The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturers' or trade names does not constitute an official endorsement or approval of the use thereof.

DESTRUCTION NOTICE: For classified documents, follow the procedures in DoD 5200.22–M, Industrial Security Manual, Section II–19 or DoD 5200.1–R, Information Security Program Regulation, Chapter IX. For unclassified, limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1991	Review article, from 7/89 to 7/90

4. TITLE AND SUBTITLE
A Theoretical Perspective on Aided Target Recognition Research

5. FUNDING NUMBERS
DA PR: AH44
PE: 61102.H4411

6. AUTHOR(S)
Dennis McGuire

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Harry Diamond Laboratories
2800 Powder Mill Road
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
HDL-TR-2194

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Laboratory Command
2800 Powder Mill Road
Adelphi, MD 20783-1145

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
AMS code: 611102H4400
HDL PR: 1AE152

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This survey closely examines the mathematical methods in use for the ill-posed problems invariably confronted at the beginning stage of the automatic/aided target recognition problem, where features relevant to the ultimate recognition goal must be extracted from images of various kinds. These methods include maximum a posteriori estimation, regularization theoretic approaches, multiple deconvolution techniques, and the simulated-annealing method called stochastic relaxation and annealing. Also examined are some of the methodological and philosophical issues relevant to the image recognition task. A relatively detailed theoretical perspective on the overall problem area is conveyed.

14. SUBJECT TERMS
ATR, automatic target recognition, target recognition, early vision, MAP estimation, MAP, regularization theory, deconvolution, simulated annealing

15. NUMBER OF PAGES
54

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	17. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Contents

1 Introduction

This review is intended as a companion to Garvin's[1] concurrent survey, which examines aided target recognition (ATR) methods primarily from the point of view of their implementations in digital-electronic and optical hardware. He concludes, among other things, that the area of this field which is undergoing the most intensive and mathematically substantive study is the front end of the ATR problem, what has been called the *early-vision* stage by computer vision researchers and the *pre-attentive* stage by cognition theorists (cognitive psychologists and artificial intelligence (AI) researchers). He further points out that while many aspects of the ATR problem have been under investigation for decades, only recently has this difficult and critical front-end aspect been vigorously attacked.

Accordingly, the purpose of this survey is to closely examine the mathematical methods in use for this stage of the ATR problem. These methods include *maximum a posteriori* (MAP) *estimation*, *regularization theory*, the *multiple deconvolution* technique of Berenstein and his collaborators[2,3] (for image resolution enhancement), and the rigorous simulated-annealing technique called *stochastic relaxation and annealing*. Other important and relevant methods that are not considered in this review include the *renormalization group* approach of Gidas,[4] the relatively new *mean field annealing* technique,[5] and the extensive methodology known as *mathematical morphology*[6] (this methodology may be the subject of future review).

Except for the morphological techniques, all these methods are aimed at solving "ill-posed" problems (defined in sect. 2), either at the formulation of solutions or at devising efficient algorithms for their computation. Ill-posed problems invariably arise with the image-processing

[1] C. G. Garvin, *Survey of Aided Target Recognition (ATR) Techniques from Digital and Optical Perspectives*, HDL technical report, to appear.

[2] C. A. Berenstein, B. A. Taylor, and A. Yger, *Sur quelques formules explicites de déconvolution*, J. Opt. (Paris) **14** (1983), 75–82.

[3] C. A. Berenstein and B. A. Taylor, *Overdetermined Systems of Convolution Equations*, Proc. Sixth Army Conf. Applied Comput. Math., Boulder, CO (June 1988).

[4] Basilis Gidas, *A Renormalization Group Approach to Image Processing Problems*, IEEE Trans. Pattern Anal. Machine Intell. **PAMI-11**, No. 2 (February 1989), 164–180.

[5] G. L. Bilbro, T. K. Miller, W. E. Snyder, D. VandenBout, and R. Mann, *Simulated Annealing using the Mean Field Approximation*, IEEE Conf. on Neural Information Processing Systems, Denver (November 1988).

[6] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press (1982).

problem known as feature extraction, which is one of the main aspects of the first stage of the ATR problem (discussed in detail in sect. 2). As a consequence, much of the mathematical technique used in the ATR field is concerned with the issue of ill-posedness, which is the thread that ties together virtually all the discussion in this review.

The report is organized as follows. In section 2 the general aspects of the first stage of the ATR problem are discussed, and the central importance of ill-posed feature extraction problems is established. Various mathematical approaches to ill-posed problems are discussed in subsequent sections. MAP estimation in its simplest form is the subject of section 3. Section 4 describes the alternative and often equivalent approach of regularization theory.

At this point the mathematical discussion is temporarily interrupted for further discussion of the problem (raised in sect. 2) of the *knowledge-theoretic base* and for a brief interim summary.

Section 6 resumes the mathematical discussion by giving a concrete example of the ill-posed nature of the deconvolution problem. Section 7 describes a relatively recent solution due to Berenstein and his collaborators.[2,3] This solution is predicated on the production of several different convolutional probings of the same unknown object, or, more pertinently, several different images of the same scene. For completeness' sake, section 8 gives an example of the application of regularization theory to the *image flow* problem. In section 9, I discuss in detail a more ambitious and sophisticated MAP estimation scheme (largely due to S. Geman and D. Geman[7]). This scheme rigorously develops a simulated annealing approach (known as stochastic relaxation and annealing) to MAP estimation problems whose prior distribution is a *Markov random field* (by prior distribution I refer to the statistics of the image before the image is obtained). Section 10 gives an overall summary.

[7]S. Geman and D. Geman, *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Trans. Pattern Anal. Machine Intell. **PAMI-6**, No. 6 (November 1984), 721–741.

2 The Nature of the Problem

What is the beginning-stage ATR problem, why is it important, and what is so difficult about it?

In the context of computer vision and from the perspective of neural network research, Kersten et al[8] have called early vision the "estimation of scene properties from image data," describing it as the "intermediate goal—between data acquisition and the final achievement of recognition—which attempts the construction of a unified representation of object surface information such as orientation, depth and reflectance, inferred from various sources of information such as stereo, motion, and color." In a recent review concerned with the regularization-theoretic approach to vision, Poggio et al[9] state that "Early vision consists of a set of processes that recover physical properties of the visible 3-dimensional surfaces from the 2-dimensional intensity arrays." They point out that these properties have been generally assumed to be context independent, i.e., involve only general constraints about the physical world and the imaging system, and that they have consequently been treated as conceptually independent modules that can be studied in isolation, at least to a first approximation, with the idea of combining them at a later stage in the overall processing.

From the perspective of cognitive psychology and AI,[10,11] the same problem arises in considering the first information-processing step that either occurs in human cognition or should occur in an intelligent machine.* The investigation of this first stage, known in these fields as the pre-attentive stage, has led to the general description of its task as the production (from the raw sensory or raw data input) of a collection of *features* and some low level of organization among them (such as the bars and edges and their associations in the *primal sketch* of Marr[11]),

[8]D. Kersten, A. J. O'Toole, M. E. Sereno, D. C. Knill, and J. A. Anderson, *Associative Learning of Scene Parameters from Images*, Appl. Opt. **26**, No. 23 (1 Dec. 1987), 4999-5006.

[9]T. Poggio, V. Torre, and C. Koch, *Computational Vision and Regularization Theory*, Nature, London, **317** (1985), 314.

[10]John R. Anderson, *Cognitive Psychology and Its Implications*, W.H. Freeman & Co., New York (1985).

[11]D. Marr, *Vision*, W.H. Freeman & Co., San Francisco (1982).

*The work of D. Marr, although directed at vision, was in a decidedly different spirit from the early-vision research discussed in the previous references (Kersten et al and Poggio et al), and is more appropriately regarded as in the AI arena.

which as a whole becomes the subsequent matter for the attentive faculty (or whatever plays this role in the intelligent machine).

Regardless of the field, be it cognitive psychology, AI, computer vision, or ATR, the first-stage problem is essentially the same. The differences in these fields pertain to the character of their raw data: whether they are organized in a linear sequence (like the phonemes of speech) or in two-dimensional arrays (like optical images), whether they arise from electromagnetic or mechanical influences, etc.

The beginning-stage ATR problem can accordingly be stated in general terms as follows: What, specifically, should first be done with the raw data obtained? This question is to be answered in view of the immediate objective of producing a collection of features and certain groupings among them, i.e., feature substructures, which all together form a scheme that generally describes the raw data field presented in reduced terms. The importance of this step is that it determines the computational feasibility of the overall task and establishes the strategy for achieving the recognition goal.*

This problem has both a knowledge-theoretic and a mathematical aspect. The epistemological issue lies in the need to establish a knowledge base for the overall identification task: what we know generally about the data field that is relevant to what we want to recognize. From this general knowledge the particulars will follow: i.e., what features of the represented environment should be sought in the raw data field. For example, the goal of computer vision research is to be able to distinguish and identify a certain class of objects that the vision system represents in the form of a two-dimensional optical image obtained from a particular viewing point. The image is a sampled distribution of light intensity or color over a two-dimensional field; it will be, in all essential respects, like the distribution of retinal stimulations humans receive when looking at a scene. In order to distinguish objects (i.e.,

*It should be noted that this general assessment, although true for all machine attempts at speech or image recognition, is an oversimplification of what is known about the recognition aspects of human cognition. The pre-attentive transformations of raw sensory influences in humans do produce partially organized collections of features—the *icons* produced from retinal stimulation by the visual apparatus and certain hard-wired pre-processing operations in the brain, for instance—but it is not correct to regard these feature schemes as sketching an ample data field in reduced terms, as though the more ample data were in fact available to the brain, because these feature schemes are all that the brain ever has to work with. The function of selectively reducing the amount of data to be further and intensively processed is performed in humans by the attentive faculty, not before the operation of this faculty.

three-dimensional figures with substance and color), we must abstract from the image such characteristics as the boundaries of objects, their colors, color textures, and relative locations, and perhaps even their movements. Achieving such a sketch of the objects in the scene would amount to pre-attentive success. (In the next step, the images of the outlined objects would be more carefully examined to identify them, by going back to the raw data emphasized by the sketch, but this is beyond the point of our example.)

The human knowledge-theoretic base, which understands the environment in terms of bodies in a spatio-temporal relation, thus appears to furnish a list of the image features needed for the early vision task: object boundaries, their associated colors, textures, etc. In this case, then, there is apparently no need to mount an investigation to establish the knowledge-theoretic base.* For other ways of forming images of an environment, however (sonar, synthetic aperture radar, and infrared images, for instance), and indeed even for the vision case, other epistemological bases may be necessary or more convenient. In any case, once the list of features has been drawn up, the mathematical aspect of the task comes into play.

Here we need to establish a mathematical theory of how the various relevant features of the environment are reflected in the raw data field; in other words, if we denote by i the mathematically structured way in which an environmental feature f (which also needs a mathematical definition) is reflected in the data field, we must begin with a functional relation between f and i, say $i = A(f)$. Suppose, for example, that we are interested in the surface boundary feature of a three-dimensional object, and that we are dealing with an optical image. Three-dimensional features are of course not fully manifested in optical images because such images contain only two-dimensional projections. Thus the surface boundary feature, $f = S$, of an environmental object appears in the overall image as a component $I = A(S)$, where A is a projection operator whose explicit form is determined by the characteristics of the

* Even this is not so clear, however. A *phenomenologist* would argue that the evidence for this knowledge may be illusory or prejudicial. According to this view, it is not clear what in a visual image is being recognized or, alternatively, what structures of retinal data ultimately induce us to imagine and see a familiar object, i.e., recognize it. Moreover, there is a method, the Husserlian phenomenological method, that can provide a very delicate evidentiary description of what actually happens when something is being seen or recognized, and such a description might provide a better list of features for the recognition task. (Personal communication from Jorge García-Gómez, a phenomenologist with expertise in the philosophy of science, Philosophy Department, Long Island University.)

viewing system. The complete image is then a superposition of various I's over the two-dimensional image field.

Because most of the difficulty in the mathematics of feature extraction is nicely illustrated by this example, I will continue with it. (There may be better examples from the standpoint of practical relevance, but the point here is only to illustrate the mathematical difficulties.) The surface boundary feature, S, that we wish to extract will present itself in the image as a closed curve, say C, giving the body's two-dimensional silhouette. It is clear that C does not give us nearly enough information to find S. But the problem is even worse, because in practice several factors combine to blur and distort our image of C: for instance, the diffraction limit of the imaging system, the pixel size associated with the system's detectors, and noise. Still another source of difficulty is the fact that the images of other bodies may be confused with the one bounded by C. Therefore, besides the projection A, we will clearly need another operator, B, to express the pixelizing and blurring of C, as well as a way to characterize noise and interference. Thus the actual image of S is more adequately given by an equation like

$$I = B[A(S)] + N, \tag{1}$$

where N denotes the noise.

It turns out that B is in general a convolution operator; it convolves functions characterizing the various aspects of the imaging system's resolution with the infinitesimally sharp silhouette C. This fact can be seen heuristically as follows. In a rough sense convolution is a smoothing operation: it turns rapidly varying functions into more slowly varying ones. In this sense it is like integration, whose inverse is differentiation; thus the inverse of a convolution operator is something like a differential operator. Now consider B. Its inverse would bring C out sharply, apart from noise and interference. In applications to visual images, where one seeks to bring out the profiles of objects by locating the relatively sharp intensity or color changes, the operation typically performed is either differentiation or something like it. Although this argument may not be very convincing, the convolutional character of B can be rigorously established by a theoretical analysis of typical viewing systems. Thus, if we ignore interference, the problem of extracting the object boundary feature from its image I is one of solving the operator equation $I = B[A(S)] + N$ (eq (1)) for S, where B is a convolution operator and A is a geometrical projection operator of some kind.

The problem expressed by equation (1) is an example of an "ill-posed" problem. What is an ill-posed problem? Consider a functional equation of the form $u = F(v)$, where v is an n-component vector, u is an m-component vector, and F is a mapping of Euclidean n-space to Euclidean m-space, regarded as an equation to be solved for v given u and F. Such an equation is called an ill-posed problem (in the sense of Hadamard) if it has more than one solution or if the inverse mapping, F^{-1}, fails to depend continuously on the given data u. (Note that when $u = F(v)$ does have a unique solution, then F^{-1} is not many-valued and so exists strictly as a mapping in its own right.) The problem posed by equation (1) is actually ill-posed in both senses of Hadamard's definition. Since A is a projection, the composite operator $B \circ A$ will not be one to one, and so the equation will have numerous distinct solutions. Moreover, even though convolution operators are invertible on reasonable function spaces (like the space of square-integrable functions, for instance), their inverses generally fail to depend continuously on the initial data.

To see how the discontinuity of an inverse can promote serious noise problems, consider our generic equation $u = F(v)$ when F^{-1} exists strictly but is not continuous, and suppose that we are asked to find v given u. The exact solution is clearly

$$v = F^{-1}(u).$$

All that would seem to be needed to complete the solution is the explicit form of F^{-1} and an obvious computation. Unfortunately, however, the matter is not that simple when u is corrupted by noise, for then the failure of F^{-1}'s continuity means that the disturbance produced in v by the noise-induced error in u cannot be bounded. We are thus forced to conclude that given an even slightly erroneous u and the explicit form of F^{-1}, we can still determine almost nothing with confidence about the actual v that produced u.

The crucial point is that virtually every feature-extraction problem is ill-posed, and the causes of this are analogous to the ones illustrated by the example we have considered. It is consequently of great importance for the beginning-stage ATR problem that methods be developed that practically circumvent the mathematical shortcomings of ill-posedness. A number of such methods have in fact been developed, and still others are under development. Let us consider some of these in a general way.

The methods being employed can be classified as probabilistic and regularization-theory approaches. Although both of these often lead to equivalent formulations of the feature extraction problem, they are quite different in terms of both practice and prospects, depending on the particular problem. In the probabilistic view of feature extraction (which I address first), the method of solution is based on the idea of MAP estimation. Its practitioners claim that MAP estimation methods are particularly well suited to a learning approach which lets developing experience with actual scene statistics guide the imposition of the constraints plainly needed for the useful solution of ill-posed problems. MAP methods also seem to mesh well with the learning capabilities of neural networks.[8]

How do MAP estimation theorists conceive the solution of ill-posed feature-extraction problems?

3 Maximum a Posteriori Estimation

The problem, as before, is

$$i = F(s) + N, \tag{2}$$

where s is an n-component vector, i is an m-component vector, N is noise, and F maps Euclidean n-space to Euclidean m-space. This equation can be viewed in alternative ways. The collection of components of s may constitute a single complex scene* feature which requires a vector description, or it may constitute a description of several scene features whose numerical specifications are somehow packed into the components of s. Depending on which is the case, the image vector, i, will represent the part of the image coming from either the single scene feature or the collection of scene features. In other words, s generally constitutes either a part or the whole of Kersten et al's[8] "unified representation" of the environmental scene, i.e., the intermediate goal of early vision. The components of s could form a partial or complete representation of a certain object or objects; they might even constitute the complete representation of all the objects of interest. In short, equation (2) describes the problem of early vision with maximum flexibility.

To approach this problem, given that it is ill-posed in either or both of the senses indicated, a MAP theorist supposes a probability or statistical model of the situation. The *elementary events* of this probability model correspond to the various ways that the viewing system can be related to the real environment, i.e., to the different positions and orientations that the system can have in the environment, and to the various lighting and other conditions that can affect the noise term in equation (2). The occurrence of such an elementary event is fully specified by the two vectors s and i, which thus furnish the immediate vector-random variables of the theory. An elementary event having taken place (i.e., the viewer having obtained an image of the scene), the problem then becomes the after-the-fact or *a posteriori* determination of the probability of various s given that i has occurred: that is, the determination of the conditional probability function $p(s|i)$. A knowledge of this function will enable us in principle to find the scene vector s whose probability conditioned on i is maximum; this s will then be the *maximum a posteriori* or MAP estimate of the scene vector that produced the observed image.

*I use the language of the vision problem for specificity.

But how can the conditional probability function, $p(s|i)$, be determined? And what if, for some actually encountered i, the maximum of $p(s|i)$ is very broad, so that other s, differing considerably from the MAP estimate, have nearly the same chance of having produced the image? The answer to the first question is the substance of the mathematical theory of the MAP approach. The answer to the second brings in the learning aspect of this approach and the potential connection to neural-network techniques.

The method of determining $p(s|i)$ is based on the application of Bayes' rule in the form

$$p(s|i) = \frac{p(i|s)p(s)}{p(i)}, \tag{3}$$

where $p(i|s)$ is the probability that the image is i when we know that the scene is s (note that the noise term of equation (2) is the only thing that makes $p(i|s)$ other than a δ-function centered on $i = F(s)$), $p(s)$ is the unconditioned probability that the scene is s, and $p(i)$ is the similarly unconditioned probability that the image is i. The use of Bayes' rule places no restrictions on the validity of this analysis, since the rule is a direct consequence of the completely general *theorem of total probabilities* and the definition of conditional probability.

We can determine $p(i|s)$ directly from our knowledge of the mapping F and the statistics of the noise term in equation (2). On the other hand, $p(s)$ represents, to the extent that it is not uniform, our *a priori* knowledge of the environment (i.e., what we know before obtaining the image). It could represent, for instance, accumulated knowledge that was gained through prior views, perhaps from different aspects, of the same environment over time. Generally speaking, $p(s)$ is our *a priori* model of the scene statistics. Likewise, $p(i)$ can be thought of as representing our prior expectations concerning the image we are about to see: i.e., $p(i)$ would be our *a priori* model of the image statistics. We need not be much concerned with $p(i)$, however, because the MAP analysis supposes that i has just been obtained, so that $p(i)$ is simply a constant (one that we never need to know, as it happens).

We further assume (following Kersten et al) that both $p(i|s)$ and $p(s)$ are multivariate Gaussian, so that

$$p(i|s) = k \exp\{-[i - F(s)]^T [i - F(s)]/2\sigma_n^2\} \tag{4}$$

and

$$p(s) = k' \exp[-s^T M s / 2\sigma_s^2]. \tag{5}$$

Here a constant diagonal covariance matrix is assumed in equation (4), s is taken to have zero mean in equation (5), and k and k' are normalization constants. Given these assumptions, if we now take the natural logarithm of equation (3), we find that the maximization of $p(s|i)$ is equivalent to the minimization of

$$[F(s) - i]^T [F(s) - i] + z s^T M s, \tag{6}$$

where z, which is a Lagrange multiplier, equals the ratio of the noise variance to the scene variance.

The essential assumptions expressed in equation (4) pertain to the noise term in equation (2). These assumptions are specifically that (a) the m-component noise vector is multivariate Gaussian, (b) the means of each of its components are zero, so that $\langle i \rangle = F(s)$, (c) distinct components of the noise vector are uncorrelated, and (d) the expectation values of the squares of the individual components are all the same. Assumptions (c) and (d) express the meaning of "a constant diagonal covariance matrix" in more detail. The only truly restrictive assumption is (a), because if we know the statistical moments of the noise present, we can translate and rotate the coordinate system in the Euclidean m-space of i (and the noise vector) so as to make the means zero and the covariance matrix diagonal in the new coordinate system. If the resulting covariance matrix fails to be constant, a simple change of scale on the coordinate axes will make it so. All these transformations would of course have to be applied to i and the mapping F as well.

Similarly, the essential assumption expressed in equation (5) is that $p(s)$ is multivariate Gaussian. In specifying this assumption as equation (5), the only specialization adopted for the coordinate system—in the Euclidean n-space of s—is that its origin coincides with $\langle s \rangle$. The relationship between the matrix M and the standard covariance matrix μ is

$$M_{ij} = \sigma_s^2 \frac{|\mu_{ij}|}{|\mu|}, \tag{7}$$

where σ_s^2 is the squared length of the variance vector, $|\mu|$ denotes the determinant of the covariance matrix, and $|\mu_{ij}|$ is the cofactor of the ij^{th} element of that determinant.

The final minimization problem for expression (6) is of course the primary focus. Given the image vector i, prior knowledge of the scene-vector statistics in the form of the matrix M, and the mapping F, we must find the global minimum of expression (6) as a function of s.

In obtaining this formulation of the MAP estimation problem, we should note the essential simplification that results from the Gaussian-statistics assumptions. These assumptions permit us to simplify matters by taking logarithms of the exponential functions, which in turn enables us to split off the effect of $p(i)$ as an additive constant term, which can then be discarded in the optimization.* More ambitious treatments of MAP estimation can be found in the literature (one of which, MAP estimation by "stochastic relaxation and annealing," is addressed in sect. 9); these more sophisticated treatments make much weaker assumptions about the prior scene statistics and the image noise. But apart from such treatments, it is interesting and potentially useful (as Kersten et al point out) that the Gaussian formulation of MAP estimation is almost equivalent to another general formulation of the typically ill-posed feature extraction problem, namely, that of "regularization theory," to which I now turn.

*Actually, the splitting off of $p(i)$ as an additive constant term and its subsequent dismissal in the optimization problem is a procedure that can be carried out regardless of the forms taken by $p(i|s)$ and $p(s)$; this is because the logarithm is an increasing function of its argument.

4 Regularization Theory

Retaining the notation of (2) for the problem to be solved, regularization theory begins with the need to adopt norms in the spaces of i and s, and to choose a particular *stabilizing mapping*, P, which, like F, maps n-space to m-space. Note the omission of the adjective "Euclidean"; in regularization theory it is sometimes more advantageous to adopt norms other than the Euclidean distance function in the image- and scene-vector spaces. Having made these choices, the theorist then typically reformulates the ill-posed problem in one of the following ways:

(1) Find the s that minimizes $\|P(s)\|$ and is such that

$$\|F(s) - i\| < \epsilon, \tag{8}$$

where the double bars denote the image-space norm and ϵ is a small constant determined from an appraisal of the image noise.

(2) Find the s that minimizes

$$\|F(s) - i\|^2 + z\|P(s)\|^2, \tag{9}$$

where z, which is understood to be real, is called the *regularization parameter.*

What is the logic of the alternative procedures (1) and (2), and what is the significance of the "stabilizing" mapping P?

Let us consider the second question first. When our original problem is ill-posed in the sense that it has too many solutions (i.e., its solution is underdetermined), the purpose of P is to provide the additional constraints needed to single out the one "right" solution. Thus we have two distinct requirements for P. It must be such that either (8) or (9) (depending on the formulation) has a unique solution from a purely mathematical standpoint, and that the unique solution so obtained is actually the one we are seeking for the concrete feature extraction problem at hand. Loosely speaking, our original problem has too few equations or too many unknowns; so the mathematical purpose of P is to provide the extra equations. However, since the reason for the underdetermination is that equation (2) does not specify a sufficient number of the actual conditions that produced the image (that is, sufficient for the unequivocal determination of the feature of interest), P must also be chosen to reflect those conditions.

A third requirement on P emerges from the answer to the first question, about the logic of formulations (8) and (9): the mapping P must express the additional *a priori* constraints so that their satisfaction amounts to the minimization of the norm of $P(s)$, either absolutely, as in (8), or relatively, as in (9). The essential difference between these two formulations is as follows: In (8) we seek an s that satisfies $i = F(s)$ to some tolerance ϵ, as measured by the image-space norm, and also separately minimizes the norm of $P(s)$. In (9), on the other hand, the degree of satisfaction of $i = F(s)$ and the degree of minimization of $\|P(s)\|$ are flexibly distributed between the two by means of the Lagrange multiplier z, whose appropriate value emerges from the solution.

In many of the pertinent applications of regularization theory, P is a differential operator expressing certain *a priori* "smoothness" requirements on the feature we are seeking to extract. As an example, suppose that we wish to take advantage of the fact that the surfaces of manmade objects are relatively smooth compared with those of most natural objects. The notion of surface smoothness can be captured in mathematical terms through the derivatives of the functions defining the surface. When we say that a surface is rough on a certain scale, we mean that it has an undulating topographic structure, characterized by some distribution of lengths which measure the distances between peaks and valleys, and that this structure can be seen on the scale of our observations. This meaning translates into the mathematical fact that certain combinations of the partial derivatives of the functions defining the surface change appreciably over the peak-to-valley distances. This observation indicates how to impose smoothness constraints on a surface: first insist on the continuity of all partial derivatives up to some order (this will rule out abrupt changes in the partials over infinitesimal distances); then prevent the combinations of derivatives that would register the appreciable changes from changing much over the relevant distance scale by minimizing the derivatives of those combinations. This is how the differential-operator character of the stabilizing mapping can be understood as arising in the particular case of a surface.

In much of the general regularization theory of feature extraction, one attempts to think of the feature of interest in the same way, that is, in terms of smoothness; but now "smoothness" is understood in the abstract sense of the mathematical theory of *smooth* or *differentiable manifolds*. Manifolds are generally defined by families of continuous functions; when these functions are continuously differentiable up to

some order, say p, the manifold is said to be smooth or differentiable to the p^{th} order. Just as with the surface example, one first attempts to define the feature of interest as a differentiable manifold of some order of smoothness; then one imposes further smoothness constraints on this feature/manifold by minimizing the norm of an appropriately chosen combination of the continuous partials, to wit, $P(s)$; this minimization is what imposes our *a priori* expectations concerning the "smoothness" of the feature. This, at any rate, is what characterizes the approach that is sometimes called standard regularization theory, where the term *standard* often also holds the additional implication that both F and P are linear mappings, so that the problem specified by (9) becomes the global minimization of a quadratic cost functional, a problem which has been exhaustively studied and well-characterized in a general way by Tikhonov and Arsenin.[12]

Not all early-vision problems have yielded to the standard theory, however. Linear F and P and the implied quadratic cost functional (9) are insufficiently flexible to embrace a number of the concrete problems of interest. For this reason there is a "trans-standard" theory, which I will not discuss, that does not limit F and P to the linear mappings and considers other formulations than (8) and (9). The Tikhonov and Arsenin theory[12] can in fact be extended, with some limitations, to nonlinear F and P, while at the same time retaining the formulation of equation (9); but even this is not enough to handle some problems.

As a final point, we note with Kersten et al[8] that the regularization-theoretic problem (9), including its Tikhonov and Arsenin generalization to nonlinear F, is formally the same as the MAP estimation problem (6), provided that the Euclidean norm is used in the image space. This can be seen as follows. First note that

$$[F(s) - i]^T [F(s) - i] = \|F(s) - i\|^2$$

when the norm on the right is Euclidean. Thus, identifying the z in equation (6) with the z in equation (9), we would like to have the equality

$$s^T M s = \|P(s)\|^2,$$

which will hold if we assume that P is linear and make the further identification

$$M = P^T P, \tag{10}$$

[12] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, Winston, Washington, DC (1977).

for then

$$s^T M s = s^T P^T P s = \|P(s)\|^2.$$

This result (from Kersten et al[8]) reveals the near-equivalence of the MAP estimation and regularization-theoretic approaches (referred to at the end of sect. 3).

The value of this observation should be evident. Although the MAP and regularization-theory approaches are not fully equivalent, it is nonetheless valuable to have two distinct and potentially fertile ways of viewing those problems for which the approaches are equivalent. In these cases we can advance on the solution in various ways. We can employ our developing knowledge of actual scene statistics as an aid in determining M (this is virtually an experimental method). We can propose reasonable forms for the stabilization matrix P and analyze their consequences (this is a heuristic method). Finally, and possibly most fruitfully, we can compare the experimental and heuristic results.

5 The Knowledge-Theoretic Base

Before considering the other mathematical methods I intend to treat, I sum up the discussion so far and further elaborate on the knowledge base issue.

The first stage of the ATR problem has two aspects: (a) choosing the collection of scene* features to be extracted from the image and (b) developing the theoretical/mathematical machinery needed for their extraction. The purpose of the set of scene features is to render the scene "intelligible" to the viewing system in a sketchy way that requires much less data to specify than the original image. Once the scene is sketched out, a much reduced quantity of the original image information is thereby circumscribed for further scrutiny and final recognition: specifically, the information belonging to the parts of the original image that have emerged as "important" in the sketch. Deciding on the particular collection of scene features to be used is both critically important and nontrivial. This choice must be made in view of what I have called a knowledge-theoretic base. What does this involve?

The viewing-system designer has his own knowledge-theoretic base, that is, his general way of understanding the world (insofar as it is pertinent to the task at hand) and his specific technical knowledge of how the objects of interest interact with the viewing system. He must use all this knowledge to select a definitive set of features for the particular problem. Because the image data presented by the viewing system are typically very complicated and are often related to a human visual image in indirect ways, two distinct but interrelated approaches to the feature-selection problem have necessarily developed: (a) the so-called model-based approach and (b) the approach that seeks to make selections based on actual experience with the kind of imagery furnished by the imaging apparatus—what might be called an empirical approach.

Model-based approaches to feature selection necessarily begin to deal with aspects of the development of the theoretical/mathematical machinery needed to extract the selected features. They thus contribute to both aspects of the overall front-end problem and are often intended to address both aspects equally. By theoretically modeling the images of the objects of interest and typical backgrounds, as well as typical

*I continue to use the language of the vision problem.

distortions and noise effects, model-based approaches launch various studies to investigate methods for selecting and extracting the required definitive scene features. Such approaches are generally preferred to the empirical alternative.

Empirical approaches are usually developed with the aid of "trained observers," that is, individuals who have had a certain degree of experience and training with the type of imagery in question and its associated "ground truth," and who can therefore usually point to the important parts of a given image. In a rough sense, such observers have been trained to perform the first-stage ATR function. With empirical approaches, the question becomes to what extent a trained observer's knowledge of the imagery can be articulated and automated. This is the essential question motivating AI research on expert systems. The two primary aspects of such approaches are (1) putting the features that trained observers look for in an explicit form, thereby providing the basis for (2) attempting to devise algorithms that can extract these features from imagery. With the first aspect, the problem is that observers are often unconscious of what they are looking for, and even in the best of cases may be unable to fully articulate it. This leads to the difficulty associated with the second aspect—that algorithms often fail to extract such features because they are not defined with sufficient precision. These problems, it should be clear, do not arise with model-based approaches; in fact, the preference for the latter is a reaction to these difficulties.

6 The Ill-Posed Nature of Deconvolution Problems

After the important features are selected and the associated mathematical theory formulated, what remains is the actual extraction of the features from the image. As mentioned earlier, this virtually always leads to an ill-posed mathematical problem.

A typical first feature-extraction problem is to determine the outlines of bodies in an image, which is generally presented in the form of an array of pixel values of either colored or grey-scale intensities. The determination of bodily outlines is always based on noting abrupt value changes between adjacent or near-adjacent pixels. (An exception arises when texture variations are the basis of boundary determinations.) When the conditions of image formation are more or less ideal (that is, when the pixel noise level is relatively low compared with the significant level changes across boundary pixels), and when the resolution afforded by the pixel size is adequate for the discernment of the important characteristics of the extracted boundary (an important characteristic might be the boundary's rectilinearity, for example), the problem of extracting bodily outlines will not be ill-posed. This is because deconvolution operations are not necessary in such situations. However, available signal-to-noise ratio and resolution are often far from ideal, and this is when deconvolution becomes essential.

An example of the type of ill-posed problem that can result when resolution and signal-to-noise ratio are less than adequate is furnished by a previously investigated problem that arose in connection with an attempt to determine the variation of certain optical characteristics of clouds near their edges.[13] These determinations were to be made from existing data obtained by probing clouds with short laser pulses and measuring the backscattered laser radiation. For these probings, in which multiple scattering effects could usually be ignored, the theoretical relationship between the optical characteristic of interest C, the laser pulse P, and the measured backscatter signal V, is adequately expressed as follows:

$$V(t) = \int_0^\infty P(t - t')R(ct'/2)C(ct'/2)\,dt', \tag{11}$$

where R is a function describing the optical receiver's relative sensitivity to backscatter occurring at various distances, and c is the speed of light.

[13] D. McGuire and M. Conner, *The Deconvolution of Aerosol Backscattered Optical Pulses to Obtain System-Independent Aerosol Signatures*, Harry Diamond Laboratories, HDL-TR-1944 (June 1981).

Both R and C are functions of x, the distance from the transceiver to the scatter point along the laser beam. On the other hand, both V and P are functions of the time, $V(t)$ being the measured return-signal pulse and $P(t)$ being the instantaneous transmitted laser power. The return signal $V(t)$ is essentially a measure of the value of $C(x)$ at various x, but with the spatial resolution of the probing laser pulse. To see this, consider equation (11) on the approximative assumption that $P(t-t')$ is proportional to a Dirac δ-function of the same argument; then

$$V(t) = P_0 R(ct/2)C(ct/2), \tag{12}$$

where P_0 is the mean transmitter power for a single laser pulse. The dependencies on t in this equation can be translated into dependencies on the distance x as follows:

$$V(2x/c) = P_0 R(x)C(x), \tag{13}$$

where the translation between time and distance is via $t = 2x/c$; i.e., the time that goes with a given distance is the round-trip time of a light signal travelling that distance. Thus the measured return pulse $V(t)$, with its argument expressed in terms of x, gives us the variation of $C(x)$ upon division by $P_0 R(x)$, both of which are presumably known characteristics of the transceiver. The two limitations on the accuracy of this measurement are the noise in the measured signal and the spatial width of the probing transmitter pulse; instead of acting as a precise delta function, the actual width of $P(t-t')$ in (11) gives an approximate version of equation (13), one that is valid for an averaged or smeared out version of $R(x)C(x)$, where the averaging or smearing interval is the spatial width of P.

Because the spatial width of $P(t)$ was about 2 m, and some cloud edges appeared (from other data) to change significantly over much shorter distances, the investigation in question sought to deconvolve equation (11) as an alternative to accepting the inadequate resolution of equation (13).

Consider the Fourier transform of equation (11). We have

$$\hat{V}(f) = \hat{P}(f)\hat{h}(f), \tag{14}$$

where f denotes frequency, RC has been abbreviated by h, and the caret indicates a Fourier-transformed function. Thus, apart from the

noise in $V(t)$, we can obtain $R(x)C(x)$ exactly as the inverse Fourier transform of $\hat{V}(f)/\hat{P}(f)$; this implies that we can determine $C(x)$ with infinitesimal resolution by dividing that inverse transform by $R(x)$. This may seem too good to be true, but there is absolutely nothing wrong with the sketchy argument just presented. Even the existence of zeros of $\hat{P}(f)$ cause the argument no difficulty, because the ratio, $\hat{V}/\hat{P}$, remains finite at such frequencies (in fact, it remains continuous). Problems do arise, however, from the combined presence of these zeros and noise, because then the ratio of $\hat{V}$ to $\hat{P}$ at the zeros of $\hat{P}$ is the nonzero noise level divided by zero: i.e., the ratio is infinite. This, moreover, gives rise to a catastrophic problem, because it is easily shown that any one of these infinities will make the integral that defines the inverse Fourier transform diverge. In this divergence we find the primary symptom of inverse-operator discontinuity for convolution operators. The only apparent alternative is to band-limit the taking of the inverse Fourier transform to within the smallest frequency at which $\hat{P}$ vanishes, but this approach leaves us with a nonzero resolution interval whose precise value depends on the shape of the transmitter pulse.*

*See reference 13 for further details.

7 Multiple Deconvolution

A less apparent approach to the problem in section 6 is due to Berenstein et al,[2,3,14,15] who considered the following general problem. Suppose that we are interested in measuring some physical property f, which is distributed over a manifold with coordinates $(x, y, ...)$, so that $f = f(x, y, ...)$. In practice the manifold is typically either the time continuum—so that $f = f(t)$ in this case—or the two-dimensional manifold over which an image, perhaps an infrared or a synthetic aperture radar image, is distributed—so that $f = f(x, y)$ in this case. It is proposed that we perform this overall measurement by performing a number, say n, of subsidiary and *different* convolutional measurements of f, namely,

$$g_i(x, y, ...) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \mu_i(u - x, v - y, ...) f(u, v, ...)\, du\, dv..., \quad (15)$$

where $\mu_i(u, v, ...)$ is the *convolution kernel* for the i^{th} subsidiary measurement, whose result is given by $g_i(x, y, ...)$. Note that we are here dealing with the use of generally multidimensional convolutional measurements, all of which have the same dimension as the manifold over which the physical quantity of interest is distributed. Given such a set of measurements, i.e., given the kernels μ_i, we seek corresponding *deconvolvers* ν_i such that

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \nu_i(x - u, y - v, ...)\mu_i(u, v, ...)\, du\, dv... = \delta(x, y, ...). \quad (16)$$

In other words, we want the sum of the n convolutions $\nu_i * \mu_i$ to produce a Dirac delta function centered on the origin of our multidimensional manifold $(x, y, ...)$. If we can find such deconvolvers for a given set of convolution kernels, it is rather straightfoward to show that the system of equations (15) can be solved for $f(x, y, ...)$ in the form

$$f(x, y, ...) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \nu_i(x - u, y - v, ...) g_i(u, v, ...)\, du\, dv..., \quad (17)$$

so that our measurements g_i and the deconvolvers ν_i are sufficient to construct f exactly, apart from the noise in our measurements.

A solution like (17) would not suffer from the noise catastrophe that befell the similar attempt (sect. 6) to find $C(x)$ exactly. In attempting

[14] C. A. Berenstein and A. Yger, *Le problème de la déconvolution*, J. Funct. Anal. **54** (1983), 113-160.
[15] C. A. Berenstein and A. Yger, *Analytic Bezout Identities*, Adv. Appl. Math. **10** (1989).

to apply the current method to the previous example, we might for instance have chosen to measure the aerosol backscatter signal with two differently shaped transmitter pulses, $P_1(t)$ and $P_2(t)$, and have obtained the two measured backscatter pulses

$$V_1(t) = \int_0^\infty P_1(t-t')R(ct'/2)C(ct'/2)\,dt', \tag{18}$$

and

$$V_2(t) = \int_0^\infty P_2(t-t')R(ct'/2)C(ct'/2)\,dt'. \tag{19}$$

Supposing that deconvolvers ν_1 and ν_2 corresponding with P_1 and P_2 could be found, we would then have been able to determine $C(x)$ exactly from

$$R(ct/2)C(ct/2) = \int_0^\infty \nu_1(t-t')V_1(t')\,dt' + \int_0^\infty \nu_2(t-t')V_2(t')\,dt'. \tag{20}$$

Because this equation simply entails convolving measured pulses with well-behaved* deconvolution kernels, it should be clear that the noise accompanying the measured pulses would not cause catastrophic problems; in fact, there would even be some smoothing of the noise transferred to the constructed function RC.

This proposed method calls forth a question: *When do deconvolvers exist?* Deconvolvers exist if and only if the convolution kernels μ_i form a *strongly coprime* set. The strict technical definition of this term is not particularly revealing to any but mathematical experts; however, an easily appreciated necessary condition for the existence of deconvolvers can be seen by taking the Fourier transform of the defining equation for the ν's—equation (16). One then gets the *Bezout equation*:

$$\sum_{i=1}^{n} \hat{\nu}_i\hat{\mu}_i = 1. \tag{21}$$

Because the convolutions and Fourier transforms involved here are generally multidimensional, one must think of the above transforms of the ν's and μ's as depending on a multidimensional wave vector, say $\vec{k}$, rather than on the usual one-dimensional frequency ω. In other words, the last equation can be written out more fully as

$$\sum_{i=1}^{n} \hat{\nu}_i(\vec{k})\hat{\mu}_i(\vec{k}) = 1, \tag{22}$$

*I have not mentioned the technical point that when deconvolvers exist they are necessarily *distributions of compact support*, which, in the one-dimensional example being considered, essentially means that they are continuous functions that vanish outside a bounded interval.

and is understood to hold for all $\vec{k}$. But if this equation is to hold throughout $\vec{k}$-space, it is evident that there can be no $\vec{k}$ at which all the $\hat{\mu}_i$ vanish, for at such a $\vec{k}$ our last equation would read $0 = 1$. Therefore, a necessary condition for the existence of deconvolvers is that the Fourier transforms of the convolution kernels do not possess any common zeros—recall that it is precisely the zeros of the Fourier-transformed convolution kernel that give rise to the catastrophic noise problems already noted in the previous direct approach to deconvolution. The stronger necessary and sufficient condition (that the convolution kernels form a strongly coprime set) requires the nonexistence of common zeros, as above, plus a little more. This "little more" has to do with how close together the separate zeros of the individual $\hat{\mu}$'s can get, principally in the far reaches of $\vec{k}$-space.

Strongly coprime sets of convolution kernels exist in great abundance and are relatively easy to find; moreover, it would be possible, without great difficulty, to implement and perform sets of strongly coprime measurements in practice. This, in turn, raises the question of finding the associated deconvolvers, as these would be needed for the explicit interpretation of the data obtained by such measurements. Here too the issues are well understood. Berenstein and his collaborators have developed explicit formulas[2] for the computation of the deconvolvers from the convolution kernels, and although these computations are generally complex, they can be done off-line and once and for all for a given experiment.

As a final point let us clarify the most relevant practical limitations of this technique. We can regard the method, at least partly, as a form of data processing that can achieve what amounts to a superresolved convolutional measurement. In image processing, this method would aim to enhance the resolution of images, which are generally diffraction limited by the aperture size of the imager's collection system, and are otherwise limited by such things as the nonzero areas of the imager's radiation detectors. For these factors, the critical feature of strongly coprime kernels is that they must be *distributions of compact support* or, less strictly, must vanish outside some bounded closed region of the image plane. Now this property is indeed possessed by the kernels that correspond to the finite areas of radiation detectors, such as are used in the image plane of infrared viewing systems, for example. The kernels here are proportional to the sensitivity of the detectors within their active areas and vanish outside these areas. On the other hand, the

convolution kernels that correspond to the aperture sizes of collecting lenses or radar antennas do not possess this property. The kernel that describes the capture of radiation by a lens is the point-spread function of the lens, which is never a function of compact support—for example, the point-spread function of a perfect circular lens is the well-known *Airy pattern*, which has nonzero values almost everywhere in the image plane. Let us leave it at this: the multiple-deconvolution method can be profitably used in image processing problems and can achieve results that at first sight seem counterintuitive; however, it cannot defeat any of the physical limitations on resolution that arise from the diffraction limit and its analogs.

8 Example of the Regularization Theory Approach

The problem of image flow is a practically relevant example of the underdetermined side of the ill-posedness issue. A typical example of this problem is the following. Consider a rigid circular hoop rotating about one of its diameters with constant angular velocity. If we project this motion onto a nearby plane that is parallel to the axis of rotation, the resulting motion observed in the plane will be that of a simple closed contour which exhibits a periodic transfiguration that ranges from a circle through a continuum of intermediate elliptical forms and back to the original circle. If all we saw was the projected contour, and had no prior knowledge of the actual motion producing the observed motion, except that it was the projected motion of a rigid body in the shape of a closed curve, how might we proceed to gain some further knowledge about the cause of what we see? The first matter we might try to resolve is the question of the true image flow.

The true image flow can be described as follows. At every instant, each point on the moving material loop has a certain velocity vector, say $\vec{v}$. The only component of $\vec{v}$ that is evident in the plane projection of the motion, however, is that parallel to the observing plane. But we cannot precisely determine even this by observing the changes in the figure of the projected contour over a small time interval (which is all that we can observe), because in observing any two distinct contours we do not know which point on one contour moved to which point on the other. The field of the parallel-velocity components over the full extent of the observed contour is called the *image flow* (the *optical* flow in the case of optical images). Consequently, the plainly underdetermined problem of intelligently estimating this parallel-velocity field is called the problem of image or optical flow.

The solution to this problem can be regarded as a first step in the determination of the motions of bodies in a scene. Both regularization-theoretic and MAP techniques—the latter with neural networks—have been successfully investigated for the determination of the image flow.[8,9] The human flavor of these solutions is nicely conveyed by the fact that among the numerous correct answers that these techniques produce, there are also erroneous ones, "optical illusions" that often coincide with well-known human optical illusions: for instance, the so-called "barber-pole illusion," where a helix rotating about its axis is seen to be moving up that axis.

Let us consider a regularization-theoretic solution of this problem to see how these methods are concretely used in matters of practical relevance. As already said, having the contour at two closely separated times would be enough to determine the image flow at that time, if we knew the right correspondence between points on the two contours. Nothing in the immediate visual evidence tells us this, however. To solve this problem, the regularization theorist uses the directions normal to the earlier contour to compute an estimate of the normal component of the image flow. The unknowns are then the corresponding tangential velocity components. He then assumes that the actual motion is that of a rigid body having a relatively smooth surface. This leads to the implied constraint that the associated velocity field is also *relatively smooth in the differentiable-manifold sense.* His problem thus becomes the formulation of this smoothness constraint in the structure of a specific stabilizing mapping, P. Letting V denote the image-flow velocity vector, the following simple stabilizing operator has been found quite adequate for a number of cases:

$$\|P(V)\|^2 = \int \left(\frac{\partial V}{\partial s}\right)^2 ds,$$

where s is the arclength along the contour, and the integral is taken along that contour. The mapping denoted by F in our previous general discussion is taken to be $F(V) = V \cdot N$, where N is the unit normal to the earlier contour, pointing from it toward the later one, and "$\cdot$" indicates the usual vector dot product. Formulation (8) is used with this P and F when the normal-velocity estimates are deemed so good that we can replace

$$\|V \cdot N - V_N\| < \epsilon$$

with $V \cdot N = V_N$, where V_N denotes the estimated normal velocity. When this is deemed insufficiently accurate, formulation (9) is used instead. Both formulations can be shown to have unique solutions. When (9) is used, the reciprocal of the emergent value of the regularization parameter, z, indicates the reliability of the data. The image flows obtained with either of these formulations are essentially approximating vector splines.

This formulation of the image-flow problem can be easily implemented and efficiently solved by computer (comparatively speaking) because of the linearity of F and the simple form of P as a linear differential operator. This comparative computational simplicity and efficiency

are not usually met with in the solution of feature-extraction problems, however, which brings us to the final problem area of the first-stage ATR task: how to efficiently implement solutions to well-formulated feature-extraction problems. Most of the difficulties we meet with here are quite familiar.

These difficulties are the usual ones associated with complex extremal problems in many variables: obtaining an effective and computationally feasible solution algorithm in view of such stumbling blocks as the existence, typically, of many local extrema. We can often develop or select an effective solution algorithm (though even this is sometimes quite difficult), but it often happens that its computational complexities and volume render it impractical. Usually, success with restructuring an algorithm to render it machine practicable hinges on making it more amenable to parallel processing, on using statistical computational techniques of the Monte Carlo genre, or both. I now turn to a recent development which engages all these issues.

9 Stochastic Relaxation and Annealing

One of the most striking and promising developments in the field of MAP estimation has been the recent emergence of the stochastic relaxation and annealing (SRA) technique of Geman and Geman.[7] This is a model-based approach to the problem of image restoration and feature extraction that ambitiously attempts to simultaneously deal with obtaining effective algorithms that produce manageable computational loads. It (a) provides a rigorous technique for the solution of a broad class of MAP estimation problems, and (b) produces a relatively parallel, statistical algorithm of the Monte Carlo type.

Because of its comprehensiveness and complexity, the work of the Gemans is given more intensive coverage than the previous topics. Section 9.1 discusses a general scheme and procedure which

1. starts from a *prior* probability distribution of a composite random field whose components are the random field of the restored (deblurred and noiseless) image to be determined, and a random field describing the feature or set of features (e.g., bodily outlines and textures) one would like to extract,
2. proceeds by a Bayesian construction to obtain the *posterior* probability distribution of the same composite random field *conditioned on the raw image obtained*, and
3. concludes by finding the state of the composite field that *maximizes* the posterior distribution and hence finds the most likely restored image and set of features consistent with the prior probability distribution and the raw image.

This last step is what the SRA algorithm is designed to perform. Its sole limitation is that it can operate only on the class of posterior distributions called *Gibbs distributions*. In this connection it is important that so-called *Markov random fields* (MRF's) and Gibbs distributions are rigorously equivalent in a certain sense, and that prior distributions which are MRF's necessarily give rise through Bayes' rule to posterior distributions that are also MRF's. Thus a Markov random field model of the prior distribution will give rise to a posterior distribution suited to the SRA algorithm.

Section 9.2 gives the formal definitions of Markov random fields and Gibbs distributions, stating their equivalence in a rigorous way. Section

9.3 describes the SRA algorithm, and succeeding sections describe how one sets about to obtain the explicit Gibbsian form of the posterior distribution required as input to the SRA algorithm.

9.1 MAP Estimation and Image Processing

Stochastic relaxation and annealing is a statistical computational means for finding the state of maximum probability of a Gibbs distribution. This maximization problem arises in the context of a MAP estimation approach to the following general image processing problem. Given an image, find a processing scheme that will simultaneously reduce noise and blurring, and as well extract such primitive features of the scene presented as the outlines of bodies, the textures of backgrounds, etc.

In this overall MAP estimation approach,[7] the composite of the image of the scene being viewed and its primitive features is regarded as a stochastic process X of the following form:

$$X = \{x_{ij}, l_{pq} : (i,j) \in Z_m, (p,q) \in D_m\}, \tag{23}$$

where Z_m and D_m are two-dimensional integer lattices, and the x_{ij} and l_{pq} are random variables associated with the corresponding lattice sites. The lattice Z_m is $m \times m$ and labels the pixels in the image, so that the x_{ij} represent the image intensities associated with the various pixels. The pixel intensities are typically in the form of an integer grey scale, but could just as well describe a color image where each x_{ij} is a vector whose components correspond to the various color elements. The lattice D_m is called the *feature lattice*. It and its associated field of random variables, l_{pq}, are best described through an example.

If we are attempting to outline the objects presented in an image, it is convenient to let Z_m be the set of pixel midpoints, and then let D_m be the set of points that are midway between all the adjacent vertical and horizontal pairs of pixel midpoints. We then associate a variable *line element* with each of the points in D_m. The presence and various admissable tilts of each line element can be represented by the values of some modular integer variable l_{pq}. For example, l_{pq} might be a binary variable where 0 indicates that no line element is present and 1 indicates the opposite. In the latter case, if (p,q) is a feature lattice site between a vertical pair of pixel sites, then the line element is understood to be horizontal; otherwise it is taken as vertical. This is just one example of

a body-outlining feature-lattice scheme with a clear generalization to one in which more than two tilt angles are used.

In the general situation, D_m is a lattice associated with the pixel lattice Z_m in some way (such as the foregoing), and the l_{pq} are random variables, associated with the sites of this feature lattice, that describe the image features of interest in accordance with some scheme (again such as the foregoing). In addition to the body-outlining schemes alluded to, there are others that describe textural features by means of a D_m with the same number of sites as Z_m and a texture label field variable.[16] It should be clear that the feature-describing potential of D_m and its associated lattice field depends primarily on our ingenuity.

The raw image that we obtain is similarly specified by its set of corresponding intensity levels,

$$\{x'_{ij} : (i,j) \in Z_m\},$$

which represent a generally blurred and noisy image of the scene of interest. Let us briefly denote this collection by x', and likewise the collections x_{ij} and l_{pq} by x and l, respectively, thinking of x and l as two variable random fields over their respective lattices. The primary objective is to determine the joint conditional probability function $P(x, l|x')$, which gives the probability of various images x, and various feature configurations l, conditioned on the pixel intensities x' of the raw image. Note that x represents a variable image that is understood to be *noiseless* and *unblurred*; that is, one of the goals of the scheme under consideration is to use MAP estimation techniques to obtain a most probable noise-reduced and deblurred version of the raw image. The other goal is to obtain the most probable configuration of the feature field l_{pq}. Part of the difference between this rather ambitious use of MAP estimation and the simpler one described in section 3 lies in the first-mentioned goal. In the simpler case a feature vector alone maximizes the posterior density, whereas in the present more ambitious case the posterior probability, $P(x, l|x')$, is maximized by both a new image vector or array, x, and an array, l, of feature indicators.

In this general MAP estimation scheme, we begin with an *a priori* joint probability density, $P(x, l)$, which represents our prior knowledge (up

[16] S. Geman and C. Graffigne, *Markov Random Field Image Models and Their Applications to Computer Vision*, Proc. International Congress of Mathematicians, Berkeley, CA (1986), 1496–1517.

to obtaining the raw degraded image) of what to expect in the way of noiseless unblurred images x and feature configurations l. We wish to determine the joint conditional probability $P(x, l|x')$, and then find the x and l that maximize this conditional probability. This will be the best one can do to restore the degraded image x' as x, and to estimate the primitive features present in the image as l. To determine $P(x, l|x')$ we use Bayes' rule in the form

$$P(x, l|x') = \frac{P(x'|x, l)P(x, l)}{P(x')}.$$

Here we presumably know the prior $P(x, l)$, can eliminate $P(x')$ in the maximization process, and must determine the reverse conditional probability $P(x'|x, l)$ from what we know.

We are thus dealing with two stochastic processes, X_b and X_a, which refer, respectively, to the statistical behavior of the composite random field $\{x, l\}$ before and after the raw image is obtained. We must now make some general assumptions about these processes in order to carry out the proposed program of determinations in a general way. The least restrictive assumption that has been found to make this program possible is to assume that X_b is a so-called *Markov random field* (MRF). Three points should be re-emphasized before I give the technical definition of an MRF: (1) Bayes' rule guarantees that X_a will be an MRF when X_b is; (2) the statistics of an MRF always admit a mathematical description in terms of a Gibbs distribution; and (3) the stochastic relaxation and annealing algorithm is aimed directly at the problem of maximizing Gibbs distributions. Therefore, since X_a necessarily has a Gibbs distribution description (because of our general assumption that X_b is an MRF), we will be able to apply the SRA technique to finding the maximum of the posterior conditional $P(x, l|x')$ once we have succeeded in establishing its Gibbsian form. The rest of my discussion of Geman and Geman[7] focuses on obtaining this Gibbsian form and describing the SRA algorithm. First, however, I discuss MRF's and Gibbs distributions, and their equivalence.

9.2 Markov Random Fields and Gibbs Distributions

Loosely speaking, an MRF on a lattice—$Z_m \cup D_m$—is a random field that exhibits only local correlations; i.e., there is some "radius" about each lattice site, whether a pixel or a feature site, such that there is significant correlation between the field value at this site with only the field values at sites within the "radius." This idea is made more precise with the concept of a *lattice neighborhood system*, which is a formal way of specifying the lattice sites that are "in the neighborhood" or within the "radius" of a given lattice site, for all the lattice sites.

Let us use a more convenient notation for the full lattice $Z_m \cup D_m$. Let this lattice have N sites altogether, where N equals m^2 plus the number of sites in the feature lattice. Let $S = \{1, 2, ..., N\}$ be some fixed enumerative labeling of all these sites, so that if $s \in S$, then s points, once and for all, to a particular one of the sites in the full lattice. Then let y_s stand for the random field variable associated with the site s, which could be either an x or an l, depending on whether s indicates a pixel site or a feature site. In this way we can denote our random field, X_b or X_a, by the new notation $Y = \{y_s : s \in S\}$. A *neighborhood system on the lattice* S is defined as follows. A neighborhood system on S is a collection, G, of subsets $G_s \subset S$, one for each site s in the lattice, such that

1. s does not belong to its own G_s, which is called the *neighborhood of* s or the *set of* s's *neighbors*, and
2. s is a neighbor of r if and only if r is a neighbor of s; i.e., the relation of being a neighbor is symmetric.

A subset $C \subset S$ is called a *clique* if all the distinct pairs of sites in C are neighbors. The set of all S's cliques is denoted by C'.

There is of course an underlying probability space, Ω, of elementary events, ω, on which we have some kind of probability measure P. Also, each y_s is a variable in some definite variable space denoted R_s. For instance, in our case, when s is a pixel site, R_s might be defined as the set of integers $\{0, 1, ..., k\}$, where k is the maximum value of the pixel grey scale; on the other hand, if s is a feature site, R_s might be defined as the set $\{0, 1\}$ describing the presence or absence of line elements, as discussed before. We can take the elementary events, ω, of the underlying probability space, Ω, to be the various distinct *configurations*

$$(y_1, y_2, ..., y_N) = \omega \tag{24}$$

of the random field Y, where each y_s ranges over its associated R_s.

We can now give the formal definition of an MRF. A random field Y is said to be a **Markov random field** relative to a neighborhood system G if the following two conditions hold.

1. The probability $P(y_1, y_2, ..., y_N)$ of the configuration $(y_1, ..., y_N)$ is everywhere positive on Ω.
2. For each $s \in S$ we have

$$P(y_s|y_1, ..., y_{s-1}, y_{s+1}, ..., y_N) = P(y_s|y_r : r \in G_s),$$

where the symbol to the right of the conditioning bar on the right-hand side of the equation means "all y_r such that r is a neighbor of s," and where all the y's are understood to vary over the whole of their domains R_s. In other words, the conditional probabilities on the left of the equation above are independent of any y's that are not associated with sites in the neighborhood of s.

These conditional probability functions are called the *local characteristics* of the MRF. The MRF concept is just one of several ways which have been devised to generalize the more familiar notion of a *Markov chain* to multidimensional lattices.

There is an equivalence between MRF's and the random lattice fields defined by Gibbs distributions, and this equivalence turns out to be of great practical value for the analysis of MRF's. Because Gibbs distributions have been used in physics for some time (in statistical mechanics), much of the general terminology used to discuss such distributions has been borrowed from physics, as can be seen in the following formal definition.

As before, we have a lattice, S, a neighborhood system, G, and a random field Y on S. In this setting a Gibbs distribution relative to G is a probability measure p, defined on Ω, which has the following form:

$$p(\omega) = \frac{1}{Z} \exp[-U(\omega)/T]. \tag{25}$$

Here Z and T are constants; U is called the *energy function* and has the form

$$U(\omega) = \sum_{C \in C'} V_C(\omega), \tag{26}$$

where each V_C is a function on Ω that depends on only those coordinates y_s of ω for which s is in the clique C. Such a family, $\{V_C : C \in C'\}$, is called a *potential*. The parameters T and Z are called the *temperature* and *partition function*, respectively, following the terminology of statistical mechanics. The term partition *function* is used for the *constant* Z because the sum of $p(\omega)$ over Ω must be unity, since p is a probability measure; this leads to the identity

$$Z = \sum_{\omega} \exp[-U(\omega)/T], \tag{27}$$

whose right-hand side will be recognized by physicists as the *canonical partition function* of statistical mechanics.

The connection between MRF's and Gibbs distributions is given by the following.

Theorem. The random field Y is an MRF relative to a neighborhood system G if and only if $P(\omega)$ is a Gibbs distribution relative to G.

This means that for an MRF the joint probability distribution $P(\omega)$ is given by the right-hand side of (25) for some potential $\{V_C : C \in C'\}$ and some constant temperature T. Thus, if we are referring to the MRF X_a (the posterior case), so that $P(\omega)$ is actually $P(x, l|x')$, then the global maximization of the latter is the same problem as finding the configuration that renders U a global minimum, because of the nature of the exponential function. To solve this problem we can use methods that were developed some time ago in statistical mechanics.[17]

[17] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equations of State Calculations by Fast Computing Machines*, J. Chem. Phys. **21** (1953), 1087–1091.

9.3 Stochastic Relaxation and Annealing Algorithms

Let us assume that we have already modeled the poster or Markovian probability function $P(x, l|x')$ with a Gibbs distribution; more specifically, suppose that we have imposed a certain neighborhood structure on the lattice, relative to which a specific modeling potential, $\{V_C : C \in C'\}$, has been constructed, and finally that the temperature parameter, T, needed to make the model distribution fit, is known (it will turn out that we do not really need to know T).

Our first need is to establish formulas for the calculation of the local characteristics of an MRF in terms of the corresponding potentials of its Gibbs distribution and the temperature. These formulas are quite easy to obtain from the Gibbsian expression for $P(\omega)$ and the definition of conditional probability. One gets

$$P(y_s|y_1, ..., y_{s-1}, y_{s+1}, ..., y_N) = Z_s^{-1} \exp\left[-\sum_{C:s\in C} V_C(\omega)/T\right], \quad (28)$$

where

$$Z_s = \sum_{y\in R_s} \exp\left[-\sum_{C:s\in C} V_C(\omega^y)/T\right],$$

and the coordinates of ω^y coincide with those of ω except in the s^{th} position, i.e., the y_s of ω^y is y, which is understood to be a variable ranging over all of R_s. An important point about this result, in explicit agreement with condition (2) above, is that the local characteristic at s depends on only y_s and those y_r with $r \in G_s$, because any site in a clique containing s must be a neighbor of s—the notation $C : s \in C$ means that the pertinent sums are over all cliques containing s. Thus the indicated calculation, for any given site, is merely a local one.

The relaxation algorithm, described next, makes repeated use of the local computation indicated by (28) to relax from an arbitrary starting configuration to a limiting configuration with the probability density $P(\omega)$. The precise technical meaning of the last phrase is given in the *relaxation theorem*, which follows the description of the algorithm.

Relaxation Algorithm. As before, I continue to use the general notation

$$Y = \{y_s : s \in S\}$$

for the random field. We discretize time according to $t = 1, 2, 3, ...$, and suppose an arbitrary starting ($t = 0$) configuration $Y(0)$. Let us envision that a separate processor is associated with each site of the pixel/feature lattice. Each such processor is connected to the processors associated with its neighboring sites and to no others. Let us further envision that a sequence

$$Y(0), Y(1), Y(2), Y(3), ...$$

of configurations of the random field will be produced by this interconnected system of processors, starting from the first, $Y(0)$, where the arguments $1, 2, 3$, etc, correspond to the above discrete time sequence. This is done by (possibly) changing the field at only one site at each time in the sequence, where the sequence of site visitations is determined by a previously chosen sequence, $\{s(t) : t = 1, 2, 3, ...\}$, of site locations. This unending sequence must be such that it revisits each site infinitely often, so that the lattice-path described by $s(t)$, as t increases without bound, will be a discrete analog of a Peanc space-filling curve. (An unending raster-scan pattern satisfies this requirement.) At each t in the discrete time sequence, the following situation occurs: $Y(t-1)$ has just been determined, and we are about to produce $Y(t)$ by (possibly) changing the value of $y_{s(t)}$. Every time another term is produced in this sequence of configurations, it will be necessary to update the configuration so that the changes become the current data available to each processor whose neighborhood has been affected. This is accomplished via the processor interconnections: a change at site s is communicated to all processors connected to s, i.e., all processors in s's neighborhood. The (possible) change of $y_{s=s(t)}$ is decided by sampling the pertinent local characteristic in accordance with equation (28). Thus each processor is programmed to make the computation specified by that equation, and executes this program each time its site is visited, according to $s(t)$. This computation is local in the sense that it involves only the field data that are associated with $s(t)$ and the sites in its neighborhood $G_{s(t)}$; moreover, because of the updating process, all these required data are immediately available to the processor. This relaxation algorithm is thus local in the sense described, relatively parallel, and statistical, since it entails repeatedly sampling the steadily updated local characteristics.

Relaxation Theorem. For any starting configuration, $Y(0)$, and any $\omega \in \Omega$, we have

$$P[Y(t) = \omega] \longrightarrow P(\omega) \text{ as } t \longrightarrow \infty.$$

The meaning of $P[Y(t) = \omega]$ needs clarification. Suppose that we generate a representative ensemble of configuations, $Y(t)$, by running the relaxation algorithm on a sufficiently broad and numerous set of starting configurations, $Y(0)$. Then $P[Y(t) = \omega]$ means the relative frequency of occurrence of the configuration ω in this ensemble. The theorem's statement can therefore be rephrased to say that we can generate a representative ensemble of configurations whose probability density coincides with $P(\omega)$ by carrying the process just described to the limit of indefinitely increasing t. Alternatively and less strictly, by picking just one starting $Y(0)$ at random, we can generate, in the $Y(t)$ that results, a good approximate sample of the distribution $P(\omega)$ by choosing t large enough. The relaxation theorem thus reveals that the relaxation algorithm provides an implementable way to sample $P(\omega)$. This result was established by Geman and Geman.[7]

A slight modification of the relaxation algorithm produces the more directly useful SRA algorithm, for which an associated convergence theorem can be proved.[7] For these it is convenient to define the quantities

$$U^* \equiv \max_{\omega} U(\omega),$$

$$U_* \equiv \min_{\omega} U(\omega),$$

and

$$\Delta \equiv U^* - U_*,$$

where the *max* and *min* are taken over all $\omega \in \Omega$. We also let

$$\Omega_0 = \{\omega \in \Omega : U(\omega) = U_*\}$$

(i.e., Ω_0 is the set of all configurations having the minimum energy), and define π_0 as the uniform probability density on Ω_0, so that all configurations without minimum energy have probability zero with respect to π_0, while the remaining configurations with minimum energy all have the same nonzero probability relative to π_0.

SRA Algorithm. To introduce "annealing" into the relaxation algorithm, we need only a prior schedule of temperatures, $T(t) : t = 1, 2, ...$, which are to be used in sequence at the appropriate times, i.e., when sampling the local characteristics. In other words, at $t = i$ we use $T(i)$ for T in equations (28) when we draw a sample at $s(i)$. The temperature sequence, which is taken as slowly decreasing, is literally an annealing schedule, because it is being used in the attempt to drive the resulting sequence of configurations toward one which renders the energy U a global minimum. Note that the latter would constitute a successful maximization of $P(x, l|x')$. The theorem supporting this algorithm is as follows.

SRA Convergence Theorem. Assume there exists an integer $\tau \geq N$ such that for every $t = 0, 1, 2, ...$, the set of sites

$$\{s(t+1), s(t+2), ..., s(t+\tau)\}$$

includes all the sites in S. Let $\{T(t) : t = 1, 2, 3, ...\}$ be any decreasing sequence of temperatures (with limit zero) such that for some integer $t_0 \geq 2$ we have

$$T(t) \geq \frac{N\Delta}{\log t} \tag{29}$$

for all $t \geq t_0$. Then, for any starting configuration and any $\omega \in \Omega$, we have

$$P[Y(t) = \omega] \longrightarrow \pi_0(\omega).$$

Again the meaning of the left-hand member is in terms of the representative ensembles, for various t, that can be generated with the SRA algorithm. This theorem therefore asserts that the relative frequency of a given configuration in the limiting asymptotic ensemble is exactly zero whenever the configuration has greater than the minimum energy. Thus, if we start from a lone arbitrary initial configuration and take t very large, the chances are excellent that the resulting $Y(t)$ will be a configuration of globally minimum energy, and thus a solution of the maximization problem for $P(x, l|x')$.

This remarkable result tells us that all we need of the posterior distribution's model is a neighborhood system and its associated energy function; with these alone, and the relatively parallel and local statistical SRA algorithm described above, we can compute the most likely x and l for the degraded image x'. A few things need to be said about this

result (and also about the relaxation theorem), but let us first observe that with these results, the issue shifts to that of modeling the G and U of the posterior density (for whatever concrete problem we wish to solve). So let us briefly consider, in general terms, how this modeling can be done, so as to bring out the remaining elements of the overall method and show how they all fit together.

9.4 Modeling the Posterior Distribution

The modeling begins with the *a priori* situation, before the degraded image is obtained. Based on the modeler's prior knowledge, the Gibbsian form of the prior density, $P(x, l)$, must first be established; that is, we must impose a neighborhood system, G, on the full set, $S = Z_m \cup D_m$, of lattice sites, and we must likewise impose a corresponding energy function, $U(x, l)$, on this neighborhood system. We must also determine the best-fit T for the resulting Gibbs distribution, but this aspect is often handled by scaling to make $T = 1$. After this generally difficult task is done, construction of the Gibbsian model of the posterior density, $P(x, l|x')$, begins. This first requires that we model the image formation process. In general, the visible image matrix, x', is related to the undegraded image matrix, x, by a formula like

$$x' = Q[H(x)] + N_d, \tag{30}$$

where H denotes the typically convolutional blurring function associated with the viewing system's optics, Q denotes the possibly nonlinear transformation that the system's detector performs on $H(x)$, the output of the optics, and N_d is the detector noise process. The noise is indicated as additive, but it could be multiplicative or some other kind; whatever the case, it should be appropriately incorporated into the above formula. This image formation model and the degraded image are next used to construct the neighborhood system and energy function of the posterior density from the corresponding G and U of the prior density. Geman and Geman[7] establish a rather general theorem pertaining to this construction (discussed in sect. 9.6). The main point about any such construction is that it allows us to incorporate our knowledge of H, Q, N_d, and x' into the modified neighborhood system and energy function for the posterior density. The final step is to run the SRA algorithm on the posterior model to obtain the MAP estimate of the most likely x and l.

9.5 Discussion of the Convergence Theorems

Let us first clarify a technical point of the SRA convergence theorem: its requirement for the existence of the integer τ. This simply means that there must be a certain minimum fixed length of the site-visitation sequence, which, when moved along that sequence to any arbitrary registration, will always encompass all the sites in the lattice. A raster-scanning pattern certainly satisfies this requirement, which in practice generally presents no problem. The same cannot be said of the requirement that the annealing temperature decrease at no faster than an inverse-logarithmic rate with t. This very slow rate of decline often makes the execution of the algorithm impractical.

The computational content of the relaxation theorem is closely analogous to the statistical-mechanical computation of the thermodynamical equations of state of physical systems—especially those with lattice symmetry—where the physics is based on the canonical-ensemble partition function, i.e., where the system is in a constant-temperature heat bath. This analogy can be outlined as follows. Both are based on formulating the statistical features of their respective problems in terms of Gibbs distributions. In the relaxation theorem, the problem addressed is the sampling of a complex Gibbs distribution; for statistical mechanics, the problem is to calculate the interesting thermodynamical properties of the system, a problem that always amounts to finding the average value (relative to the Gibbs distribution) of the system's internal energy as a function of the appropriate independent macroscopic variables. Though these problems appear quite different, the computational techniques for solving them are quite similar.

Before fast computing machines became available, the aim of statistical physics was to perform fully tractable theoretical calculations of the thermodynamic properties of interest. This, however, was seldom possible except in the simplest cases: systems of noninteracting particles or molecules (like the ideal gas), and those relatively weakly interacting systems that could be approached via expansions based on the ideal case as its zero-order term (like the near-ideal gas via the virial expansion). The notorious intractability of most statistical-mechanical calculations (and the difficulties of extracting useful results in closed form from apparently simple macroscopic systems) is strikingly illustrated by the case of the famous Onsager solution of the two-dimensional Ising model of ferromagnetism. The Gibbs distribution for this lattice problem gives rise to a canonical partition function which could almost not

be simpler: a two-dimensional cubic lattice with only nearest-neighbor spin-spin interactions to form the energy function. Nonetheless, the theoretical derivation of the ferromagnetic equations of state from this partition function (by Onsager's method and even by other somewhat simpler ones following Onsager) is an enormous labor.*

This Onsager derivation is indeed a kind of frontier post representing the limits of our ability to perform such theoretical analyses. The computer, however, has made achieving somewhat more modest goals rather commonplace. By statistically modeling the action of complex partition functions on the computer, we can obtain numerical representations of equations of state and even investigate such things as phase transitions. The so-called Monte Carlo methods that have evolved along these lines are arguably the most important statistical computational techniques we have today. These methods emerged in the fifties from the work of Metropolis et al[17] which addressed the basic issue of how to efficiently compute (canonical) statistical-mechanical averages of the general form

$$\langle f \rangle = \frac{\int P(\omega) f(\omega)\, d\omega}{\int P(\omega)\, d\omega}, \tag{31}$$

where f is the dynamical variable of interest and $P(\omega)$ is the relevant Gibbs distribution. Because the $d\omega$ in such computations stands for an exceedingly long string of separate differentials, the numerical computation of such multiple integrals by brute-force techniques is out of the question. The Metropolis method consists of replacing the brute-force method (which would entail evaluating f on a regular lattice of points in the space of ω, and then summing the appropriately weighted results to get an estimate of the integral) with one based on a random sampling of the values of f at a much reduced number of points. The problem with the brute-force method is that an impossibly large number of points are needed to get accurate results. With the Metropolis algorithm, however, the computation becomes possible because virtually only points that make a significant contribution to the integral are chosen, and a greatly reduced number of them will give good accuracy. The chosen points are those with a high probability according to $P(\omega)$, i.e., they are chosen by effectively sampling $P(\omega)$, which will tend to produce ω with high $P(\omega)$, and hence ω that make a relatively greater contribution to the integral. Since the relaxation algorithm of

*I once witnessed the unfolding of such a derivation over about a dozen consecutive lecture hours, a derivation characterized by the most compact reasoning. It was during about the eighth session that the lecturer identified the mode of demonstration: "proof by intimidation."

Geman and Geman[7] is an effective computational technique for sampling a Gibbsian $P(\omega)$, it should not be surprising that this algorithm is a close analog of the Metropolis algorithm.

9.6 Construction of Posterior Distribution

I now describe the explicit construction of $P(x, l|x')$ in its Gibbsian form from the elements thus far assembled. This final step provides the input for the stochastic relaxation and annealing algorithm whose output is the solution of the overall problem.

Given the neighborhood system and energy function, G and U, for the prior density $P(x, l)$, and given the blurring function H, the detector transformation Q, and the nature of the accompanying noise process, we may proceed as follows to construct a new neighborhood system G^P, and a new energy function U^P (relative to G^P), where G^P will be the appropriate neighborhood system and U^P the appropriate Gibbsian energy function for the posterior conditional $P(x, l|x')$. In what follows, assume that U is measured in units such that $T = 1$. As a consequence, the corresponding temperature parameter for U^P is also one, as will be seen. Let us begin with the construction of G^P, which is closely tied to the convolutional blurring function H.

The function H can be expressed in matrix form as follows:

$$H_{ij}(x) = \sum_{i',j'} H_{i-i',j-j'} x_{i'j'}, \tag{32}$$

where $H_{k,k'}$ is a $(2m-1) \times (2m-1)$ matrix of constant elements, and k and k' run from $-(m-1)$ to $m-1$. Generally, the matrix H has nonzero elements only within a certain "window" of its central element $H_{0,0}$. Equation (32) therefore asserts that the blurred value $H(x)$ at a given pixel site (i, j) is essentially a weighted average of x_{ij} and the corresponding x's at all the surrounding sites within H's window of (i, j). Take, for example, the case of a rectangular window, where all the matrix elements of H vanish when the magnitudes of k and k' exceed certain positive-integer constants giving the window's horizontal and vertical extents. In the square-window case where $H_{k,k'} = 0$ whenever either $|k|$ or $|k'|$ exceeds 1, the only nonzero values of $H_{k,k'}$ would be the nine in the center of the array, namely,

$$\begin{matrix} H_{-1,-1} & H_{-1,0} & H_{-1,-1} \\ H_{0,-1} & H_{0,0} & H_{0,1} \\ H_{1,0} & H_{1,1} & H_{1,-1} \end{matrix}$$

where $H_{0,0}$ weights x_{ij}, and the rest weight the "true" grey levels at the eight adjacent surrounding pixels. The nonzero values of $H_{k,k'}$ are of course positive, and the effect of (32) would be a true average if H were normalized to make the sum of its matrix elements equal to 1.

We can now use the blurring function window and the prior neighborhood system to construct a new system of neighborhoods for $Z_m \cup D_m$. For each pixel site s (i.e., for each $s \in Z_m$), let K_s denote the set of pixels in the blurring-function window around s. We then put

$$K_s^2 \equiv \bigcup_{r \in K_s} K_r. \tag{33}$$

The new neighborhood system $G^P = \{G_s^P : s \in S\}$ can now be defined as follows. If $s \in D_m$, put

$$G_s^P = G_s;$$

in other words, the neighborhoods of the feature sites remain unchanged. On the other hand, if $s \in Z_m$, put

$$G_s^P = G_s \cup K_s^2 - \{s\}. \tag{34}$$

It is not difficult to see that G^P is a neighborhood system for $Z_m \cup D_m$.

The next step is to form an energy function U^P for G^P, one which will be a valid Gibbsian energy function for $P(x, l|x')$ over the neighborhood system G^P. For this let us make two assumptions, one necessary, the other not: namely, that the noise in (30) is additive white Gaussian noise of mean μ and standard deviation σ (which is the unnecessary assumption), and that this noise is independent of $X = \{x, l\}$ (which is the necessary one). The first assumption means simply that the grey-level noise at each pixel is a Gaussian random variable of mean μ and standard deviation σ, and that its mode of combination with the pixel grey level is addition. This assumption is unnecessary because the following proof can be modified to apply to any type of noise and any mode of noise combination that is invertible—that is, any noise process that can be expressed as some function of x' and $Q[H(x)]$, such as $x' - Q[H(x)]$, which is the case with additive noise.

Let $\tilde{\mu}$ denote the $m \times m$ matrix all of whose elements are equal to μ. We define the nonnegative function $U''(x)$ by

$$U''(x) = \frac{1}{2\sigma^2} \|\tilde{\mu} - x' + Q[H(x)]\|^2, \tag{35}$$

where the double bars indicate the standard matrix norm—the square root of the sum of the squares of the matrix elements. Then we have the following Bayesian result, which is established by Geman and Geman.[7]

Bayesian Theorem. For every fixed image, x', the posterior density $P(x, l|x')$ is a Gibbs distribution over $\{S, G^P\}$ with energy function

$$U^P(x, l) = U(x, l) + U'(x) \tag{36}$$

and temperature parameter unity.

Proof. Because the proof is instructive I give it in full. We start from Bayes' rule, namely,

$$P(x, l|x') = \frac{P(x'|x, l)P(x, l)}{P(x')}. \tag{37}$$

Now from (25) we have

$$P(x, l) = \frac{1}{Z}\exp[-U(x, l)]; \tag{38}$$

moreover, $P(x')$ is just a constant depending on x'. What needs elucidation, then, is $P(x'|x, l)$, i.e., the probability of getting the blurred and noisy image we did get, given that it was produced by (x, l). Since the noise is assumed independent of x and l, it is clear that $P(x'|x, l)$ is just the probability that the noise matrix is equal to $x' - Q[H(x)]$. And since we have assumed that the noise is Gaussian, white, etc, it follows that

$$P(x'|x, l) = k\exp[-U'(x)], \tag{39}$$

where k is a normalization constant. Consequently,

$$P(x, l|x') = \frac{1}{Z^P}\exp[-U^P(x, l)], \tag{40}$$

where Z^P is a normalizing constant depending on x'. It thus remains to show that U^P is a Gibbs energy function relative to the neighborhood system G^P.

This is most readily seen by analyzing the local characteristics of $P(x, l|x')$ as given by (40). If $s \in S$ is a feature site, the relevant local characteristic is, by definition,

$$P(l_s|x, \{l_r : r \neq s, r \in D_m\}, x') =$$

$$= \frac{\exp[-U^P(x,l)]}{\sum_{l_s} \exp[-U^P(x,l)]} = \frac{\exp[-U(x,l)]}{\sum_{l_s} \exp[-U(x,l)]}, \tag{41}$$

where the sums are over all values in the range of l_s, and the second step follows from (36). This is precisely the result we would get for the corresponding local characteristic of $P(x,l)$, and this shows that the neighborhood structure of U^P for feature sites is the same as for the prior U, as in the definition of G^P. Similarly, if s is a pixel site, then the relevant local characteristic of the posterior is

$$P(x_s|l, \{x_r : r \neq s, r \in Z_m\}, x') = \frac{\exp[-U^P(x,l)]}{\sum_{x_s} \exp[-U^P(x,l)]}, \tag{42}$$

but now the cancellations in the above ratio are less evident than in (41). To clarify which cancellations can be made, we decompose the energy in the form

$$U^P(x,l) = \sum_{C:s\in C} V_C(x,l) + \sum_{C:s\notin C} V_C(x,l) +$$

$$\sum_{r:s\in K_r} \{\mu - [x' - Q(H(x))]_r\}^2/2\sigma^2 + \sum_{r:s\notin K_r} \{\mu - [x' - Q(H(x))]_r\}^2/2\sigma^2,$$

where the first two terms give $U(x,l)$, and the second two give $U'(x)$. We can see that the sum of the second pair of terms gives $U'(x)$ as follows: by (35), the quantity being summed in either of the last two terms is the contribution to U' of the r^{th} pixel site; but since the total r-range of the two sums taken together is precisely all $r \in Z_m$, this contribution is clearly $U'(x)$. Note that the second and fourth terms of this decomposition of U^P do not depend on x_s, so that (42) becomes

$$P(x_s|l, \{x_r : r \neq s, r \in Z_m\}, x') =$$

$$= \frac{\exp[-\sum_{C:s\in C} V_C(x,l) - \sum_{r:s\in K_r} \{\mu - [x' - Q(H(x))]_r\}^2/2\sigma^2]}{\sum_{x_s} \exp[-\sum_{C:s\in C} V_C(x,l) - \sum_{r:s\in K_r} \{\mu - [x' - Q(H(x))]_r\}^2/2\sigma^2]}. \tag{43}$$

To see that the fourth term does not depend on x_s, we merely write the dependencies of $[x' - Q(H(x))]_r$ explicitly as follows:

$$[x' - Q(H(x))]_r = \Phi_r(x'_r, \{x_t : t \in K_r\}). \tag{44}$$

Thus the ratio in (42) depends on only the first two terms of the decomposition of U^P. Now the first of these terms depends solely on the

(x, l) for sites in G_s (because $s \in C \Rightarrow C \subset G_s$), and the second term depends solely on the sites in

$$\bigcup_{r:s\in K_r} K_r = \bigcup_{r\in K_s} K_r \equiv K_s^2.$$

This is enough to show that the neighborhood structure of U^P for pixel sites is as given in (34).

End of proof.

With this we have produced a complete description of the general construction of the posterior conditional $P(x, l|x')$. The items needed for this construction are (1) the prior probability density, $P(x, l)$, (2) the model of the image-formation process, and (3) the raw image. Thus the method under consideration is now totally explicit. To produce a concrete example, we need only a model of the prior probability density $P(x, l)$.

I conclude this overall section by summarizing the status of this approach to the image processing problem in terms of the issues which have emerged in its description, and also with some indication of alternatives. I begin with the last point mentioned, namely, the issue of providing explicit models for $P(x, l)$.

To compose a Gibbsian model of $P(x, l)$ that incorporates our prior knowledge, one must first decide on the neighborhood system G. The criteria for specifying G must come from an appraisal (for the specific image processing problem being addressed) of the correlation distance expected in the image, the two-dimensional analog of the width of the autocorrelation function of a random signal. After such an appraisal, a reasonable specification of G will be possible. One must then go through the details of what is essentially a combinatorial problem, namely, the enumeration of the full set of G's cliques: the set C'. For all but the smallest neighborhood structures (such as nearest-neighbor neighborhoods), this will be an arduous task that must nonetheless be done before we can assign local energy functions, V_C, to all the cliques $C \in C'$. This is the last point at which our *a priori* knowledge is injected. The criteria and guidelines to be used for this task are presently obscure, aside from making intuitively reasonable guesses. Much work must yet be done on this aspect of the problem.

Having completed the specification of $P(x, l)$, we must next construct $P(x, l|x')$, or equivalently, G^P and U^P. Though the task to be performed here is clear, achieving it would entail much ingenuity and systematic labor in all but the simplest cases. Such effort is nonetheless required to provide the data needed to execute the SRA algorithm and obtain the results sought.

As has been mentioned, the SRA scheme was developed to make it computationally feasible to solve a very complex problem, and SRA has made a great deal of progress in this direction. Nevertheless, the rather slow annealing schedules allowed by rigorous SRA most often produce impractically long computation times. It is thus apparent that we need to determine whether more rapid annealing schedules are viable, and to exploit the potential of further parallelization of the SRA algorithm.

Unfortunately, there appears to be little chance for significant improvement in annealing schedules, as the inverse-logarithmic lower bound on the rate of temperature decline appears tight. (This has not been proved, however.) In addition, attempts at further parallelization of the algorithm have produced some bizarre results, namely, the loss of *all image information* in the product of the parallelized version of the algorithm. This, however, seems to indicate some basic conceptual problems in algorithm theory, rather than any flaw in the SRA technique. Consequently, a speed-enhanced SRA algorithm will likely have to await the resolution of a paradox in pure algorithmics. It is therefore reassuring that alternatives to SRA exist.

The primary alternatives to SRA include the *renormalization group* approach of Gidas,[4] the less fundamental approach called *mean field annealing*,[5] and finally the methodology of *mathematical morphology*.[6] Like SRA, both the renormalization group and mean field annealing approaches are based on a Markov random field model as described in section 9.1. The Gidas method consists of a radical reorganization of the needed computations analogous to the renormalization group analysis developed in statistical mechanics. The less fundamental mean field annealing technique attempts to approximate and simplify the required annealing calculations (again in analogy with a theoretical technique used in physics), so that a sufficiently accurate and computationally feasible result can be obtained. Both approaches are quite promising.

The techniques of morphological image processing are quite different from all the methods discussed thus far. Mathematical morphology

sets out from an altogether distinct set of principles aimed at the direct analysis of the shapes and textures in an image. It is based on the consideration of a general class of *morphological transformations* of the closed subsets of the image manifold. The subsets acted upon by these transformations can be either the direct image sets in the case of binary images, or the cross sections of the intensity function defining a grey-tone image. The class of morphological transformations is defined in accordance with the principles of translational and scale "invariance," a "principle of local knowledge," and a certain continuity requirement. All these can be viewed as general requirements that emerge from the image-processing goal of shape and texture analysis. Morphological transformations differ from the more familiar image-processing transformations of windowing, cross-correlation, and Fourier transformation, in that they are intrinsically nonlinear and involve the probing of the image with user-specified *structuring elements.* The method can be generally viewed as a type of nonlinear filtering that leads to the direct extraction of shape information. This approach also holds great promise.

10 Conclusion

This report has attempted to convey a relatively detailed perspective on the beginning stage of the ATR problem, which simply concerns what needs to be done with the image data at the beginning of processing to determine the presence or absence of certain important objects. This primary task must at least result in limiting the wealth of data that make up a typical image to those which might be reflections of these important objects. It is simply not possible to intensively process, for example, a 4000 by 4000 pixel image for any relatively immediate practical purpose. I have accordingly considered what the basic elements of the beginning-stage task are: (1) the determination of a definitive set of image features for the particular recognition task, and (2) the establishment of the various mathematical methods needed to extract these features from the image data. This report has shown in detail that much progress has been achieved in these areas, but it has also made it clear that the problems of feature identification and the establishment of methodologies for systematic feature extraction are far from fully resolved. Powerful mathematical methods have been devised to deal with the primary difficulties of the ill-posed problems presented by feature extraction, and an array of important feature extraction problems have been adequately solved; nevertheless, the solutions of a number of these problems tend to bog down in the complexities of carrying out the procedures of the method, primarily those concerned with fashioning realizable and effective computer implementations. There is little choice here but to continue attempting to advance on these complexity issues, and a number of apparently fruitful avenues are being pursued.

In feature identification, which has not been considered in any methodical detail, a great deal of work needs to be done. Although not claiming any particular expertise in feature identification, on this subject I observed a surprising lack of literature, possibly because the problem does not fall under a single technical discipline. Questions arising in this area inevitably lead to comparisons with the human process of recognition, and these in turn lead to psychological or philosophical matters. As these are normally not subjects in which engineers, physicists, and mathematicians have expertise, it is not surprising that their applied literature does not deal with these questions systematically. The practitioners of the "hard" applied sciences tend (with some justification) to avoid philosophical concerns, thinking or hoping that the "recognition problem" does not really have a substantive philosophical or psycho-

logical element. But this may not be true. As a possible symptom of the shortsightedness of this view, consider the often reported frustration experienced with apparently algorithm-resistant imagery in which trained observers can nonetheless point to and rather simply describe what is important for recognition purposes. The elusiveness of what they describe may be due to an innocent difference between what they do in recognizing and what they *say* they do. A trained observer's verbal description of an object's important features has to be based, at least in part, on a (possibly mistaken) conceptual prejudice, that is, on an *explanation* that he unwittingly gives to himself of what that object is in terms of its recognizable modes of appearance. How else could he formulate his observations of what he does in recognizing an object? We may thus be forced to heed the *phenomenologist's* dictum that we should be quite sure that a thing has been adequately *described* before we try to *explain* it.

DISTRIBUTION

Administrator
Defense Technical Information Center
Attn DTIC-DDA (2 copies)
Cameron Station, Building 5
Alexandria, VA 22304-6145

Commander
Communications-Electronics Command
R&D Technical Library
Ft Monmouth, NJ 07703-5018

Commander
US Army Computer Systems Command
Attn Technical Library
Ft Belvoir, VA 22060

Chief
US Army Research Office (Durham)
Attn SLCRO-MA, Dir Mathematics Div
PO Box 12211
Research Triangle Park, NC 27709-2211

Director
Defense Communications Agency
Attn Command & Control Center
Washington, DC 20305

Director
Defense Communications Engineering Center
Attn Code R123, Technical Library
1860 Wiehle Ave
Reston, VA 22090

Director, Defense Nuclear Agency
Attn Tech Library
Washington, DC 20305

Dept of the Air Force, HQ
Radar Target Scatter Facility
6585th Test Group (AFSC)
Attn LTC R.L. Kercher, Chief
Holloman AFB, NM 88330

US Army Electronics Technology
and Devices Laboratory
Attn SLCET-DD
Ft Monmouth, NJ 07703-5601

Director
NASA
Attn Technical Library
John F. Kennedy Space Center
Kennedy Space Center, FL 32899

Director
NASA
Attn Technical Library
Langley Research Center
Hampton, VA 23665

Director
NASA
Attn Technical Library
Lewis Research Center
Cleveland, OH 44135

Director
NASA, Goddard Space Flight Center
Attn 250, Tech Info Div
Greenbelt, MD 20771

Under Secretary of Defense for Research
& Engineering
Attn Research & Advanced Tech
Washington, DC 20301

Brown University
Div. Applied Mathematics
Attn S. Geman
Providence, RI 02912

George Mason University
Attn ECE Dept., R.A. Athale
Fairfax, VA 22030

DISTRIBUTION (cont'd)

Engineering Societies Library
Attn Acquisitions Dept
345 E. 47th Street
New York, NY 10017

Institute for Telecommunications Sciences
National Telecommunications & Info Admin
Attn Library
Boulder, CO 80303

University of Maryland
Ctr. Environmental and Estuarine Studies,
Chesapeake Biological Lab
Attn E.V. Patrick
Solomons, MD 20688-0038

University of Maryland
Mathematics Dept. and Systems Research Ctr.
Attn C.A. Berenstein
College Park, MD 20742

University of Massachusetts
Dept. Mathematics and Statistics
Attn D. Geman
Amherst, MA 01003

The American University
Department of Mathematical Statistics
Attn S.D. Casey
4400 Massachusets Avenue, NW
Washington, DC 20016-8050

Long Island University
Philosophy Department
Attn J. Garcia-Gomez
Southampton, NY 11968

Installation Support Activity
Attn SLCIS-CC, Legal Office

US Army Laboratory Command
Attn AMSLC-DL, Director, Corporate Labs

USAISC
Attn AMSLC-IM-VA, Admin Ser Br
Attn AMSLC-IM-VP, Tech Pub Br (2 copies)

Harry Diamond Laboratories
Attn Laboratory Directors
Attn SLCHD-TL, Library (3 copies)
Attn SLCHD-TL, Library (WRF)
Attn SLCHD-NW, Chief
Attn SLCHD-NW-CS, Chief
Attn SLCHD-NW-E, Chief
Attn SLCHD-NW-EH, Chief
Attn SLCHD-NW-EP, Chief
Attn SLCHD-NW-ES, Chief
Attn SLCHD-NW-P, Chief
Attn SLCHD-NW-R, Chief
Attn SLCHD-NW-RP, Chief
Attn SLCHD-NW-RS, Chief
Attn SLCHD-NW-TN, Chief
Attn SLCHD-NW-TS, Chief
Attn SLCTO, N. Berg
Attn SLCTO, B. Weber
Attn SLCHD-ST, R. Johnson
Attn SLCHD-ST-OP, A. Filipov
Attn SLCHD-ST-OP, C. Garvin
Attn SLCHD-ST-OP, J. Goff
Attn SLCHD-ST-OP, N. Gupta
Attn SLCHD-ST-OP, L. Harrison
Attn SLCHD-ST-OP, D. Kafig
Attn SLCHD-ST-OP, J. Mait
Attn SLCHD-ST-OP, D. Prather
Attn SLCHD-ST-OP, B. Sadler
Attn SLCHD-ST-OP, D. Smith
Attn SLCHD-ST-OP, M. Taylor
Attn SLCHD-ST-SA, J. Dammann
Attn SLCHD-ST-SA, B. Stann
Attn SLCHD-ST-SS, D. Gerstman
Attn SLCHD-ST-SS, J. Griffin
Attn SLCHD-ST-SS, D. McCarthy