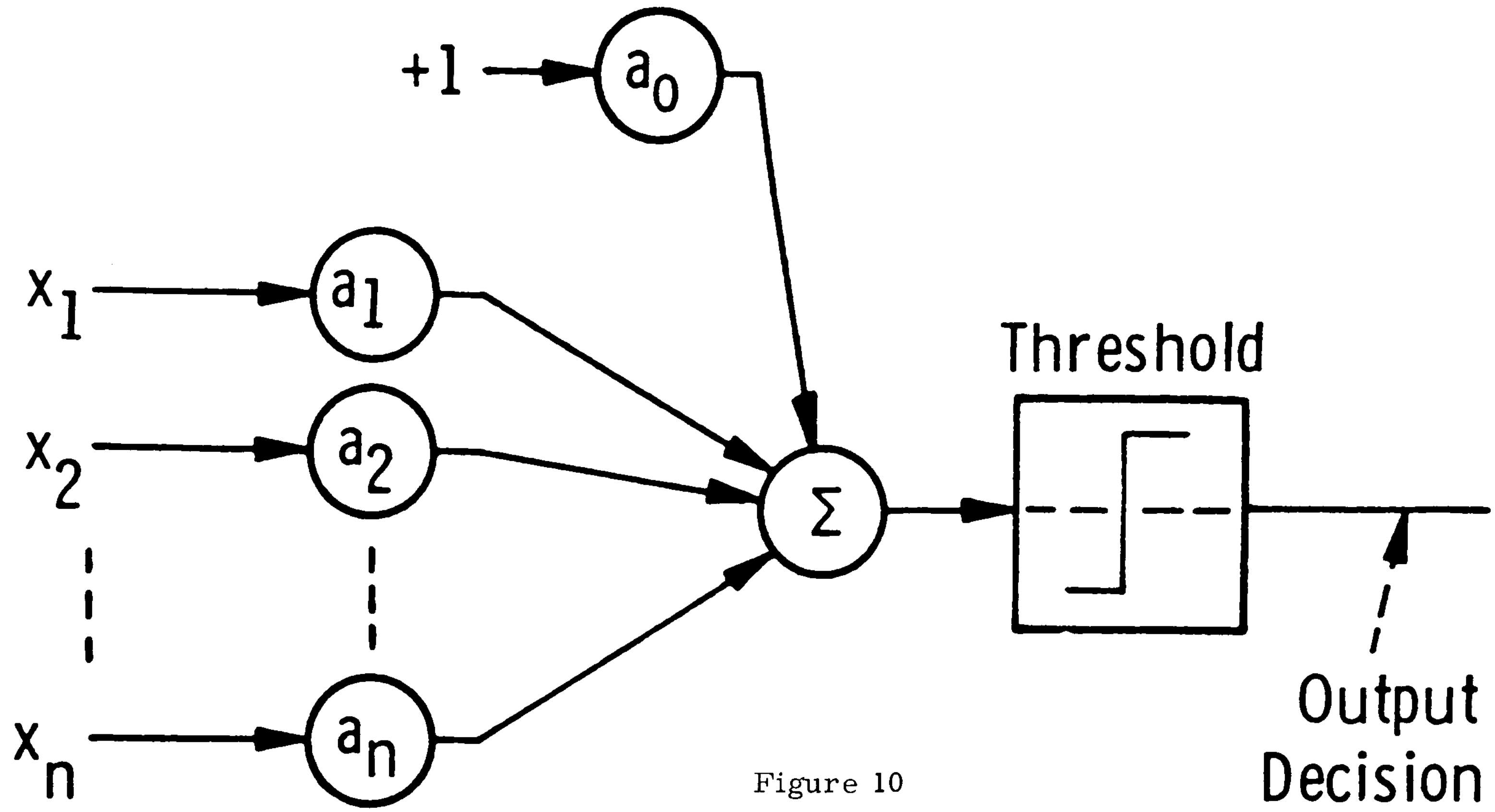
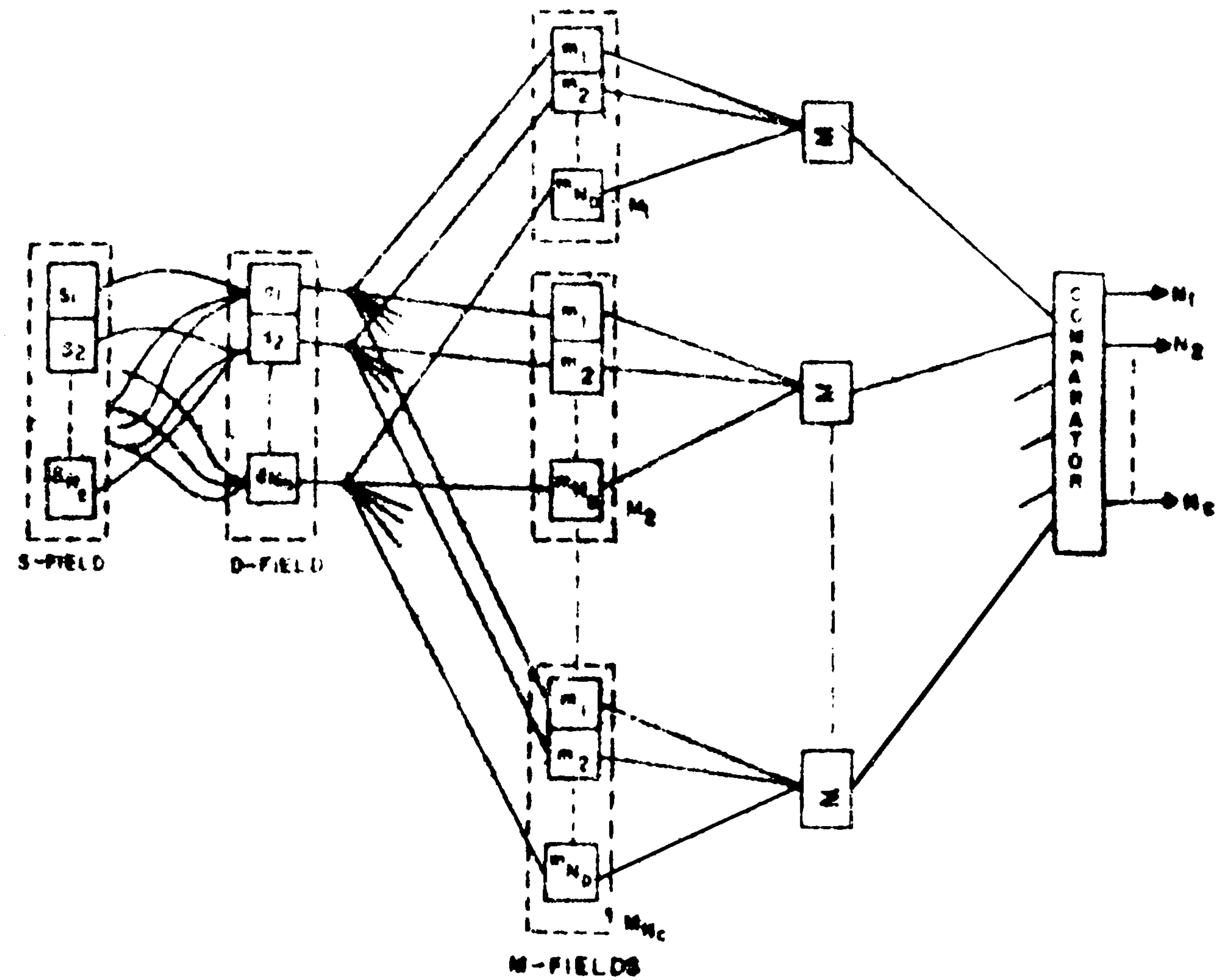


Figure 10

A decision element or vote-taker for binary inputs of +1 or -1

C50-88



ORGANIZATION OF THE CONDITIONED REFLEX MODEL.

Figure 11