Ill-Posed Problems in Early Vision

MARIO BERTERO, TOMASO A. POGGIO, ASSOCIATE, IEEE, AND VINCENT TORRE

Invited Paper

The first processing stage in computational vision, also called early vision, consists of decoding two-dimensional images in terms of properties of 3-D surfaces. Early vision includes problems such as the recovery of motion and optical flow, shape from shading, surface interpolation, and edge detection. These are inverse problems, which are often ill-posed or ill-conditioned. We review here the relevant mathematical results on ill-posed and ill-conditioned problems and introduce the formal aspects of regularization theory in the linear and nonlinear case. Specific topics in early vision and their regularization are then analyzed rigorously, characterizing existence, uniqueness, and stability of solutions.

INTRODUCTION

Vision systems, whether artificial or biological, are confronted with the problem of inferring geometrical and physical properties of surfaces around the viewer. The available data—the images—consist of two-dimensional arrays of light intensity values measured by an eye or a camera. For tasks such as navigation, manipulation, and visual recognition, vision systems have to first recover 3-D properties of surfaces from the 2-D images. Typical 3-D properties are the distance between the surfaces and the viewer, their orientation, texture, reflectance, and motion parameters (from a temporal sequence of images).

The visual skills that provide us with this kind of information have been explored in animals and humans with physiological and behavioral techniques. With the recent development of computer vision, these problems have been formulated rigorously and given, by now, familiar names such as, *structure from stereo, structure from motion, structure from texture, shape from shading, edge detection, visual interpolation, and computation of optical flow.* The computational modules that solve them together constitute the core of *early vision*, and provide spatial and geometrical information about the 3-D world. The results of this first stage of processing are then used for *higher* level tasks such as navigation in the environment, manipulation of objects and, of course, object recognition as well as reasoning about objects. Unlike high level vision, early vision is mostly considered as a set of *bottom-up* processes that do not rely upon specific high-level information about the scene to be analyzed. It is commonly argued, on the basis of computational and psychophysical considerations, that these different modules of early vision can be analyzed independently of each other, to a first approximation. Their most natural implementation is in terms of distinct pieces of hardware, whose outputs will be integrated at a later stage, possibly using more "intelligent" procedures.

Even a superficial analysis of these problems reveals their common inverse nature: they can be regarded as *inverse optics* since they attempt to recover physical properties of 3-D surfaces from the 2-D images they generate. This observation characterizes the field of *early vision* as the solution of problems of inverse optics [1], [2]. The same observation makes clear that the data of the problems (the 2-D images) contain in general limited information about the solutions (the 3-D properties). This lack of information implies that the problems of early vision are very often ill-posed in the original sense of Hadamard [3], [4]: the solution may not be unique (giving an ambiguous reconstruction) or it does not exist, or it does not depend continuously on the data. As a consequence of the ill-posedness of the problems of early vision, the effect of noise, which is always present in a physical measurement, is very important: even a small error in the data can produce an extremely large error in the solution. Notice also that, since practical problems are always made discrete and therefore are reduced to the inversion of a matrix (in the linear case), non-uniqueness and numerical instability can have very similar effects.

Inverse and ill-posed problems are very important in several domains of applied science such as medical diagnostics, seismic exploration, atmospheric remote sensing, radioastronomy, microscopy and so on. The relevance of these problems has stimulated, since the beginning of the 1960s, the development of theoretical and practical methods for determining approximate and stable solutions. Most of these methods have now been unified in a theory which

Manuscript received June 10, 1987; revised March 21, 1988. The research in this paper was performed in part at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contract N00014-85-K-0124. Some support for T. Poggio is provided by a gift from the Artificial Intelligence Division of the Hughes Aircraft Corporation. NATO provided some support for V. Torre. The research was also funded by EEC (ESPRIT P940).

M. Bertero is with the Dipartimento di Fisica and the Istituto Nazionale di Fisica Nucleare, 16146 Genova, Italy.

T. Poggio is with the Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA.

V. Torre is with the Dipartimento di Fisica, 16146 Genova, Italy.

IEEE Log Number 8822794.

0018-9219/88/0800-0869$01.00 © 1988 IEEE

is called the *regularization theory of ill-posed problems* [5], [6]. On the other hand, it has been recognized only recently that several problems of early vision are ill-posed and that methods developed independently by researchers active in this field are in fact specific examples of regularization theory [1], [2], [7]–[17]. Even if the theory of ill-posed problems has not yet significantly contributed to early vision, the development of this theoretical framework is important for at least two reasons. First, the synthesis of methods developed independently in different scientific domains within a general framework always contributes to a deeper understanding of the problems. Second, regularization theory provides several methods and algorithms that have not yet been applied to early vision.

The aim of this paper is to give a rigorous formulation and development of the ideas outlined above. It is organized in two parts. In the first part, for the convenience of the reader, we sketch the theory of ill-posed problems. In particular we characterize the difference between well-posed and ill-posed problems (Section II) and between well-conditioned and ill-conditioned problems (Section IV). The notions of generalized inverses (Section III) and of regularization methods (Section V) are then introduced. Section VI contains some results related to inverse nonlinear problems. Several monographs are already available on the subject [5], [6], [18]–[21]. They stress the mathematical aspects of the theory rather than its practical applications. More physical presentations are given in [22], [23].

In the second part we show that several approaches, recently proposed by many authors to solve problems of early vision, using smoothness constraints and variational techniques, can be obtained directly and justified in the general framework of regularization theory. Five problems in early vision are studied in detail: edge detection and numerical differentiation (Section VII), optical flow (Section VIII), surface interpolation (Section IX), shape from shading (Section X), and stereo matching (Section XI). The solutions that we will describe for edge detection and stereo using the regularization approach are new, though they are practically equivalent to previous methods. In the case of optical flow, surface interpolation and shape from shading, regularization leads to the same solution already obtained by previous workers. We derive, however, more complete results about uniqueness and the properties of the solution.

The problems of early vision are in general mildly ill-posed. Roughly speaking, this means that a reduction of the errors in the data can produce a significant improvement of the solutions. This is a lucky situation because many inverse problems are severely ill-posed, in the sense that a reduction of the noise even of several orders of magnitude will not induce a significant improvement of the solution. More precise mathematical definitions of these concepts are given in [24], where mildly ill-posed problems are called well-behaved.

The prototypical problem for early vision is surface reconstruction. This is the problem of approximating a "surface" from noisy and possibly sparse data. As we will see later, even problems that do not suffer from being underdetermined (like the new formulation of the optical flow in Section VIII-E) still require regularization because the measurements are noisy and sparse. The main role of regularization in vision is therefore as an approximation technique that exploits *a priori* information to counter noise in the data and to fill-in wherever data are missing or not reliable.

Part One

I. Outline

In this part of the paper we review some of the methods which have been developed for the approximate solution of ill-posed problems. The linear case is discussed in detail since a well-developed theory is available. We also make some comments on nonlinear problems.

In Section II we define the class of well-posed problems, stressing that a well-posed problem is not necessarily robust against noise. A well-posed problem, in order to have solutions that are robust against noise, must also be well-conditioned (see Section IV). For ill-posed, linear, inverse problems, well-posedness can be restored by generalized solutions if the range of the operator (which has to be inverted) is closed (see Section III). When the range of the operator is not closed, or when the problem is seriously ill-conditioned, regularization techniques have to be used (Section V) in order to avoid the instability of the solution against noise. Therefore, since images are intrinsically noisy, these techniques represent the ideal tool for early vision problems. Some results on inverse nonlinear problems are presented in Section VI.

II. Well-Posed and Ill-Posed Problems

Hadamard [3], [4] defined a mathematical problem to be *well-posed* when:

a) for each datum g in a given class of functions Y there exists a solution u in a prescribed class X (*existence*);
b) the solution u is unique in X (*uniqueness*);
c) the dependence of u upon g is continuous, i.e., when the error on the data g tends to zero, the induced error on the solution u tends also to zero (*continuity*).

The requirement of *continuity* is related to the requirement of *stability* or *robustness* of the solution (see, for instance, [25]). Continuity, however, is a necessary but not sufficient condition for stability. A well-posed problem can be ill-conditioned (see Section IV).

All the classical problems of mathematical physics, such as the Dirichlet problem for elliptic equations, the forward problem for the heat equation, and the Cauchy problem for hyperbolic equations, are well-posed in the sense of Hadamard. Also, the "direct" problem in scattering (or imaging) theory, namely the computation of the scattered radiation (image) from a known constitution of the sources and of the targets, is well-posed.

"Inverse" problems usually are *not* well-posed. In most cases an "inverse" problem can be obtained from the "direct" one by exchanging the role of solution and data. For instance, in the case of scattering theory, the inverse problem consists of the computation of the characteristics of the targets from the knowledge of the sources and of the scattered radiation.

As an example of an inverse problem in early vision, let us consider the problem of edge detection. One part of the problem is equivalent to numerical differentiation which is ill-posed because the solution does not depend contin-

uously on the data. The intuitive reason for the ill-posed nature can be seen by considering a function $f(x)$ perturbed by a very small noise term $\epsilon \sin \Omega x$. The functions $f(x)$ and $f(x) + \epsilon \sin \Omega x$ can be arbitrarily close for very small ϵ, but their derivatives may be very different if Ω is large enough. This simply means that differentiation "amplifies" high frequency noise.

The need to investigate problems that are not well-posed, but are of interest in applied science, originated two interesting branches of mathematical analysis: the first is the theory of *generalized inverses* [26], [27] which is an extension of the theory of the Moore-Penrose inverse of a matrix; the second is the regularization theory of *ill-posed* (or improperly posed) problems [5], [6], [18]-[21]. At present, the term ill-posed is used generally (but not only) for those problems that do not satisfy the requirement of continuity. Examples of ill-posed problems are analytic continuation, the Cauchy problem for elliptic equations, backsolving the heat equations, superresolution, computer tomography, Fredholm integral equations of the first kind, and, as we will see, many problems in early vision.

III. Generalized Inverses

Most linear inverse problems can be formulated as follows: assume that functional spaces X, Y (for instance, Hilbert spaces) are given and that a linear, continuous operator L from X into Y is also given; then the problem is to find, for some prescribed $g \in Y$, a function $u \in X$ such that

$$g = Lu. \tag{3.1}$$

In this formulation, the direct problem is just the computation of g, given u. Therefore, continuity of L is equivalent to well-posedness of the direct problem.

The problem of numerical differentiation discussed in the previous section takes the form (3.1) if we introduce the integral operator

$$(Lu)(x) = \int_{-\infty}^{x} u(y)\, dy. \tag{3.2}$$

Thus u is the derivative of the data g. The operator (3.2) is not continuous in $L^2(-\infty, +\infty)$ but continuity can be restored by an appropriate choice of the space X.

The problem (3.1) is well-posed if and only if the operator L is injective (i.e., the equation $Lu = 0$ has only the trivial solution $u = 0$ (uniqueness)), and it is *onto* Y (existence). Then general theorems of functional analysis (for instance, the "closed graph theorem") ensure that the inverse mapping L^{-1} is also continuous (continuity).

Assume now that the equation $Lu = 0$ has nontrivial solutions. The set of these solutions is a closed subspace of X, which is called the null space $N(L)$ of L. This is the subspace of the "invisible objects," since they produce a zero image g. Assume also that the range $R(L)$ of L, namely the set of the g which are images of some $u \in X$, is a *closed* subspace of Y. An example is provided by the integral operator corresponding to the perfect low pass filter

$$(Lu)(x) = \int_{-\infty}^{+\infty} \frac{\sin \Omega(x - y)}{\pi(x - y)}\, u(y)\, dy. \tag{3.3}$$

In such a case, if we take $X = Y = L^2(-\infty, +\infty)$, the null space is the set of all the functions u whose Fourier transform is zero on the band $[-\Omega, \Omega]$, while the range of L is the set of the band-limited functions with bandwidth Ω, which is a closed subspace of $L^2(-\infty, +\infty)$. Notice that L is a projection operator, the so-called band-limiting operator.

A way of restoring existence and uniqueness of the solution under the conditions above is to redefine both the solution space X and the data space Y. We take a new space X' which is the set of all the functions orthogonal to $N(L)$ (in the case of (3.2), X' is the space of square integrable Ω-bandlimited functions), and we take $R(L)$ as the new data space Y' (in the case (3.2) again, the space of the square integrable Ω-bandlimited functions). Then for any $g \in Y'$ there exists a unique $u \in X'$ such that $g = Lu$, (in the case of (3.2) the solution is trivial: $u = g$) and therefore the new problem is well-posed.

The redefinition of the space X, Y outlined above usually is quite difficult (almost impossible) in practical problems. Therefore, it is useful to have a method, based on the solution of variational problems, which produces the same result. This is just the method of *generalized inverses* [26], [27].

A. Least Squares Solutions or Pseudosolutions

Consider first the case in which L is *injective* but not *onto* (i.e., the existence condition is not satisfied). The functions $u \in X$ that solve the variational problem

$$\|Lu - g\|_Y = \text{minimum} \tag{3.4}$$

where $\|\cdot\|_Y$ denotes the norm of Y, are called the *least squares solutions* (or pseudosolutions) of problem (3.1). These solutions can be easily obtained considering the first variation of the functional (3.4)

$$2Re\,(Lu - g, Lh)_Y \tag{3.5}$$

where Re denotes the real part, h is an arbitrary function of X and $(\cdot, \cdot)_Y$ the inner product of the Hilbert space Y. Setting (3.5) equal to zero, we obtain the Euler equation

$$L^*Lu = L^*g \tag{3.6}$$

where L^* is the adjoint of the operator L (L^* is a mapping from Y into X). When $R(L)$ is closed, (3.6) always has solutions but the solution is not unique when $N(L)$ is nontrivial. Notice that the set of solutions of (3.6) coincides with the set of solutions of the equation

$$Lu = Pg \tag{3.7}$$

where P is the projection onto $R(L)$. Therefore, solving (3.5) is equivalent to assuming $Y' = R(L)$ or to projecting g onto Y'. When the operator L is injective, the solution of (3.6) is unique and well-posedness has been restored.

B. Normal Pseudosolutions or Generalized Solutions

Consider now the case in which L is not injective (i.e., the uniqueness condition is not satisfied and the problem is underconstrained). Then, one looks for the solution of (3.6) which has minimal norm

$$\|u\|_X = \text{minimum}. \tag{3.8}$$

This solution is unique and is denoted by u^+. u^+ is usually called the *generalized solution* (or normal pseudosolution) of problem (3.1). u^+ is orthogonal to $N(L)$ and therefore this procedure is equivalent to taking $X' = N(L)^{\perp}$.

Since there exists a unique u^+ for any $g \in Y$, a linear map-

ping L^+ from Y into X is defined by

$$u^+ = L^+ g. \tag{3.9}$$

The operator L^+ is the *generalized inverse* of L and it is continuous. Therefore, the problem of computing the generalized solution of (3.1) is well-posed if and only if $R(L)$ is closed. The essential reason for this result is that in this case the space Y can be decomposed as

$$Y = R(L) \oplus R^{\perp}(L) \tag{3.10}$$

where $\oplus$ means direct sum and $R^{\perp}(L)$ is the orthogonal complement of $R(L)$. This decomposition can be made if, and only if, $R(L)$ is closed.

C. C-Generalized Solutions

In several inverse problems, the generalized solution is trivial or does not satisfy some physical requirements such as smoothness. Examples are provided in Section IX. Then an extension of the generalized solution proceeds as follows: let $p(u)$ be a norm or a seminorm on X of the following style:

$$p(u) = \| Cu \|_Z \tag{3.11}$$

where C is a linear operator from X into the Hilbert space Z (the constraint space). The operator C may not be defined everywhere on X. For instance, suppose X is a space of square-integrable functions and C is a differential operator. Therefore, in general, $p(u)$ is defined on a subset of X, i.e., the domain of C, denoted as $D(C)$. When the null space of C is trivial (containing only the null element of X), then $p(u)$ is a norm on $D(C)$; otherwise, $p(u)$ is a seminorm.

If there exists a unique least-squares solution that minimizes $p(u)$, we denote it by u_C^+ and we call it a *C-generalized solution*. The mapping $g \mapsto u_C^+$ defines a linear operator L_C^+ from Y into X, which will be called the C-generalized inverse of L. It is obvious that u_C^+ can have a nonzero component onto $N(L)$ (the subspace of the "objects" that are "invisible" under the action of the operator L). Therefore, this procedure is physically plausible only when the constraint describes some physical property of the solution of the problem.

Necessary and sufficient conditions for the existence of u_C^+ for any g have been given in the case where $R(L)$ is closed and C is a bounded operator with $R(C)$ also closed [26]. However, the assumption of a bounded constraint operator C may not cover the interesting case of a differential operator. Furthermore, when $D(C)$ is a subset of X, it is obvious that u_C^+ does not exist for any $g \in Y$. If we denote by $LD(C)$ the set of all the functions $g \in Y$ such that $g = Lu$ with $u \in D(C)$, then $LD(C)$, in general, does not coincide with $R(L)$. Under these circumstances, if $Pg \notin LD(C)$, the intersection between the set of the least squares solutions and $D(C)$ is empty and u_C^+ does not exist. In other words, the problem of determining the C-generalized solution may be ill-posed even when $R(L)$ is closed.

Sufficient conditions which assure the existence of u_C^+ for any g such that $Pg \in LD(C)$ are the following [21]:

i) The intersection of $N(L)$ and $N(C)$ contains only the null element of X, i.e., the set of equations

$$Lu = 0, \qquad Cu = 0 \tag{3.12}$$

has only the common trivial solution $u = 0$ (uniqueness condition);

ii) The operator $C: X \mapsto Z$ is closed with $D(C)$ dense in X and $R(C) = Z$;

iii) The set of functions g such that $g = Lu$ and $Cu = 0$, i.e., the set $LN(C)$, is closed in Y.

The third condition is always satisfied in the case of seminorms defined in terms of differential operators because in that case $N(C)$ is a finite dimensional subspace of X and L is a continuous operator.

When the constraint operator C satisfies conditions i)–iii) and furthermore is bounded, u_C^+ exists for any $g \in Y$ and the C-generalized inverse L_C^+ is bounded.

D. Generalized Solutions for Problems with Discrete Data

We conclude this section by noting that problems with discrete data can be formulated as (3.1), g being now an n-dimensional vector in a Euclidean space. In fact, ignoring the errors in the data, a linear inverse problem with discrete data can be formulated as follows [28]: given a set $\{F_i\}_{i=1}^n$ of linear functionals defined on X and a set $\{g_i\}_{i=1}^n$ of numbers, find a function $u \in X$ such that

$$g_i = F_i(u), \qquad i = 1, \cdots, n. \tag{3.13}$$

In particular, when the functionals F_i are continuous on X, by Riesz Theorem [80], there exist functions $\psi_1, \psi_2, \cdots, \psi_n$ such that

$$F_i(u) = (u, \psi_i)_X \tag{3.14}$$

where $(\cdot, \cdot)_X$ is the inner product of X.

This problem is a special case of the problem (3.1) if we consider the data g_i as the components of a vector $\vec{g}$ in a n-dimensional Euclidean space Y and if we define an operator L from X into Y by means of the relation

$$(Lu)_i = (u, \psi_i)_X, \qquad i = 1, \cdots, n. \tag{3.15}$$

The operator L is not injective: $N(L)$ is the infinite dimensional closed subspace of all the functions u orthogonal to the subspace spanned by the functions ψ_i. On the other other hand, the range of L, $R(L)$, is closed: $R(L)$ is just Y when the functions ψ_i are linearly independent; otherwise, it is a subspace with dimension $n' < n$.

Along the lines described above one can introduce generalized solutions or C-generalized solutions for problems with discrete data. Their determination is always a well-posed problem in the strict mathematical sense. However, numerical stability cannot be guaranteed (see the next section).

As a final remark, we point out that the problem of interpolation by means of spline functions can be formulated as a problem of determining a generalized or C-generalized solution in a suitable reproducing kernel Hilbert space (see, for instance, [28], [29]). As a simple example we shall discuss the problem of linear interpolation.

Let X be a space of differentiable functions, defined on the interval [0, 1] and having a square integrable first derivative. X is a Hilbert space if we define a scalar product by means of the formula

$$(u, v)_X = u(0)\, v(0) + \int_0^1 u'(x)\, v'(x)\, dx. \tag{3.16}$$

Let $x \in [0, 1]$ be a fixed, arbitrary point; then, from the elementary relation

$$u(x) = u(0) + \int_0^x u'(x')\, dx' \tag{3.17}$$

it follows that

$$u(x) = (u, Q_x)_X \tag{3.18}$$

where

$$Q_x(x') = 1 + \min\{x, x'\}. \tag{3.19}$$

Clearly $Q_x \in X$ for any x, and therfore all the evaluation functionals (i.e., the functionals which associate to a function u its value in a given point) are continuous.

A Hilbert space of continuous functions having the previous property is called a reproducing kernel Hilbert space. The reproducing kernel $Q(x, x')$ is defined by

$$Q(x, x') = Q_x(x') = Q_{x'}(x), \tag{3.20}$$

and its name is due to the relation

$$(Q_x, Q_{x'})_X = Q(x, x'). \tag{3.21}$$

Assume now that a function $u \in X$ is specified at the points $x_1, x_2, \cdots, x_N (x_n \in [0, 1])$ and let $g_1, g_2, \cdots, g_N$ be its values. The interpolation problem (i.e., find $u \in X$ such that $u(x_n) = g_n$ for $n = 1, \cdots, N$) can be formulated, thanks to (3.18), as the problem of determining $u \in X$ such that

$$(u, Q_{x_n}) = g_n, \qquad n = 1, \cdots, N \tag{3.22}$$

and therefore it takes the form (3.13), (3.14). If we recall that the generalized solution is orthogonal to $N(L)$ (Section III) and that $N(L)$ is the orthogonal complement of the subspace spanned by the functions

$$\psi_n(x) = Q(x_n, x) \tag{3.23}$$

(L is defined as in (3.15)), we conclude that the generalized solution must be a linear combination of the functions ψ_n

$$u^+(x) = \sum_{n=1}^{N} c_n Q(x_n, x). \tag{3.24}$$

From (3.19) it follows that $u^+(x)$ is just the linear interpolation of the data g_n.

Interpolation by means of splines of degree $m = 2k - 1$ ($k \geq 1$) can be obtained along similar lines by a suitable definition of the reproducing kernel Hilbert space X [28]. Interpolation by means of natural splines of the same degree [29] can be formulated as the problem of determining, in the same space, a C-generalized solution that minimizes the L^2 norm of the derivative of order k.

IV. Well-Conditioned and Ill-Conditioned Problems

As already remarked in previous sections, continuous dependence of the solution on the data does not yet mean that the solution is robust against noise. Generalized solutions of inverse problems with discrete data can provide striking evidence of this fact. Therefore, it is necessary to investigate more carefully error propagation from the data to the solution when solving problem (3.1).

We assume, as in Section III, that $R(L)$ is closed, so that the generalized inverse L^+ is continuous. We denote by Δg a variation of the data g and by Δu^+ the corresponding variation on the generalized solution u^+. Then the standard analysis of error propagation proceeds as follows:

From (3.8), because of the linearity of L^+, we get $\Delta u^+ = L^+ \Delta g$, which implies

$$\|\Delta u^+\|_X \leq \|L^+\| \|\Delta g\|_Y \tag{4.1}$$

where $\|L^+\|$ denotes the norm of the continuous (bounded) operator L^+. Analogously, from (3.1), with $u = u^+$, it follows that

$$\|g\|_Y \leq \|L\| \cdot \|u^+\|_X. \tag{4.2}$$

Combining (4.1) and (4.2) we obtain the inequality

$$\frac{\|\Delta u^+\|_X}{\|u^+\|_X} \leq \|L\| \|L^+\| \frac{\|\Delta g\|_Y}{\|g\|_Y}. \tag{4.3}$$

It is important to point out that this inequality is precise in a certain sense. When L is an $N \times M$ matrix or L corresponds to an inverse problem with discrete data, then equality can hold. If L is an operator on infinite dimensional spaces, then one can always prove that the left-hand side (LHS) of (4.3) can be arbitrarily close to the right-hand side (RHS).

The quantity

$$\alpha = \|L\| \|L^+\| \geq 1 \tag{4.4}$$

is called the *condition number* of the problem. When α is not far from 1, the problem is said to be *well-conditioned*, while when α is large the problem is said to be *ill-conditioned*.

It is obvious that these definitions are not as precise as that of well-posedness. However, what is important in practice is the estimation of the condition number since it provides insight into the numerical stability of the problem. In the case where L is an $N \times M$ matrix, $\|L\|$ is the square root of the maximum eigenvalue of the $M \times M$ positive semidefinite and symmetric matrix L^*L (notice that the positive eigenvalues of this matrix coincide with the positive eigenvalues of the matrix LL^*) and $\|L^+\|$ is the inverse of the square root of the minimum positive eigenvalue of the same matrix, i.e.,

$$\alpha = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}. \tag{4.5}$$

Inverse problems with discrete data are always well-posed in the sense that the generalized solution depends continuously on the data. They can be, however, ill-conditioned and also extremely ill-conditioned. When the discrete problem is a discrete version of an ill-posed problem formulated in infinite dimensional spaces, then the ill-conditioning of the generalized solution depends on the number of data points and, in general, it increases by increasing the number of data points.

V. Regularization Methods

When the range of L, $R(L)$, is not closed, then the inverse L^{-1} or the generalized inverse L^+ is not defined everywhere on Y and it is not continuous. Therefore, both the requirements of existence and continuity do not hold true. This is the most difficult case and appropriate techniques are required. An example of operators in this class is provided by compact operators (not of finite rank; see [26] for the definition of a compact operator). It is easy to see that an

ill-posed problem has a condition number $\alpha = \infty$. Therefore, extremely ill-conditioned problems behave in practice as ill-posed problems and have to be treated by the same techniques.

A. Tikhonov Regularization

The most investigated approach to ill-posed problems is the *regularization method* of Tikhonov [30]. The key idea is to introduce a family of continuous "approximations" of a noncontinuous operator. More precisely, a regularization algorithm for the generalized solution of (3.1) is given in terms of a one-parameter family of continuous operators R_λ, $\lambda > 0$, from Y into X, such that for any given $g \in R(L)$

$$\lim_{\lambda \to 0} R_\lambda g = L^+ g. \tag{5.1}$$

Therefore, when applied to noise-free data g, R_λ provides an approximation of u^+ which becomes better and better as $\lambda \to 0$. However, when R_λ is applied to noisy data $g_\epsilon = g + n_\epsilon$, where n_ϵ represents experimental errors or noise, we have

$$R_\lambda g_\epsilon = R_\lambda g + R_\lambda n_\epsilon, \tag{5.2}$$

and the second term typically is divergent when $\lambda \to 0$. It follows that a compromise between "approximation" (the first term) and "error propagation" (the second term) is required. This is the problem of the "optimal choice" of the *regularization parameter* λ.

One of the most studied regularization techniques consists of minimizing the functional

$$\|Lu - g\|_Y^2 + \lambda \|Cu\|_Z^2 = \text{minimum}, \tag{5.3}$$

where C is a constraint operator, satisfying for instance the conditions stated in Section III. In the original paper of Tikhonov, it is given by

$$\|Cu\|_Z^2 = \sum_{r=0}^{\gamma} \int c_r(x) \, |u^{(r)}(x)|^2 \, dx \tag{5.4}$$

where the weights $c_r(x)$ are strictly positive functions and $u^{(r)}(x)$ indicates the rth-order derivative of $u(x)$. If u_λ is the solution of (5.3), and if we put

$$u_\lambda = R_\lambda g \tag{5.5}$$

then

$$R_\lambda = (L^*L + \lambda C^*C)^{-1} L^*. \tag{5.6}$$

Notice that u_λ is unique when (3.11) have only the trivial solution $u = 0$ and that when $\lambda \to 0$, $g \in R(L)$, $R_\lambda g$ converges to $L_C^+ g$ [21].

Three methods have been proposed for the choice of λ in (5.6) and in the case of noisy data g_ϵ:

i) Among all u such that $\|Cu\|_Z \le E$ find u that minimizes $\|Lu - g_\epsilon\|_Y$ [31]. Using the method of Lagrange multipliers the solution of this problem can be reduced to the solution of (5.3), with λ arbitrary, and to the search of the unique λ such that

$$\|Cu_\lambda\|_Z = E. \tag{5.7}$$

ii) Among all u such that $\|Lu - g_\epsilon\|_Y \le \epsilon$, with given ϵ, find u that minimizes $\|Cu\|_Z$ [32], [33]. Again, the solution of the problem is equivalent to finding the unique λ such that

$$\|Lu_\lambda - g_\epsilon\|_Y = \epsilon. \tag{5.8}$$

This is also called *Morozov's discrepancy principle*.

iii) Among all u such that $\|Lu - g_\epsilon\|_Y \le \epsilon$, $\|Cu\|_Z \le E$, find a u of the type (5.5). This is equivalent [34], [35] to taking

$$\lambda = (\epsilon/E)^2. \tag{5.9}$$

The first method consists of finding the function u that satisfies the constraint $\|Cu\|_Z \le E$ and best approximates the data. The second method computes the function u that is sufficiently close to the data (ϵ depends on the estimate of the errors) and is most "regular." In the third method, one looks for a compromise between the degree of regularization and the closeness of the solution to the data.

B. Regularization and Filtering

The regularized solution (5.5), (5.6) takes a very simple form in the case where L is a convolution operator

$$(Lu)(x) = \int_{-\infty}^{+\infty} K(x - y) \, f(y) \, dy \tag{5.10}$$

(notice that the operator (3.2) is an operator in this class) and the constraint operator C is the identity operator, $C = I$. Then, in terms of Fourier transforms we obtain

$$u_\lambda(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{|\hat{K}(\xi)|^2}{|\hat{K}(\xi)|^2 + \lambda} \frac{\hat{g}(\xi)}{\hat{K}(\xi)} e^{ix\xi} \, d\xi \tag{5.11}$$

where $\hat{K}(\xi)$ and $\hat{g}(\xi)$ are the Fourier transform of the kernel $K(x)$ and of the data function $g(x)$ respectively. It follows that the regularized solution is essentially a "filtered" version of the non-regularized (generalized) solution of (3.1), which is given by

$$u^+(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\hat{g}(\xi)}{\hat{K}(\xi)} e^{ix\xi} \, d\xi. \tag{5.12}$$

This remark suggests that, in this case, one can define regularization algorithms in terms of filter functions $\Phi(\lambda; \xi)$:

$$u_\lambda(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(\lambda; \xi) \frac{\hat{g}(\xi)}{\hat{K}(\xi)} e^{ix\xi} \, d\xi \tag{5.13}$$

satisfying the conditions: a) $0 \le \Phi(\lambda; \xi) \le 1$; b) $\Phi(\lambda; \xi) \to 1$ for any ξ when $\lambda \to 0$; c) $\Phi(\lambda; \xi)/\hat{K}(\xi)$ is a bounded function of ξ for any $\lambda > 0$.

Such a procedure is often used in the problem of edge detection (Section VII). The proof that (5.13) defines a regularization algorithm, in the sense specified in Section V-A, can be found, for instance, in [21].

C. Smoothing and Interpolation

As already remarked, regularization algorithms can be used for ill-conditioned problems. A well-known example is the smoothing of a function whose values, specified on a finite set of points, are affected by errors [36]. It is interesting to compare smoothing and interpolation by means of cubic splines using the framework outlined above. Interpolation of a function $u(x)$, $x \in [0, 1]$, is the problem of searching for a function which takes the prescribed values

$$u(x_i) = g_i, \qquad i = 1, \cdots, n \tag{5.14}$$

and minimizes the seminorm [29]

$$p(u) = \int_0^1 |u''(x)|^2 \, dx. \tag{5.15}$$

Therefore, the interpolation problem is equivalent to the computation of a generalized solution. On the other hand, the smoothing problem is formulated again as the minimization of the seminorm (5.15) [36], but condition (5.14) is replaced by

$$\sum_{i=1}^{n} |u(x_i) - g_i|^2 \leq \epsilon^2 \tag{5.16}$$

(for simplicity, we have assumed that the errors on the data have the same variance). Therefore, the smoothing problem corresponds to method ii) for the choice of the regularization parameter.

D. Cross Validation and Generalized Cross Validation

We conclude this section with a short description of the *cross validation method* [37]–[39]. This is a method for the choice of the regularization parameter and it has been applied to smoothing problems and also to the solution of Fredholm integral equations of the first kind in the framework of the method of collocation (or moment-discretization). However, it applies to any linear inverse problem with discrete data, as formulated in Section III.

The idea behind cross validation is to allow the data points themselves to choose the value of the regularization parameter by requiring that a good value of the parameter should predict missing data points. In this way, no *a priori* knowledge about the solution and/or the noise is required.

Let $(Lu)_i$ be defined as in (3.15) and let $u_\lambda^{[k]}$ be the minimizer of the functional

$$F_\lambda^{[k]}[u] = \frac{1}{n} \sum_{\substack{i=1 \\ i \neq k}}^{n} |(Lu)_i - g_i|^2 + \lambda \| u \|_X^2. \tag{5.17}$$

Then the cross validation function $V_o(\lambda)$ is defined by

$$V_o(\lambda) = \frac{1}{n} \sum_{k=1}^{n} |(Lu_\lambda^{[k]})_k - g_k|^2 \tag{5.18}$$

and the cross validation method consists in determining the value of λ, say $\bar{\lambda}$, which minimizes (5.18). The computation of the minimum is based on the relation

$$V_o(\lambda) = \frac{1}{n} \sum_{k=1}^{n} \frac{|(Lu_\lambda)_k - g_k|^2}{|1 - A_{kk}(\lambda)|^2}, \tag{5.19}$$

where u_λ is the minimizer of the functional

$$F_\lambda[u] = \frac{1}{n} \sum_{i=1}^{n} |(Lu)_i - g_i|^2 + \lambda \| u \|_X^2, \tag{5.20}$$

and $A_{kk}(\lambda)$ is the kkth entry of the $n \times n$ matrix

$$A(\lambda) = LL^*(LL^* + \lambda I)^{-1} \tag{5.21}$$

where LL^* is the Gram matrix of the functions ψ (see (3.15)).

It has been shown [39] that, from the point of view of minimizing predictive mean-square error, the minimization of $V_0(\lambda)$ must be replaced by the minimization of the generalized cross-validation function defined by

$$V(\lambda) = \left(\frac{1}{n} Tr\,[I - A(\lambda)] \right)^{-2} \left(\frac{1}{n} \| (I - A(\lambda))\, \vec{g} \|^2 \right) \tag{5.22}$$

where $\| \cdot \|$ denotes the Euclidean norm and Tr is the trace operation. An important property of $V(\lambda)$ is the invariance with respect to permutations of the data.

VI. Regularization of Nonlinear Problems

The case of nonlinear ill-posed problems is quite difficult and, for the moment, no general approach seems to exist.

If A is a nonlinear operator from a Hilbert space X into a Hilbert space Y, we have the equation

$$g = A(u). \tag{6.1}$$

Obviously, a solution to this equation exists if and only if g is in the range of the operator A.

A. Linearization

The simplest way of treating (6.1) is to try to linearize the problem. This is the case of a *differentiable operator* [40]. The nonlinear operator A has a first derivative at the point u_o if there exists a linear operator L_o: $X \to Y$ such that, for any $u \in X$,

$$\lim_{t \to 0} \frac{1}{t} [A(u_o + tu) - A(u_o)] = L_o u. \tag{6.2}$$

The operator L_o is called the first derivative of A at the point u_o and one usually writes

$$L_o = A'(u_o). \tag{6.3}$$

An operator which is differentiable at the point u_o is also continuous at that point.

If an approximation u_o of the solution of (6.1) is known and if the operator A is differentiable at u_o, then (6.1) can be approximated by the linear equation

$$\delta g_o = L_o \delta u_o \tag{6.4}$$

where $\delta g_o = g - A(u_o)$, $\delta u_o = u - u_o$, and L_o is the derivative of A at u_o. Obviously, the procedure is consistent if the solution δu_o of (6.4) is a "small" correction to the approximate solution u_o.

The procedure can be iterated. By means of the solution δu_o of (6.4), one gets a new approximation, $u_1 = u_o + \delta u_o$, of the true solution u. Then one considers the linear equation $\delta g_1 = L_1 \delta u_1$, where $L_1 = A'(u_1)$, $\delta g_1 = g - A(u_1)$, and $\delta u_1 = u - u_1$. By solving this equation one gets a new approximation $u_2 = u_1 + \delta u_1$ and so on. It is easily recognized, by writing (6.1) in the form $P(u) = 0$ with $P(u) = A(u) - g$, that this method is just an extension to functional equations of a method which, in the case of real equations, is known as Newton's method or the method of tangents. Such an extension is also known as the Newton–Kantorovich method and it is one of the few practical methods for the actual solution of a nonlinear functional equation.

The iterative algorithm can be put in the following form:

$$u_{n+1} = u_n + [A'(u_n)]^{-1}[g - A(u_n)], \tag{6.5}$$

and a simplified algorithm is given by

$$u_{n+1} = u_n + [A'(u_o)]^{-1}[g - A(u_n)]. \tag{6.6}$$

Sufficient conditions for the convergence of both iterative algorithms have been given [40].

They include the continuity of the inverse of the derivative of the operator A. In several problems this condition is not satisfied. It has been suggested [41] to use, at each

step of the algorithm, a regularized approximation of the inverse of the derivative of the operator A. Convergence results for such a modified algorithm are not yet available.

B. Generalized and Regularized Solutions

Extensions of regularization theory to ill-posed nonlinear problems have also been proposed: the case of nonlinear integral equations has been investigated by Tikhonov and an abstract approach is given by Morozov.

We assume that $A: X \to Y$ is a continuously differentiable operator, i.e., that A has a derivative at each point $u \in X$ and that this derivative is a linear continuous operator. However, even in the case of such a simplifying assumption, a well-developed theory of generalized inverses does not exist. One can introduce least-squares solutions of (6.1) by solving the variational problem

$$\|A(u) - g\|_Y = \text{minimum} \tag{6.7}$$

analogous to the problem (3.4). When a solution of such a problem exists for any $g \in Y$, one says that (6.1) is strictly normally solvable. A sufficient condition for strict normal solvability is that the range of A is weakly closed in Y [43]. Notice that this condition may be stronger than the condition of closure of range that applies to the case of linear operators (Section III). Weakly closed sets are (strongly) closed, but the converse is not always true.

If, for a given g, the set of least squares solutions is not empty, one could try to select one of these solutions by means of another variational principle as in Section III-A, i.e., by minimizing a norm or seminorm such as (3.11). In contrast to the case where the operator A is linear, the generalized or C-generalized solution defined in such a way may not exist and, even if it does exist, is not necessarily unique. Such a lack of uniqueness applies also to the case of regularized solutions (in which case, however, existence can easily be assured).

The basic point in the definition of regularized solutions is again the minimization of a functional similar to (5.3); i.e.,

$$\Phi_\lambda[u] = \|A(u) - g\|_Y^2 + \lambda \|Cu\|_Z^2. \tag{6.8}$$

The uniqueness of the minimum of $\Phi_\lambda[u]$ usually is not proven (but see [42] for a special case where uniqueness holds true). However, it is not difficult to prove the existence of at least one local minimum.

Assume that the operator $A: X \to Y$ is continuous everywhere and that the constraint operator $C: X \to Z$ is *linear* and has a compact inverse C^{-1}. (This condition is satisfied, for instance, by the differential operator (5.4).) Then, *for any* $\lambda > 0$, *the functional (6.8) has at least one minimum point* u_λ. The proof of this result can easily be done just by adapting to the general case the proof given in [42] for the case of nonlinear integral equations.

As stated above, in general nothing can be said about the uniqueness of the minimum of the functional (6.8). However, if we assume that:

a) for a given g, (6.1) has a unique solution u in the domain of C;
b) in a neighborhood of u, the operator A has everywhere continuous first and second derivatives;
c) the derivative of A at u, $A'(u)$, is invertible;

then, by a rather easy generalization of the theorems contained in [42], one can prove that if g_ϵ is a noisy data, with $\|g - g_\epsilon\|_Y \leq \epsilon$, and if in the functional (6.8), with g replaced by g_ϵ, we choose the regularization parameter λ in such a way that $\lambda = \gamma\epsilon^2$, where γ is an arbitrary constant, then any minimum point of such a functional converges to u when $\epsilon \mapsto 0$; therefore, for sufficiently small values of ϵ, there exists only one minimum point.

Part Two

In this part we will consider several problems in early vision in the light of the mathematical results outlined in Part One. We will discuss edge detection, computation of optical flow, surface reconstruction, shape from shading, and stereo matching. Lastly we will discuss learning. Several of these problems have recently been solved using smoothness constraints or variational techniques, without an explicit reference to regularization theory. We will show that many of these results and several new ones, in particular existence and uniqueness of solutions, are direct consequences of the mathematical results of regularization theory presented in Part One.

The different modules that are part of early vision may reflect separate processing stages occurring in the brain, where we simultaneously make use of different visual procedures: we can extract sharp changes in image brightness (edge detection); we can understand the motion of objects from the changing images (computation of optical flow); we recover the 3-D structure of a scene from a pair of images (structure from stereo); and we can construct a dense description of 3-D surfaces from sparse features (visual surface interpolation).

As we mentioned in the introduction, problems in early vision are ill-posed because the available information is not sufficient to obtain a good solution, i.e., one which is physically correct and robust against noise. In this context regularization theory represents the correct tool for extracting the available information. Caution, however, is necessary: regularization theory can provide optimal techniques to reduce the effects of noise but cannot produce new information if it is not originally available. As we will see, edge detection, or numerical differentiation, is an ill-posed problem and there is little doubt that regularization theory is very useful in solving it. When we discuss the computation of optical flow, we will show that recent results [44]–[46] can also be seen as straightforward consequences of regularization methods, but we will also show that a better solution to the optical flow problem can be obtained by a more appropriate use of the available image data without relying exclusively on regularization theory. On the other hand even this solution needs to be regularized because the optical flow that it delivers is typically noisy and occasionally sparse.

Part Two is divided into five sections, each dealing with one module of early vision. In Section VII we present the ill-posed nature of numerical differentiation and of edge detection. In Section VIII we discuss the computation of optical flow. In Section IX we discuss recent approaches to surface interpolation, illustrating how variational principles can be viewed as regularized solutions to discrete ill-posed problems. In Section X we review recent variational approaches to shape from shading, in the framework of regularization theory. In Section XI we discuss stereoscopic vision.

VII. Edge Detection and Numerical Differentiation

Edge detection [47]-[51] is a key first step in early vision. This apparently simple problem of measuring sharp brightness changes in the image has proved to be difficult. It is now clear that edge detection should indicate not simply finding "edges" in the image, an ill-defined concept in general, but also measuring appropriate derivatives of the brightness data. This involves the task dependent use of different two-dimensional derivatives. In many cases, it is appropriate to mark locations corresponding to some critical points of the specific derivative such as its maxima or zeros. In some cases, later algorithms based on these binary features—presence or absence of edges—may be equivalent or very similar to algorithms that directly use the continuous value of the derivatives. From this point of view the low level problem commonly called edge detection consists of a) choosing a differential operator appropriate for the later tasks (say stereo), and b) computing correct and stable numerical derivatives of the image data.

Regularization theory is capable of indicating optimal ways for obtaining good numerical derivatives but cannot suggest the best differential operator. The choice of the differential operator depends on geometrical and topological properties of detected edges. In Part One we have already seen why numerical differentiation is ill-posed.

A. Regularization of Differentiation

Possibly the most natural use of regularization for the case of numerical differentiation is to interpolate or approximate the data with an analytic function and subsequently to compute the analytical derivative of the interpolating or approximating function [52].

Consider a one-dimensional "image" $y_i = f(x_i) + \epsilon_i$, where y_i is the data and ϵ_i represent errors in the measurement. We want to estimate f so we choose a Tikhonov stabilizing functional $\| Cf \|^2 = \int (f''(x))^2 \, dx$, where f'' is the second derivative of f. Physically, this choice corresponds to a constraint of smoothness on the intensity profile. Its physical justification is that the (noiseless) image is smooth because of the imaging process: the image is a bandlimited function and, therefore, has bounded derivatives. We look for an approximating function f minimizing

$$\sum_i (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 \, dx. \tag{7.1}$$

When the data are given on a regular grid and satisfy appropriate boundary conditions the solutions can be obtained by convolution with an appropriate filter R [52]. Differentiation can then be accomplished by convolutions of the data with the appropriate derivative of the filter. The resulting filters, which are spline filters for discrete data and Butterworth-like filters for continuous data (the two become indistinguishable in practice) are very similar to the derivatives-of-a-gaussian extensively used in recent years [52].

In the 2-D case, if the regularizing functional $\| Cf \|$ is

$$\int \int (\nabla^2 \operatorname{grad} f)^2 \, dx \, dy \tag{7.2}$$

where ∇^2 indicates the Laplacian and grad $f(x, y)$ the gradient of $f(x, y)$, then it has been shown [52] that the solution $f(x, y)$ can be obtained by convolving the data $g(x, y)$ with the filter

$$R_2(x, y) = \frac{1}{2} \int_0^\infty \frac{J_0(\omega z)}{\lambda \omega^6 + 1} \, \omega \, d\omega \tag{7.3}$$

where J_0 is the zero order Bessel function and $z = \sqrt{x^2 + y^2}$.

Differentiation can also be regularized using the filtering techniques described in Section V-B. Then, in the case of the inversion of the operator (3.2), condition c) of Section V-B is equivalent to requiring that the filter function $\Phi(\lambda; \xi)$ is such that $i\xi \, \Phi(\lambda; \xi)$ is a bounded function of ξ for any $\lambda > 0$. Therefore, these regularizing filters are essentially low pass filters. Three main types of filtering have been used in computer vision to perform edge detection. We list their main properties below.

B. Band-Limited, Support-Limited, and Minimal Uncertainty Filters

Band-limited filters are an obvious choice for regularizing differentiation, since the simplest way to avoid harmful noise is to filter out high frequencies that are amplified by differentiation. Linear and circular prolate functions constitute an interesting class of band-limited filters [53], [54] and have already been used in edge detection [50]. These filters satisfy all conditions of Tikhonov needed to regularize differentiation if we take the inverse of the bandwidth as the regularization parameter.

All real filters have a finite extent and are support-limited. A class of support-limited filters that has been used in edge detection [47] is the so-called difference of boxes (DOB). These filters are Haar functions [55] that form a basis for square integrable functions on a bounded interval. However, these filters do not satisfy condition c) of Section V-B and therefore cannot be used to regularize differentiation. This conclusion derives from the fact that the Haar functions are discontinuous. As a consequence, the limit of their Fourier transform as ξ goes to infinity tends to zero as ξ^{-1}. It is possible, however, to introduce smooth support-limited filters whose Fourier transform tends to 0 as desired as $\xi \to \infty$. If the inverse Fourier transform of the filter $\phi(\lambda; x)$ has, for instance, continuous derivatives up to order p and the $(p + 1)$th derivative is integrable, then $\Phi(\lambda, \xi)$ tends to zero as $|\xi|^{-(p+1)}$. Furthermore, if $\phi(\lambda; x)$ is C^∞, then $\Phi(\lambda; \xi)$ tends to zero more rapidly than any inverse power of ξ. An example is provided by the function

$$\phi(\lambda; x) = \begin{cases} C_\lambda \exp \dfrac{1}{1 - (x/\lambda)^2}, & |x| < \lambda \\ 0, & |x| > \lambda \end{cases} \tag{7.4}$$

where C_λ is a constant such that $\Phi(\lambda; 0) = 1$. Therefore, the best support-limited filter for edge detection and numerical differentiation is not the DOB but the filter (7.4), which is often used in digital signal processing when aliasing needs to be reduced.

The Gaussian function minimizes the product of spread in the space and the frequency domains [56] and can be viewed as a filter with minimal uncertainty. Filtering with a Gaussian function regularizes differentiation, because the Gaussian function $\phi(\lambda; x) = \exp(-x^2/2\lambda)$ satisfies all conditions of Section V-B. Moreover, filtering with a Gaussian transforms a continuous and bounded function into an entire function.

Therefore numerical differentiation can be regularized in a number of ways that are all consequences of the results presented in Part One. There are two main possibilities: filtering the data with appropriate derivatives of Tikhonov filters; or interpolating (or approximating) the discrete data with splines and then performing an analytical derivation. These two regularizing procedures are equivalent.

C. The Differential Operator in Edge Detection

Edges [51] in real images can be detected either as maxima of a first-order derivative or as zeros of a second-order derivative. In a two-dimensional image edges detected as maxima or as zeros have different geometrical properties. Fig. 1 shows an image from which the edges shown in Fig. 2 were extracted.

Fig. 1. An image of an interior in the Department of Physics.

The original image was first smoothed by the convolution with a 2-D symmetrical Gaussian function with a small value of λ (Fig. 2(a) and (b)) and a larger value of λ (Fig. 2(c) and (d)). Edges in (b) and (d) were extracted as zeros of the Laplacian and in (a) and (c) as maxima of the first-order derivative in the direction of the gradient image brightness.

In order to have a fair comparison between different schemes, edges were thresholded so as to have the same number of edges in all panels of Fig. 2. Two main observations argue in favor of a gradient scheme [49]:

i) A gradient scheme is generally more robust against noise because it uses only first-order derivatives and not second-order derivatives as a zero-crossing scheme. Since a zero of a second-order derivative does not necessarily coincide with an extremum of the first-order derivative, the null space of second-order derivatives is larger than the space of extrema of first-order derivatives, and therefore we expect a lower proportion of false edges in a gradient scheme. For these two reasons, edges detected as extrema of a first-order derivative are more reliable.

ii) A zero-crossing scheme, as shown in Fig. 2, cannot detect properly a trihedral vertex or a T-junction, because it introduces a spurious edge line. This behavior is a consequence of topological properties of zero-crossing contours that are intersections of structurally stable intersections of smooth surfaces [51], [57], [58]. Therefore we expect a better localization with a gradient scheme, which minimally distorts vertexes and junctions.

VIII. Computation of Optical Flow

The recovery of the motion of visible surfaces is a major task of both biological and artificial vision systems. The recovery of motion can be used to obtain a variety of additional information about the viewed scene, for example, depth, by using parallax effects, and the segmentation of the surrounding world into regions corresponding to distinct rigid objects.

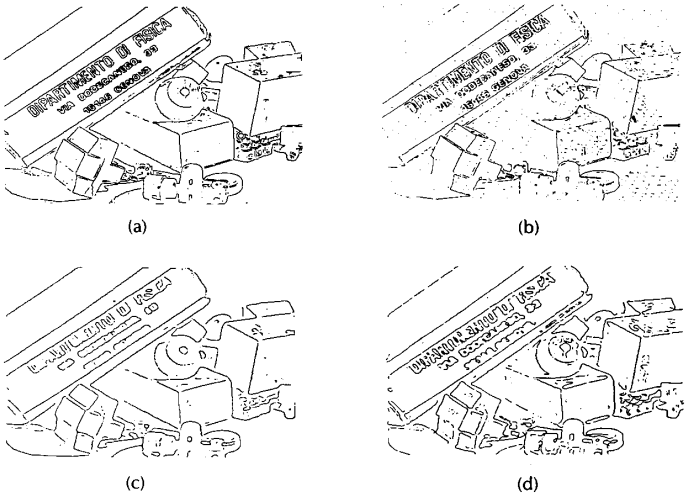


Fig. 2. Edges extracted either as zero-crossings of the original image of Fig. 1 convolved with the Laplacian of a Gaussian filter (b) and (d), or as maxima of the directional derivative in the gradient direction (a) and (c). In (a) and (b) a gaussian function with $\lambda = 1$ was used and in (c) and (d) with $\lambda = 3$. The threshold on edges was set so to have the same number of edges in each panel.

The motion of visible surfaces originates a 3-D velocity field which is projected by the imaging device into a 2-D velocity field. In order to recover information about the 3-D velocity field from the changing images, several authors [44]–[46] have introduced the notion of optical flow, defined as the distribution of apparent velocities of movement of brightness patterns in an image. There is no *a priori* reason to guarantee that the optical flow, as defined by these authors, and the 2-D velocity field are similar and even in some way related. It has been shown recently [59] that the various definitions of the optical flow and the "true" 2-D velocity field coincide only under very special conditions.

The recovery of the optical flow is usually regarded as plagued by the aperture problem [60]: when a straight moving edge is observed through a narrow aperture only the component of motion perpendicular to the edge can be measured. This view has been formalized in an extreme form by the elegant approaches of Horn and Schunck, and Hildreth.

Horn and Schunck [46] derived equations relating the change in image brightness $E(x, y, t)$ at a point $\{x, y\}$ and time t to the motion of brightness pattern. Their key definition is that the brightness of a particular point in the moving pattern is constant, so that the total derivative of $E(x, y, t)$ is zero:

$$\frac{dE}{dt}(x, y, t) = 0. \tag{8.1}$$

Then, from local measurements of the partial derivatives of $E(x, y, t)$ with respect to space coordinates and time, it is possible to estimate the component of the velocity field parallel to the gradient of $E(x, y, t)$. In this definition, the normal component is never determined, even in the case of edges that are not straight, and it must be recovered (see [59] for an analysis of the validity of the underlying assumptions).

Hildreth [44], [45] suggested computing the optical flow not over the entire image but only along 1-D contours. In real images, these 1-D contours are edges corresponding to sharp changes in image brightness (see Section VII). Hildreth [44], [45] observed that it was possible to obtain the normal vectors along the contour by a simple inspection of the extracted edges: if $E(x, y, t)$ is again the image brightness, then the normal component $v^\perp$ of the local velocity vector $\vec{V}$ at the points of the contour Γ is given by

$$v^\perp = \frac{\partial}{\partial t} \nabla^2 E \Big|_\Gamma \tag{8.2}$$

where ∇^2 is the Laplacian. A better estimate of $v^\perp$, however, is

$$v^\perp = \frac{\partial}{\partial t} \frac{\partial^2 E}{\partial n^2} \Big|_\Gamma \tag{8.3}$$

where $\partial^2/\partial n^2$ is the second derivative along the direction of the gradient [51].

In Sections VIII-A–D we discuss these two approaches in the framework of regularization theory. In Section VIII-E we show how a different approach better exploiting the available information can circumvent this extreme form of the aperture problem and is able to provide a 2-D vector field very close to the true 2-D velocity field.

A. Optical Flow Along a Contour

We first consider the problem of determining the two-dimensional optical flow along a contour Γ in the image assuming that local motion measurements along the contour provide only the component of the velocity in the direction perpendicular to the contour. We assume that the component of velocity tangential to the contour is invisible to local detectors that examine a restricted region of the contour. The local velocity vector $\vec{V}(s)$ is decomposed into a perpendicular and a tangential component to the curve

$$\vec{V}(s) = v^\top(s)\, \vec{t} + v^\perp(s)\, \vec{n}. \tag{8.4}$$

Here s is the arclength and $\vec{t}, \vec{n}$ are unit vectors respectively tangent and normal to the contour Γ

$$\vec{t} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \qquad \vec{n} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \tag{8.5}$$

where θ is the angle between $\vec{t}$ and the unit vector of the x axis. They depend also on s, but we omit this dependence for simplicity of notation.

The component $v^\perp(s)$ and the vectors $\vec{t}, \vec{n}$ are given by direct measurements and, therefore, are the data of the problem. We will denote by $g(s)$ the measured values of $v^\perp(s)$ and by $\vec{g}(s)$ the corresponding velocity field

$$\vec{g}(s) = g(s)\, \vec{n}. \tag{8.6}$$

Then the problem can be formulated as the inversion of a projection operator in the space $X = Y = L^2(\Gamma) \oplus L^2(\Gamma)$ ($L^2(\Gamma)$ denotes the space of square integrable functions defined over Γ). The norm of a velocity field $\vec{V} \in X$ is defined by

$$\begin{aligned} \|\vec{V}\|_X^2 &= \int_\Gamma \vec{V}(s) \cdot \vec{V}(s)\, ds \\ &= \int_\Gamma |v\ (s)|^2\, ds + \int_\Gamma |v^\perp(s)|^2\, ds. \end{aligned} \tag{8.7}$$

The projection operator is

$$L\vec{V}(s) = v^\perp(s)\, \vec{n}, \tag{8.8}$$

and the set of the solutions of the equation

$$L\vec{V} = \vec{g} \tag{8.9}$$

with $\vec{g}$ defined by (8.6), is the set of the velocity fields $\vec{V}$ given by

$$\vec{V}(s) = \psi(s)\, \vec{t} + g(s)\, \vec{n} \tag{8.10}$$

where $g(s)$ is the given data function and $\psi(s)$ is an arbitrary function in $L^2(\Gamma)$. The generalized solution, or solution of minimal norm, exists for any data function $g(s)$, but it is trivial since it is given by

$$\vec{V}^+ = \vec{g}. \tag{8.11}$$

In other words, the generalized solution restores well-posedness, but it gives a solution that does not have any physical relevance. Therefore, one has to look for suitable C-generalized solutions corresponding to physically acceptable velocity fields.

B. A C-Generalized Solution for the Optical Flow Along a Contour

A seminorm introduced by Hildreth gives a very useful constraint for the recovery of the optical flow. Put $Z = X = L^2(\Gamma) \oplus L^2(\Gamma)$ and introduce the operator

$$C\vec{V} = \dot{\vec{V}} \tag{8.12}$$

where the dot means derivation with respect to s. Then the C-generalized solution is the velocity field of the form (8.10) that minimizes the functional

$$\|C\vec{V}\|_X^2 = \int_\Gamma \dot{\vec{V}} \cdot \dot{\vec{V}}\, ds. \tag{8.13}$$

It is easy to show that existence and uniqueness of the C-generalized solution can be derived from the general result given in Section III-C.

First, consider the question of uniqueness. We know that the C-generalized solutions is unique if and only if the intersection of $N(C)$ and $N(L)$ is the null element (condition i) of Section III-C). Now $N(C)$ is the set of the constant velocity fields (or translations), say $\vec{V} = \vec{a}$. Furthermore, $N(L)$ is the set of the velocity fields orthogonal everywhere to $\vec{n}$, i.e., $\vec{v} \cdot \vec{n} = 0$. This condition can be satisfied by $\vec{a} \neq 0$ only if $\vec{n}$ is constant; that is, only if Γ is a straight line. Therefore if Γ is not a straight line, the intersection of $N(C)$ and $N(L)$ is always the null element, and uniqueness is restored by the use of the C-generalized solution (8.13).

The existence of the solution follows from the fact that the operator (8.12) satisfies conditions ii) and iii) of Section III-C. Condition ii) is a rather general property of differential operators, and condition iii) is also verified because $N(C)$ is a two-dimensional subspace of $X = L^2(\Gamma) \oplus L^2(\Gamma)$. Therefore, we can conclude that the C-generalized solution exists whenever $\vec{g} \in LD(C)$.

In order to see more precisely the meaning of the last condition, assume that the contour Γ consists of a finite number of regular arcs, so that the tangent is continuous on Γ with the exception of a finite number of points, $s_1, s_2, \cdots, s_n$, where the tangent has both right and left limits. Then a solution $\vec{V}(s)$ of the form (8.10) is in the domain of the constraint operator C if $\psi(s)\vec{t}$ and $g(s)\vec{n}$ are differentiable on each regular arc and furthermore they satisfy suitable conditions at the discontinuity points s_i in order to ensure the continuity of $\vec{V}(s)$. We can derive these conditions from the equations

$$\vec{V}_+(s_i) = \vec{V}_-(s_i), \qquad i = 1, \cdots, n \tag{8.14}$$

where $+$ and $-$ denote respectively right and left limit. It follows that

$$\begin{aligned} \psi_+(s_i) &= (\sin \phi_i)^{-1}[g_+(s_i) \cos \phi_i - g_-(s_i)] \\ \psi_-(s_i) &= (\sin \phi_i)^{-1}[g_+(s_i) - g_-(s_i) \cos \phi_i] \end{aligned} \tag{8.15}$$

where $\sin \phi = \vec{t}_- \cdot \vec{n}_+ = -\vec{t}_+ \cdot \vec{n}_-$, $\cos \phi = \vec{n}_+ \cdot \vec{n}_- = \vec{t}_+ \cdot \vec{t}_-$ ($\vec{t}_+$ is the right limit of the tangent, etc.). Therefore, if $g(s)$ admits a right and left limit at the points s_i, it is possible to derive from (8.15) the right and left limit of ψ. All these conditions characterize the subset $D(C)$ which contains the unique solution that minimizes the seminorm (8.13). Of course, if $\vec{g}$ is not differentiable on the regular arcs or does not have left and right limits at the discontinuity points, the C-generalized solution does not exist. It follows that the problem is ill-posed.

Before discussing this point, we want to point out that if the data $\vec{g}$ are not affected by noise, the C-generalized solution coincides with the true solution in two important cases [44], [45]: the first is translation of an arbitrary contour and the second is arbitrary motion of a rigid polygon. These results can be derived from the Euler equation for the C-generalized solution.

Assume that the regular arcs have a differentiable curvature. From the following relations, which are true on each regular arc

$$\dot{\vec{t}} = \dot{\theta}\vec{n}, \qquad \dot{\vec{n}} = -\dot{\theta}\vec{t} \tag{8.16}$$

where $\dot{\theta}$ is just the curvature, one can derive from (8.10)

$$\dot{\vec{V}}(s) = [\dot{\psi}(s) - \dot{\theta}(s)\, g(s)]\vec{t} + [\dot{g}(s) + \dot{\theta}(s)\, \psi(s)]\vec{n} \tag{8.17}$$

and therefore, when ψ satisfies the conditions (8.15)

$$\begin{aligned} \|C\vec{V}\|_X^2 = &\int_\Gamma \{|\dot{g}(s)|^2 + |\dot{\theta}(s)\, g(s)|^2\}\, ds \\ &+ \int_\Gamma \{|\dot{\psi}(s)|^2 + |\dot{\theta}(s)\, \psi(s)|^2 + 2\dot{\theta}(s)\, \dot{g}(s)\, \psi(s) \\ &- 2\dot{\theta}(s)\, g(s)\, \dot{\psi}(s)\}\, ds. \end{aligned} \tag{8.18}$$

This is a functional of ψ, which is an arbitrary function except for being differentiable and satisfying conditions (8.15). Then, by annihilating the first variation of this functional, it follows that, on each regular arc, the function ψ which minimizes the functional is a solution of the differential equation

$$-\ddot{\psi}(s) + |\dot{\theta}(s)|^2\psi(s) + 2\dot{\theta}(s)\, \dot{g}(s) + \ddot{\theta}(s)\, g(s) = 0. \tag{8.19}$$

In the case of a closed contour, the C-generalized solution is given by the unique solution of (8.19) satisfying the conditions (8.15). If the contour is regular everywhere, then one has to add boundary conditions such as

$$\psi(0) = \psi(l), \qquad \dot{\psi}(0) = \dot{\psi}(l). \tag{8.20}$$

When the contour is open, one needs boundary conditions at the end points of the contour. These can be obtained directly, through a partial integration, from the annihilation of the first variation of (8.18)

$$\dot{\psi}(0) = \dot{\theta}(0)\, g(0), \qquad \dot{\psi}(l) = \dot{\theta}(l)\, g(l). \tag{8.21}$$

However, these conditions are correct only in the case of pure translation. In the general case it is necessary to measure the tangential velocity of the endpoints and take

$$\psi(0) = v\ (0), \qquad \psi(l) = v\ (l) \tag{8.22}$$

where $v\ (0)$ and $v\ (l)$ are the measured values.

If the motion of the contour is pure translation and

$$\vec{a} = \{a_1, a_2\} \tag{8.23}$$

is the constant velocity field, the noise-free data are given by

$$g(s) = -a_1 \sin \theta + a_2 \cos \theta. \tag{8.24}$$

Then, if we put

$$\psi(s) = a_1 \cos \theta + a_2 \sin \theta, \tag{8.25}$$

taking into account that $\dot{\psi} = \dot{\theta}g$ and $\dot{g} = -\dot{\theta}\psi$, it is easy to verify that ψ satisfies (8.19). In the case of an open contour, also the boundary conditions (8.21) are satisfied (the boundary conditions (8.15) are obvious since the velocity field is continuous).

In the case of a rigid polygon [61], since arbitrary rigid motion consists of translation plus rotation, on each segment of the polygon both the normal and tangent velocity are linear functions of the arclength s. But, on a segment of a straight line, (8.19) becomes $\dot{\psi}(s) = 0$ and therefore $\psi(s)$ is a linear function of s. The boundary conditions (8.20) (plus the boundary condition (8.22) in the case of an open polygon) give the correct values of the constants provided that also in this case the measured values are noise-free.

As we already remarked, the difficulty of this approach is that the problem of determining such a C-generalized solution is ill-posed. For this reason, in the case of noisy data, one has to look for a regularized approximation of the C-generalized solution, which can be obtained by minimizing the functional [44], [45]

$$\Phi_\lambda[\vec{V}] = \|L\vec{V} - \vec{g}\|_X^2 + \lambda\|C\vec{V}\|_X^2. \tag{8.26}$$

If we denote by $\vec{V}_\lambda$ the minimum of the functional (8.26) and if we put

$$\vec{V}_\lambda = \psi_\lambda(s)\,\vec{t} + \phi_\lambda(s)\vec{n} \tag{8.27}$$

then it is easy to show that, on each regular arc, ψ_λ and ϕ_λ must be solutions of the system of differential equations

$$-\ddot{\psi}_\lambda(s) + 2\dot{\theta}(s)\,\dot{\phi}_\lambda(s) + |\dot{\theta}(s)|^2\,\psi_\lambda(s) + \ddot{\theta}(s)\,\phi_\lambda(s) = 0 \tag{8.28a}$$

$$-\lambda[\ddot{\phi}_\lambda(s) + 2\dot{\theta}(s)\,\dot{\psi}_\lambda(s) + |\dot{\theta}(s)|^2\,\phi_\lambda(s) + \ddot{\theta}(s)\,\psi_\lambda(s)] + \phi_\lambda(s) = g(s) \tag{8.28b}$$

plus boundary conditions similar to those discussed in the previous case (continuity of $\vec{V}_\lambda$ at the discontinuity points, etc). The determination of the parameter λ can be performed using one of the methods discussed in Section V.

In practice, the most economical method for the computation of $\vec{V}_\lambda$ is perhaps the conjugate gradient method. Regularizing properties of this method [62], [63] can also be used in order to avoid the minimization of (8.26).

In the previous treatment we have neglected the errors in the determination of the contour which imply an approximate knowledge of the operator L (8.8). However, if the equation $L\vec{V} = \vec{g} + \delta\vec{g}$, where $\delta\vec{g}$ is the error on the data, is replaced by the equation $(L + \delta L)\vec{V} = \vec{g} + \delta\vec{g}$, where δL is the error on the operator, it appears that the two equations are equivalent in the sense that only the error on $\vec{g}$ is different in the two cases (in one case it is $\delta\vec{g}$ and in the other case it is $\delta\vec{g} - (\delta L)\vec{V}$). This point of view assumes that the errors in the determination of the contour have been included in the errors on the data.

C. Two-Dimensional Optical Flow

As we already recalled at the beginning of this section, Horn and Schunck [46] attempted to recover the optical flow in the entire image and not just on a one-dimensional contour. Their basic equation is (8.1), which, written explicitly, provides the relationship

$$\vec{\nabla}E \cdot \vec{V} = -\partial_t E \tag{8.29}$$

where $\vec{\nabla}E = \{\partial_x E, \partial_y E\}$ is the gradient of the brightness distribution in the image, $\vec{V}$ is the velocity field (optical flow), and $\partial_t E$ is the partial time derivative of the brightness. Therefore, a measurement of ∇E and $\partial_t E$ gives the component of $\vec{V}$ parallel to $\vec{\nabla}E$.

We assume that the brightness distribution $E(x, y, t)$ is defined in a bounded region Ω whose boundary $\partial\Omega$ is a contour with an everywhere continuous tangent. Furthermore, we will also assume, for simplicity, that $\vec{\nabla}E$ is never zero in Ω and that the level lines of $E(x, y, t)$ have everywhere differentiable tangents and normals. We denote by $\vec{t}$ and $\vec{n}$ the tangent and normal to the level line at the point $\{x, y\}$

$$\vec{t} = |\vec{\nabla}E|^{-1}\begin{pmatrix}\partial_y E\\ -\partial_x E\end{pmatrix}, \qquad \vec{n} = |\vec{\nabla}E|^{-1}\begin{pmatrix}\partial_x E\\ \partial_y E\end{pmatrix}. \tag{8.30}$$

Then the velocity field $\vec{V}(x, y)$ can be everywhere represented as follows

$$\vec{V}(x, y) = v\,(x, y)\,\vec{t} + v^\perp(x, y)\vec{n}. \tag{8.31}$$

The problem can again be formulated as the inversion of a projection operator: taking $X = Y = L^2(\Omega) \oplus L^2(\Omega)$ and

$$(L\vec{V})(x, y) = v^\perp(x, y)\vec{n}, \tag{8.32}$$

the data will be given by

$$\vec{g}(x, y) = g(x, y)\vec{n} \tag{8.33}$$

where $g(x, y)$ is the measured value of $-\partial_t E/|\vec{\nabla}E|$. Then the set of solutions of the equation $L\vec{V} = \vec{g}$ is the set of velocity fields

$$\vec{V}(x, y) = \psi(x, y)\,\vec{t} + g(x, y)\vec{n} \tag{8.34}$$

where ψ is an arbitrary function in $L^2(\Omega)$. The generalized solution $\vec{V}^+$ is trivial also in this case, since $\vec{V}^+ = \vec{g}$.

D. A C-Generalized Solution for the Two-Dimensional Optical Flow

As in the case of the optical flow along a contour, it is necessary to look for C-generalized solutions. The method suggested in [46] can be formulated in this framework.

Introduce the constraint space $Z = X \oplus X$ and define an operator C: $X \mapsto Z$ as

$$C\vec{V} = \begin{pmatrix}\partial_x\vec{V}\\ \partial_y\vec{V}\end{pmatrix} \tag{8.35}$$

with the associated seminorm

$$\|C\vec{V}\|_Z^2 = \int_\Omega \{\partial_x\vec{V} \cdot \partial_x\vec{V} + \partial_y\vec{V} \cdot \partial_y\vec{V}\}. \tag{8.36}$$

Written in terms of the cartesian components of $\vec{V}$ this is just the integral of the quantity called the measure of the departure from smoothness in the velocity flow [46].

First consider the question of uniqueness. The null space $N(C)$ is the set of the constant velocity fields, say $\vec{V} = \vec{a}$, while the null space $N(L)$ is the set of the velocity fields which are orthogonal everywhere to $\vec{n}$, i.e., $\vec{V} \cdot \vec{n} = 0$. The intersection is the set of constant velocity fields such that $\vec{a} \cdot \vec{n} = 0$ and this condition cannot be satisfied by $\vec{a} \neq 0$ if the level lines are not parallel straight lines everywhere.

It is easy to verify that conditions i)–iii) of Section III-B are satisfied and the existence of the solution is guaranteed. It may be interesting however to write the Euler equation for the C-generalized solution. After some lengthy but elementary computations, using the orthogonality relations $\vec{n} \cdot \partial_x\vec{n} = \vec{n} \cdot \partial_y\vec{n} = 0$, $\vec{t} \cdot \partial_x\vec{t} = \vec{t} \cdot \partial_y\vec{t} = 0$, $\partial_x\vec{n} \cdot \partial_x\vec{t} =$

0 we obtain

$$\|C\vec{V}\|_Z^2 = \int_\Omega \{|\vec{\nabla} g|^2 + (|\partial_x \vec{n}|^2 + |\partial_y \vec{n}|^2)|g|^2\}\, dx\, dy + \int_\Omega \{|\vec{\nabla}\psi|^2 + (|\partial_x \vec{t}|^2 + |\partial_y \vec{t}|^2)|\psi|^2 + 2[(\vec{n} \cdot \partial_x \vec{t})\partial_x g + (\vec{n} \cdot \partial_y \vec{t})\partial_y g]\psi + 2g[(\vec{t} \cdot \partial_x \vec{t})\partial_x \psi + (\vec{t} \cdot \partial_y \vec{n})\partial_y \psi]\}\, dx\, dy. \tag{8.37}$$

In order to find the Euler equation of the functional (8.37) one has to consider a variation of ψ, $\psi \mapsto \psi + h$ and put equal to zero the term of first order in h. Then, using the divergence theorem in order to eliminate the partial derivatives of h, transforming the fourth term in (8.37) by means of the identities $\vec{t} \cdot \partial_x \vec{n} = -\vec{n} \cdot \partial_x \vec{t}$, $\vec{t} \cdot \partial_y \vec{n} = -\vec{n} \cdot \partial_y \vec{t}$, and using the fact that h is arbitrary, one finds that the unique function ψ that minimizes the functional (8.37) is the unique solution of the boundary value problem:

$$\nabla\psi + (|\partial_x \vec{t}|^2 + |\partial_y \vec{t}|^2)\psi + 2[(\vec{n} \cdot \partial_x \vec{t})\partial_x g + (\vec{n} \cdot \partial_y \vec{t})\partial_y g] + (\vec{n} \cdot \Delta \vec{t})g = 0 \tag{8.38}$$

$$\left.\frac{\partial \psi}{\partial \nu}\right|_{\partial\Omega} = \left.\left(\vec{n} \cdot \frac{\partial \vec{t}}{\partial \nu}\right) g\right|_{\partial\Omega} \tag{8.39}$$

where ν is the normal to $\partial\Omega$. Notice that this boundary value problem is just the extension in the 2-D case of the problem (8.19) with the boundary conditions (8.21). The boundary condition (8.39) can be replaced by the value of ψ if the tangent velocity can be measured on $\partial\Omega$.

It is also easy in the present case to verify that if the motion is pure translation (i.e., a constant velocity field), and if the data function is noise-free, then the C-generalized solution coincides with the exact velocity field.

It is also obvious that in this case the C-generalized solution is ill-posed and one must introduce regularized approximations. These can be obtained by minimizing the analog of the functional (8.26), and this is precisely the method used in [46].

E. A Solution to the Aperture Problem

In Section VIII-B we have seen that in the case of a rigid polygon [61] the C-generalized solution (8.13) gives the correct solution, possibly suggesting that the minimization of this functional captures some basic properties of rigid motion. Unfortunately this result has not been extended to the two-dimensional optical flow, and the use of C-generalized solution minimizing the functional (8.36) does not have obvious physical plausibility.

Familiarity with regularization theory may suggest to reconsider whether the original problem is really ill-posed. The available information is the time varying image brightness $E(x, y, t)$ from which we want to obtain a time-varying 2-D vector field as close as possible to the 2-D velocity field. The key point argued in [59] is that the definition of an optical flow is rather arbitrary and one cannot obtain the "true" velocity field but only an approximation to it, with the same qualitative properties.

When the problem is stated in this way, the aperture problem disappears because there are many 2-D vector fields which can be defined in terms of $E(x, y, t)$ without using C-generalized solutions. In particular, it has been recently shown that the 2-D vector field obtained by solving:

$$\frac{d}{dt} \operatorname{grad} E(x, y, t) = 0 \tag{8.40}$$

provides an excellent approximation to the 2-D velocity field [64]. Equation (8.40) can be rewritten as

$$\frac{\partial^2 E}{\partial x^2} v_x + \frac{\partial^2 E}{\partial x \partial y} v_y + \frac{\partial^2 E}{\partial x \partial t} = 0 \tag{8.41a}$$

$$\frac{\partial^2 E}{\partial x \partial y} v_x + \frac{\partial^2 E}{\partial y^2} v_y + \frac{\partial^2 E}{\partial y \partial t} = 0 \tag{8.41b}$$

where v_x and v_y are the two components of the optical flow. When $|\det \operatorname{Hess} E(x, y, t)|$ is different from zero, then it is possible to obtain from (8.41a) and (8.41b) the two components of the optical flow.

Equation (8.40) is a vector equation, not a scalar equation as (8.1), and does not have the aperture problem in the same extreme way as (8.1).

The use of the vector equation (8.40) can be justified in a number of ways:

i) Equation (8.40) gives the exact vector field of a rigid black 2-D pattern moving on the image plane.
ii) Equation (8.40) provides a vector field which is not usually equal to the true 2-D velocity field but is similar in the majority of cases.

Fig. 3 shows a sequence of four frames of a printed board translating towards the camera. Fig. 4 reproduces the 2-D vector field obtained from the sequence shown in Fig. 3 by using (8.40) and a further smoothing of the optical flow. It is evident that the obtained optical flow is very close to the true 2-D motion field. The exact definition of closeness is the one used in structural stability and it refers to topological properties of solutions [58].

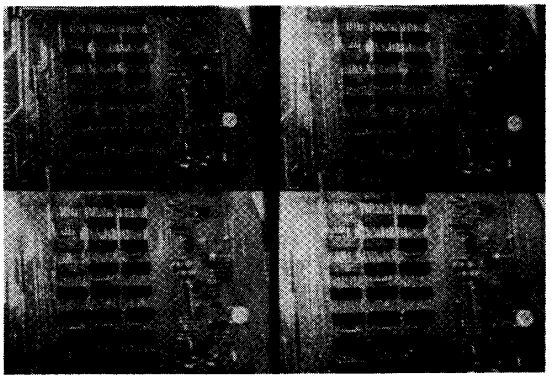
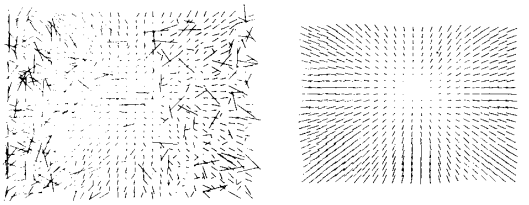


Fig. 3. A sequence of four images of a printed board translating towards the camera.

The use of (8.40) to compute the optical flow suggests that this problem is not ill-posed but may be ill-conditioned when $|\det \operatorname{Hess} E(x, y, t)|$ is very small.



(a) (b)

Fig. 4. The optical flow was obtained by first smoothing the images in Fig. 3 with a 2-D symmetrical gaussian function ($\lambda = 1$) and using (8.40) to obtain the two components of the optical flow (a). In (b) the optical flow shown in (a) was smoothed by the convolution with a 2-D symmetrical gaussian function ($\lambda = 5$). Derivatives were computed by using a Taylor series expansion [51].

IX. Surface Reconstruction

Most algorithms able to recover depth from pairs of stereo images [65]–[67] provide depth values only for special points in the viewed scene. This sparse 3-D map can be sufficient for many goals in robotics, such as navigation or recognition, where the redundant information does not require a very dense 3-D map. In many other cases such as in aerial photogrammetry or in terrain reconstruction a dense map is required. Therefore it is useful to consider the problem of recovering a visual surface $f(x, y)$ from 3-D sparse data.

A. Surface Interpolation

The original data are a finite set of depth values $z_i = f(x_i, y_i)$, $i = 1, \cdots, n$ (which are assumed to be exact; that is, noise-free) and the problem is the recovery of a smooth function $f(x, y)$ interpolating z_i at $(x_i, y_i) = t_i$ contained in Ω. Grimson [66], [67] proposed to find f such that it minimizes the seminorm

$$\|Cf\|^2 = \int \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right] dx\, dy. \quad (9.1)$$

Uniqueness of solution is guaranteed by the existence of at least four noncoplanar points $z_i = f(x_i, y_i)$ [66], [67].

This procedure can be seen as an application of generalized inverses in the case of discrete data (see Section III-D): in this case, uniqueness of the solution is guaranteed when the intersection of the null space of C, ($N(C)$) and the null space of $L(N(L))$ is empty, where L is the operator defined in Section III-D.

The null space $N(C)$ is composed of the set of functions $f(x, y) = ax + by + c$ with a, b, and c constants. These functions consist of all planar surfaces defined in Ω. The null space $N(L)$ has been defined in Section III-D and consists of the set of functions such that $f(x_i, y_i) = 0$ for $i = 1, \cdots, n$. Therefore, it is easy to see that when $i \geq 4$ and the points $t_i = (x_i, y_i)$ are distinct, the intersection of $N(C)$ and $N(L)$ is empty. In other words, uniqueness is guaranteed if there are at least four noncoplanar points, as required in [66], [67].

B. Surface Approximation with Noisy Data

It is also useful to consider the case in which the data are noisy, that is, when the original data are $g_i = f(t_i) + \epsilon_i$, $i = 1, \cdots, N$ and ϵ_i is additive noise. In this case, it is reasonable to look for a solution close to the original data g_i, but smooth. This approach can be seen as an application of regularization theory. In Part One we showed that interpolation is an ill-posed problem which can be solved by the use of a generalized inverse. We will now present an approach to interpolation directly originating from regularization theory [71], [72], which clarifies the relationship between splines, regularization theory, and gives a different framework to the results on visual interpolation [67]–[70].

We can consider the case in which we want to estimate a smooth function $f(t)$, $t \in \Omega \subset R^2$, given a finite number of observations of linear functionals of f. In the case of spatial interpolation, our functionals are

$$g_i = F_i(f) + \epsilon_i = f(t_i) + \epsilon_i, \qquad i = 1, \cdots, n \quad (9.2)$$

where ϵ_i is additive noise. A regularized estimate $f_{n,\lambda}$ is obtained by solving the minimization problem

$$\sum_{i=1}^{n} (f(t_i) - g_i)^2 + \lambda J_m(f) \quad (9.3)$$

in which $J_m(\cdot)$ is a seminorm in H_m (H_m is a reproducing kernel Hilbert space of functions defined in Ω) defined by

$$J_m(f) = \int \int_{-\infty}^{+\infty} \sum_{\nu=0}^{m} \binom{m}{\nu} \left(\frac{\partial^m f}{\partial x^\nu \partial y^{m-\nu}} \right)^2 dx\, dy, \quad (9.4)$$

(here m indexes the highest square integrable derivative) and λ controls the tradeoff between the degree of approximation of the solution to the data and the smoothness of solution. The value of λ can be computed by the method of generalized cross-validation [37]–[39]. If $m = 2$ we have the functional (9.1). The solution of this minimization problem is one of the "thin plate splines," so called because $J_2(f)$ is the bending energy of a thin plate.

In [71] it was shown that a unique solution exists for any $\lambda > 0$ provided:

1) $m > 1$;
2) $n \geq M = \binom{m+1}{2}$;
3) the "design" $t_1, \cdots, t_n$ is unisolvent, that is if $\{\phi_\nu\}_{\nu=1}^{n}$ is a basis for the M dimensional space of polynomials of total degree $m - 1$ or less, then $\Sigma_{\nu=1}^{n} \alpha_\nu \phi_\nu(t_i) = 0$ ($i = 1, \cdots, n$) implies that the $\alpha_\nu \equiv 0$.

If $m = 2$, then we need at least three points which do not lie on the same straight line (to satisfy the requirement of a unisolvent design), which is the same requirement as found in [66] and [67]. Moreover, the solution has an explicit representation [71] as:

$$f_{n,m,\lambda}(t) = \sum_{j=1}^{m} c_j E_m(t, t_j) + \sum_{\nu=1}^{n} d_\nu \phi_\nu(t) \quad (9.5)$$

where

$$E_m(s, t) = \theta_m |s - t|^2 \log |s - t| \quad (9.6)$$

with

$$s = (x_1, y_1) \qquad t = (x_2, y_2)$$

$$|s - t| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

and

$$\theta_m = 1/2^{2m-1}\pi[(m-1)!]^2. \tag{9.7}$$

The coefficients $c = (c_1, \cdots, c_n)$ and $d = (d_1, \cdots, d_n)$ are determined by the solution of the algebraic linear system:

$$\begin{aligned}(K + \rho)C + Td &= g \\ T^TC &= 0\end{aligned} \tag{9.8}$$

where K is the $n \times n$ matrix with $K_{jk} = E_m(t_j, t_k)$, $\rho = n\lambda$, T is the $n \times m$ matrix with $T_{\nu i} = \phi_\nu(t_i)$ and $g = (g_1, \cdots, g_n)$.

C. Surface Interpolation on a Regular Grid

While surface interpolation from sparse data requires an arbitrary grid of knots, other problems of machine vision require the approximation of a 3-D surface through points given on a rectangular grid. For example, when a smooth function f interpolating intensity values on the regular grid of a CCD camera is regularized, it is possible to use doubly cubic splines or a tensor product of splines, giving an interpolating function that minimizes

$$\int\int (\partial^4 f/\partial x^2 \partial y^2)^2 \, dx \, dy. \tag{9.9}$$

In this case different kinds of doubly cubic splines can be used, according to the available data [73]. The algorithms are then convolution algorithms (see Section VII-D).

X. Shape from Shading

It is a common experience to notice our ability to recover the shape of an object from its shading. Convexity or concavity of viewed objects are easily understood by looking at the profile of radiating light. Here we have another classical problem of early vision, "shape from shading," which has stimulated elegant mathematical approaches. The problem of shape from shading was initially formulated in [74], [75], and [78] as the solution of five ordinary differential equations called the characteristic strip equations. Of considerable use in this problem has been the reflectance map $R(p, q)$ [76], [77] which specifies the radiance of a surface patch as a function of its orientation, determined by the pair (p, q). If $z(x, y)$ is the surface of the object, p and q are defined as

$$p = \frac{\partial z}{\partial x} \quad \text{and} \quad q = \frac{\partial z}{\partial y}, \tag{10.1}$$

and the unit normal $\vec{n}$ to the surface is

$$\vec{n} = \frac{1}{\sqrt{1 + p^2 + q^2}} \{-p, -q, 1\}. \tag{10.2}$$

The reflectance map can be computed from the bidirectional reflectance-distribution function and the light source arrangement [77].

Formally, given an image $E(x, y)$ and a reflectance map $R(p, q)$, the shape from shading problem may be regarded as the recovery of a smooth surface $z(x, y)$ satisfying the image irradiance equation

$$E(x, y) = R\left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) = R(p, q) \tag{10.3}$$

over some domain Ω of the image. Since there are two unknown functions (p and q) and only one equation the solution is not unique and the problem is underconstrained (and ill-posed). Uniqueness of the solution can be recovered by the use of photometric stereo, which takes multiple images of the same scene from the same position with different illumination [78]. In this approach, several equations of the type of (10.3) are available, with different reflectance maps since the illumination source is different. Three different light sources can be used to obtain a unique solution.

If only one source of illumination is available, uniqueness can be restored by variational techniques similar to those previously seen. Assuming that the object has a Lambertian surface and is illuminated by a planar wave of light (and the unit vector $\vec{s} = (s_1, s_2, s_3)$ points to the light source), then the Lambertian reflectance map becomes

$$R(p, q) = \vec{n} \cdot \vec{s}. \tag{10.4}$$

If, instead of using the pair $\{p, q\}$, the new variables $\{f, g\}$ are introduced

$$f = \frac{2p}{1 + \sqrt{1 + p^2 + q^2}} \qquad g = \frac{2q}{1 + \sqrt{1 + p^2 + q^2}}, \tag{10.5}$$

the reflectance map becomes

$$R(f, g) = \frac{4 - (f^2 + g^2)}{4 + (f^2 + g^2)} \cdot \left(-\frac{4f}{4 - (f^2 + g^2)}, -\frac{4g}{4 - (f^2 + g^2)}, 1\right) \cdot \vec{s}. \tag{10.6}$$

The problem of shape from shading can be formulated either using the unknown $\vec{n}$ or the pair $\{p, q\}$ or $\{f, g\}$.

A. The Variational Approach to Shape from Shading

When the unknown $\vec{n}$ is used, the variational approach is to find $\vec{n}(x, y)$ such that it minimizes

$$\int_\Omega (E(x, y) - \vec{n} \cdot \vec{s})^2 \, dx \, dy + \lambda \int_\Omega \left(\frac{\partial \vec{n}}{\partial x}\right)^2 + \left(\frac{\partial \vec{n}}{\partial y}\right)^2 dx \, dy, \tag{10.7}$$

with the constraint $\|\vec{n}\| = 1$. In this case, the variational problem is quadratic in the unknown $\vec{n}$, but the constraint $\|\vec{n}\| = 1$ is unusual.

When the pair $\{f, g\}$ is used, we seek functions f and g minimizing:

$$\int_\Omega |(E(x, y) - R(f, g)|^2 \, dx \, dy + \lambda \int_\Omega \left[\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial y}\right)^2\right] dx \, dy, \tag{10.8}$$

with $R(f, g)$ given by (10.6). The variational problem is not quadratic in the unknown $\{f, g\}$ and the results of nonlinear inverse problems must be used.

B. Regularization of Shape from Shading

We give an application of the result stated in Section VI by formulating the problem in terms of the pair $\{p, q\}$. We define the space X as the direct sum $L^2(\Omega) \oplus L^2(\Omega)$, i.e., u is

a pair $\{p, q\}$ of square integrable functions:

$$\|u\|_X^2 = \int_\Omega p^2(x, y)\, dx\, dy + \int_\Omega q^2(x, y)\, dx\, dy. \quad (10.9)$$

Let the space Y be also a space of square integrable functions (we now call $g(x, y)$ the image $E(x, y)$), and from (10.2) and (10.4) we define a nonlinear operator A: $X \to Y$ as follows:

$$(Au)(x, y) = \frac{s_3 - ps_1 - qs_2}{\sqrt{1 + p^2 + q^2}}. \quad (10.10)$$

Because $\vec{n}$ and $\vec{s}$ are unit vectors, it is obvious that $|(Au)(x, y)| \leq 1$ for any $\{x, y\} \in \Omega$. It follows that the domain of A is X and that the range of A is contained in the set of $g(x, y)$ such that $|g(x, y)| \leq 1$ in Ω. Furthermore, it is not difficult to prove that the operator A is continuous everywhere, i.e., if u is any element of X and if $\{u_n\}$ is a sequence convergent to u, then Au_n converges to Au. Indeed, using the inequalities

$$|s_3 - ps_1 - qs_2| \leq \sqrt{1 + p^2 + q^2}, \sqrt{1 + p_n^2 + q_n^2} \geq 1, \quad (10.11)$$

it follows that

$$|Au - Au_n| \leq |s_1||p - p_n| + |s_2||q - q_n| + |\sqrt{1 + p^2 + q^2} - \sqrt{1 + p_n^2 + q_n^2}|. \quad (10.12)$$

Then, using the inequality $(q_1 + \cdots + q_n)^2 \leq n(q_1^2 + \cdots + q_n^2)$ (with $n = 2, 3$), we get

$$|Au - Au_n|^2 \leq (|p - p_n|^2 + |q - q_n|^2). \quad (10.13)$$

By integrating over Ω we get the continuity of the operator A.

Finally, we consider the constraint operator C defined by

$$\|Cu\|_Z^2 = \int_\Omega \left[c_o(p^2 + q^2) + \left(\frac{\partial p}{\partial x}\right)^2 + \left(\frac{\partial p}{\partial y}\right)^2 + \left(\frac{\partial q}{\partial x}\right)^2 + \left(\frac{\partial q}{\partial y}\right)^2 \right] dx\, dy \quad (10.14)$$

where c_o could take the value $c_o = 0$ and give the stabilizer used by Ikeuchi and Horn. We can seek a solution to the problem of shape from shading by minimizing the functional

$$\int_\Omega |(Au)(x, y) - g(x, y)|^2\, dx\, dy + \lambda \|Cu\|_Z^2 \quad (10.15)$$

where the first term in (10.15) is (10.10) and the constraint operator C is defined in (10.14). Because the operator A is continuous and the constraint operator has a compact inverse, the results presented in Section VI indicate the existence of at least a local minimum of the functional (10.15). Furthermore, if $g \in R(A)$, the u_λ converges to an exact solution when $\lambda \mapsto 0$. Note that the problem of uniqueness of the regularized solution remains open.

XI. Stereo Matching

Not all inverse problems of early vision can be solved using the regularizing techniques introduced in Part One. For example, stereopsis [65]–[67], which is the process that computes depth from two images of the same scene obtained by two eyes or cameras, appears as an inverse problem that may be approached with standard regularization techniques. It turns out that this is, however, quite difficult. The critical problem in stereopsis is the correspondence problem, that is, the matching of corresponding features in the two images. Let us consider the 1-D matching problem, by considering the intensity profile—or some corresponding feature map—on conjugated epipolar lines [67]. In this case, the obvious way to match the right image $R(x)$ with the left one $L(x)$ is to find the disparity $d(x)$ such that the two intensity profiles $L(x)$ and $R(x + d(x))$ are as close as possible. We can formalize this in the following way: let us define an operator P_R that depends on the image as

$$P_R f(x) \mapsto R(x + f(x)). \quad (11.1)$$

The disparity function that we want could be seen as the solution to the inverse problem:

$$L(x) = P_R d(x). \quad (11.2)$$

The operator in (11.2) which has to be inverted depends on the data and is not known *a priori*. This class of problems is not covered by the available mathematical results. We could still try to determine $d(x)$ by minimizing

$$\|L(x) - R(x + d(x))\|. \quad (11.3)$$

A sufficient condition for the solution of (11.3) to be unique is that $L(x)$ and $R(x)$ are strictly monotonic functions of x. This is clearly a very restrictive condition, almost never satisfied by real images. In general, the problem admits many solutions unless constraints are imposed on $d(x)$. If we use constraints of the Tikhonov type, we can look for a solution $d(x)$ that minimizes

$$\|L(x) - R(x + d(x))\| + \lambda \|d'(x)\|. \quad (11.4)$$

The second term in (11.4) is the disparity gradient, which is thus introduced as a direct consequence of regularization methods.

One important property of the disparity is that $d(x)$ can be discontinuous. Furthermore, there are often occlusions, that is regions in one image that do not correspond to any part in the other image. In this case, $d(x)$ is not defined.

Because of the presence of occlusions and discontinuities in the disparity, (11.4) does not provide a physically plausible solution. Equation (11.4) requires $d(x)$ to be continuous and differentiable. Equation (11.4) is, however, valid if the disparity gradient is strictly less than 2 (Julesz' definition): in this case there are no occlusions and (11.4) provides a physically plausible solution.

Another problem with (11.4) is that in many instances matching is not performed between the intensity profiles in the two images, but between features maps. In this case, $L(x)$ and $R(x)$ are not continuous functions of x.

XII. Discussion

We believe that algorithms in early vision can be described as solutions to problems of inverse optics. These inverse problems are usually ill-posed or ill-conditioned, but their "degree of ill-posedness" is different in each different instance. Classical problems in inverse optics, such as super-resolution, bandwidth extrapolation, and limited angle tomography can be seriously ill-posed. In many instances ill-posed problems in early vision can become

mildly ill-posed if appropriate devices and techniques are used. This is the case of edge detection. Since in the case of a mildly ill-posed problem, it is important to reduce the amount of noise present in the imaging process, if a low noise camera is chosen, possibly with a cooled sensor, and if a high quality A/D system is used, one may obtain fairly good solutions to the problem of edge detection.

When many views of the same scene are available the problem of shape from shading becomes a well-posed problem and possibly even over-determined. Similarly the best way to solve ill-posed problems in early vision is to obtain additional information or data, or to act on the experimental set-up (active vision) and reduce the instrumental noise. The techniques described in this paper are purely mathematical techniques, which have to be used after a careful evaluation of the physical nature of the problem.

A. Physical Plausibility of the Solution

When the origin of ill-conditioning is the lack of continuous dependence of the solution on the data, regularization techniques, such as those used for edge detection or surface reconstruction, are likely to guarantee an optimal use of available data. On the contrary when a C-generalized solution is used because the solution is not unique or because some relevant information seems missing, the physical plausibility of the solution must be proved.

For instance let us consider the computation of the optical flow, where a C-generalized solution minimizing (8.13) gives the correct solution in the case of a rigid polygon: there is no reason why a similar solution gives a correct or approximately correct solution in a more general case. We have seen that a more appropriate analysis of the computation of the optical flow, using (8.40), reveals that the problem is ill-conditioned only when $|\det \text{Hess } E(x, y, t)|$ is very small.

Physical plausibility of the solution is the most important criterion to select a good solution. The decision regarding the choice of the appropriate stabilizing functional cannot be made judiciously from purely mathematical considerations. A physical analysis of the problem and of its generic constraints play the main role. Standard regularization theory provides a framework within which one has to seek constraints that are rooted in the physics of the visual world, but offers a restricted universe of possible constraints since only certain a priori assumptions can be translated into the language of Tikhonov stabilizers.

B. Well-Posedness and Structural Stability

Robustness against noise implied by well-posedness (or, more precisely, by well-conditioning) means continuity of the solution on the input data. This notion or definition of robustness against noise is not necessarily the only one or even the most useful in early vision. It may be useful to compute qualitative features of images or of processed images and to ask which of these features are unaltered when the original image is slightly perturbed or degraded. If we consider the optical flow or the 2-D motion field, it is of some relevance to look for features of this vector field that are invariant under small perturbations [79]. This problem leads naturally to the analysis of structural stable properties of the vector field, that is qualitative or topological features that are robust against noise [57], [58]. The difference between well-posedness and structural stability is that the former notion is essentially metric (the definition uses norms) while the latter is qualitative (the definition uses topological techniques). A description of the optical flow in terms of foci, spirals, nodes and limit cycles can be used to understand the qualitative features of the motion as limbs and cusps can be used to understand the shape of objects. In essence we may see as a complementary module of early vision the qualitative analysis of images and the theory of structural stability as the right framework for this analysis.

C. Regularization and Learning

The problem of learning a mapping between an input and an output space is essentially equivalent to the problem of synthesizing an associative memory that retrieves the appropriate output when presented with the input and *generalizes* when presented with new inputs. It is also equivalent to the problem of estimating the system that transforms inputs into outputs given a set of examples of input-output pairs. A classical framework for this problem is *approximation theory*.

Approximation theory deals with the problem of approximating or interpolating a continuous function $f(X)$ by an approximating function $F(W, X)$ having a fixed number of parameters W (X and W are real vectors $X = x_1, x_2, \cdots, x_n$ and $W = w_1, w_2, \cdots, w_m$). For a choice of a certain F, the problem is then to find the set of parameters W that provides the best possible approximation of f. This is the *learning* step. Needless to say, it is very important to choose an approximating function F that is as compatible as possible with f. There would be little point in trying to learn an approximation if the chosen approximation function $F(W, X)$ could only give a very poor representation of $f(X)$, even with optimal parameter values.

Of course any reconstruction (or approximation) problem of this type is ill-posed in the sense that the information in the data is not sufficient to uniquely reconstruct the mapping in regions where data are not available. In addition, the data are usually noisy. *A priori* assumptions are needed about the mapping. Generalization is not possible if the mapping is completely *random* or *local*. For instance, knowing examples of the mapping represented by a telephone directory (people's names into telephone numbers) does not help estimating the telephone number corresponding to a new name. Generalization is based on the fact that the world in which we live is usually—at the appropriate level of description—redundant. In particular, it may be *smooth:* small changes in some input parameters determine a correspondingly small change in the output (it may be necessary in some cases to accept *piecewise smoothness*). This is the most general constraint that makes approximation possible, and thus this very simple form of generalization. It establishes an interesting connection between learning on one hand and regularization, splines and Bayesian approaches on the other hand [86].

D. Stochastic Route to Regularization

When *a priori* knowledge of statistical properties of the signal and of the noise is available, a probabilistic version of regularization methods is possible [22], [23], [81], [83]. Several authors have stressed the stochastic interpretation of spline approximation in which the smoothness properties of splines correspond to suitable prior probabilities.

Bertero, Poggio and Torre [84] have discussed a Bayesian approach which has the advantage of showing the connection between Markov Random Field models and standard regularization as developed in this paper. In particular, they show how standard regularization can be regarded as a special case of MRF models and is itself equivalent to Wiener filtering.

These techniques, though computationally expensive, represent a powerful extension of the methods described in this paper [87], [88]. Furthermore, approximate efficient algorithms may be devised for each specific problem [83], [87].

E. Future

This paper has attempted to review the recent development of a regularization framework for computational vision. The review is not exhaustive, and we only mentioned in a cursory way several important papers that are related to regularization. Since the image understanding field is undergoing rapid development, we expect that many more useful connections between vision problems and regularization methods will soon be discovered and exploited in algorithms. A natural area for future work is to apply formal regularization techniques to other problems of early vision such as the computation of surface color, shape-from-texture and spatio-temporal approximation [1]. A more fundamental problem that arises in almost every vision problem is the problem of *scale*, that is, the resolution at which to operate. Methods that have been proposed to deal with the problem include scale-space techniques that consider the behavior of the result across a continuum of scales. From the point of view of regulation theory, the concept of scale is related quite directly to the parameter λ [89]. It is tempting to conjecture that methods used to obtain the optimal value of λ may provide, either directly or after suitable modification, the optimal scale associated with the specific instance of certain problems.

An outstanding problem at present in the area of early vision is the detection and localization of discontinuities. Because of the equivalence between regularization and generalized splines, it is impossible to deal directly with discontinuities in the framework of the classical theory. Different methods, such as Markov Random Fields, seem capable of performing approximation and reconstruction while preserving and detecting discontinuities [80]–[83], [85]. There are promising approaches to the problem of integrating different visual modules such as stereo, motion, color, and texture that rely on coupled Markov Random Field models and their capability to detect and represent discontinuities. Though they use Monte Carlo methods, they indicate that deterministic algorithms (in some cases, of the relaxation type) may provide very good approximative solutions. A significant challenge for regularization theory in computational vision is thus to extend the classical formalism to deal with discontinuities. Lee and Pavlidis' work [11], [16] is an example of this for the 1-D case. The two-dimensional case is significantly more difficult. The approaches of [7]–[10], [85] to surface reconstruction and to edge detection respectively, though not explicitly framed in the context of classical regularization, represent some promising initial steps in the direction of extending the 2-D theory.

Acknowledgment

Dr. E. De Micheli provided us with Figs. 1 and 2, and Dr. B. Caprile, Dr. F. Girosi, and Dr. A. Verri provided us with Figs. 3 and 4.

References

[1] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, pp. 314–319, Sept. 1985.

[2] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 4, pp. 413–424, July 1986.

[3] J. Hadamard, "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton University Bulletin*, vol. 13, 1902.

[4] ——, *Lectures on the Cauchy Problem in Linear Partial Differential Equations*. New Haven, CT: Yale University Press, 1923.

[5] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston & Sons, 1977.

[6] C. W. Groetsch, "The theory of Tikhonov regularization for Fredholm equations of the first kind," *Research Notes in Mathematics*, vol. 105. Boston, MA: Pitman, 1984.

[7] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, MA: MIT Press, 1987.

[8] A. Blake, A. Zisserman, and A. V. Papoulias, "Weak continuity constraints generate uniform scale-space descriptions of plane curves," in *Proc. ECAI*, 1986.

[9] A. Blake and A. Zisserman, "Some properties of weak continuity constraints and the GNC algorithm," in *Proc. CVPR*, pp. 656–661, 1986.

[10] ——, "Invariant surface reconstruction using weak continuity constraints," in *Proc. CVPR*, pp. 62–67, 1986.

[11] D. Lee and T. Pavlidis, "One dimensional regularization with discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, in press.

[12] P. Amandan, "A unified perspective on computational techniques for the measurement of visual motion," in *Proc. Intl. Conf. on Computer Vision*, pp. 719–732, 1987.

[13] P. Amandan and R. Weiss, "Introducing a smoothness constraint in a matching approach for the computation of displacement fields," in *DARPA IV Workshop Proc.*, pp. 186–195, 1985.

[14] T. E. Boult and J. R. Kender, "Visual surface reconstruction using sparse depth data," in *Proc. CVPR*, pp. 68–76, 1986.

[15] T. E. Boult, "Using optimal algorithms to test model assumptions in computer vision," in *Proc. Intl. Conf. on Computer Vision*, pp. 921–929, 1987.

[16] D. Lee, "Some computational aspects of low-level computer vision," this issue.

[17] T. Boult, "What is regular in regularization," in *Proc. Intl. Conf. Computer Vision*, pp. 457–462, 1987.

[18] M. Lavrentiev, *Some Improperly Posed Problems of Mathematical Physics*. Berlin, West Germany: Springer-Verlag, 1967.

[19] L. E. Payne, "Improperly posed problems in partial differential equations," presented at SIAM Regional Conf. Series in Applied Math, 1975.

[20] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*. Berlin, West Germany: Springer-Verlag, 1984.

[21] M. Bertero, "Regularization methods for linear inverse problems," in *Inverse Problems*, C. G. Talenti, Ed. Berlin, West Germany: Springer-Verlag, 1986.

[22] V. F. Turchin, V. P. Kozlov, and M. S. Malkevich, "The use of mathematical-statistics methods in the solution of incorrectly posed problems," *Sov. Phys. Usp.*, vol. 13, pp. 681–840, 1971.

[23] M. Bertero, C. DeMol, and G. A. Viano, "The stability of inverse problems," in *Inverse Scattering Problems in Optics*, H. P. Baltes, Ed. Berlin, West Germany: Springer-Verlag, 1980.

[24] F. John, "Continuous dependence on data for solutions of partial differential equations with a prescribed bound," *Commun. Pure Appl. Math.*, vol. 13, pp. 551–585, 1960.

[25] R. Courant and D. Hilbert, *Methods of Mathematical Physics Vol. II*. London, England: Interscience, 1962.

[26] M. Z. Nashed, Ed., *Generalized Inverses and Applications*. New York, NY: Academic Press, 1976.

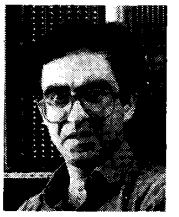
[27] C. W. Groetsch, *Generalized Inverses of Linear Operators*. New York, NY: Dekker, 1977.
[28] M. Bertero, C. DeMol, and E. R. Pike, "Linear inverse problems with discrete data: I-general formulations and singular system analysis," in *Inverse Problems*, vol. 1, pp. 301–330, 1985.
[29] T. N. E. Greville, Ed., *Theory and Application of Spline Functions*. New York, NY: Academic Press, 1969.
[30] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Math. Dokl.*, vol. 4, pp. 1035–1038, 1963.
[31] V. K. Ivanov, "On linear problems which are not well-posed," *Soviet Math. Dokl.*, vol. 3, pp. 981–983, 1962.
[32] ——, "The approximate solution of operator equations of the first kind," *USSR Comp. Math. Math. Phys.*, vol. 6, pp. 197–205, 1966.
[33] V. A. Morozov, "On the solution of functional equations by the method of regularization," *Soviet Math. Dokl.*, vol. 7, pp. 414–417, 1966.
[34] K. Miller, "Least squares methods for ill-posed problems with a prescribed bound," *SIAM J. Math. Anal.*, vol. 1, pp. 52–74, 1970.
[35] J. N. Franklin, "On Tikhonov's method for ill-posed problems," *Math. Comp.*, vol. 28, pp. 889–907, 1974.
[36] C. H. Reinsch, "Smoothing by spline functions," *Numer. Math.*, vol. 10, pp. 177–183, 1967.
[37] G. Wahba, "Practical approximate solutions to linear operator equations when the data are noisy," *SIAM J. Numer. Anal.*, vol. 14, 1977.
[38] ——, "Ill-posed problems: Numerical and statistical methods for mildly, moderately and severely ill-posed problems with noisy data," Tech. Rep. 595, Univ. of Wisconsin, Madison, WI, 1980.
[39] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, pp. 377–403, 1979.
[40] L. V. Kantorovich and G. P. Akilov, *Functional Analysis in Normed Spaces*. New York, NY: Pergamon Press, 1964.
[41] A. Roger, "Newton-Kantorovich algorithm applied to an electro-magnetic inverse problem," *IEEE Trans. Antennas Propagat.*, vol. AP-20, pp. 232–238, 1981.
[42] A. N. Tikhonov, "Solution of nonlinear integral equations of the first kind," *Soviet Math. Dokl.*, vol. 5, pp. 835–838, 1964.
[43] M. Z. Nashed, "Generalized inverses, normal solvability, and iteration for singular operator equations," in *Nonlinear Functional Analysis and Applications*, L. B. Rall, Ed. New York, NY: Academic Press, 1971.
[44] E. C. Hildreth, *The Measurement of Visual Motion*. Cambridge, MA: MIT Press, 1984.
[45] ——, "Computation of the velocity field," *Proc. R. Soc. Lond. B.*, vol. 221, pp. 189–220, 1984.
[46] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.
[47] A. Herskovitz and T. O. Binford, "On boundary detection," Artificial Intelligence Laboratory Memo 183, Massachusetts Institute of Technology, Cambridge, MA, 1980.
[48] D. Marr and E. C. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B*, vol. 207, pp. 187–217, 1980.
[49] J. F. Canny, "Finding edges and lines in images," Artificial Intelligence Laboratory Tech. Rep. 720, Massachusetts Institute of Technology, Cambridge, MA, 1983.
[50] K. F. Shanmugan, F. M. Dickey, and J. A. Green, "An optimal frequency domain filter for edge detection in digital pictures," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 44, pp. 99–149, 1965.
[51] V. Torre and T. Poggio, "On edge detection," Artificial Intelligence Laboratory Memo 768, Cambridge, MA, Massachusetts Institute of Technology, 1984. Also printed in *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 147–163, 1986.
[52] T. Poggio, H. Voorhees, and A. Yuille, "Regularizing edge detection," Artificial Intelligence Laboratory Memo 776, Massachusetts Institute of Technology, Cambridge, MA, 1984.
[53] B. R. Frieden, "Linear and circular prolate functions," in *Progress in Optics IX*, E. Wolf, Ed. Amsterdam, The Netherlands: Elsevier North Holland, 1971, pp. 312–408.
[54] H. J. Landau and H. O. Pollack, "Prolate spherical wave functions, Fourier analysis and uncertainty-II," *Bell Syst. Tech. J.*, vol. 40, pp. 65–84, 1961.
[55] H. H. Hermuth, *Transmission of Information by Orthogonal Functions*. Berlin, West Germany: Springer-Verlag, 1972.
[56] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York, NY: McGraw-Hill, 1965.
[57] R. Thom, "Quelques proprietes globales des varietés differentiables," *Comm. Math. Helv.*, vol. 28, pp. 17–86, 1954.
[58] T. Stuart and I. Poston, *Catastroph Theory and its Applications*. Boston, MA: Pitman, 1978.
[59] A. Verri and T. Poggio, "Motion field and optical flow: Qualitative properties," AIL Memo 917, Massachusetts Institute of Technology, Cambridge, MA, 1986.
[60] D. Marr, *Vision*. Boston, MA: Freeman, 1982.
[61] A. Yuille, "The smoothest velocity field and token matching schemes," Artificial Intelligence Laboratory Memo 724, Massachusetts Institute of Technology, Cambridge, MA, 1983.
[62] W. J. Kammerer and M. Z. Nashed, "On the convergence of the conjugate gradient method for singular linear operator equations," *SIAM J. Numer. Anal.*, vol. 9, pp. 165–181, 1972.
[63] M. Bertero, P. Brianzi, M. Defrise, and C. DeMol, "Iterative inversion of experimental data in weighted spaces," in *Proc. URSI International Symp. on Electromagnetic Theory*, 1986.
[64] S. Uras, F. Girosi, A. Verri, and V. Torre, "A computational approach to motion perception," *Biological Cybernetics*, 1988, in press.
[65] D. Marr and T. Poggio, "A theory of human stereo vision," *Proc. Roy. Soc. Lond. B*, vol. 204, pp. 301–328, 1979. Also published as Artificial Intelligence Laboratory Memo 451, Massachusetts Institute of Technology, Cambridge, MA, 1977.
[66] W. E. L. Grimson, *From Images to Surfaces: A Study of the Human Early Visual System*. Cambridge, MA: MIT Press, 1981.
[67] ——, "A visual theory of visual surface interpolation," *Phil. Trans. R. Soc. Lond. B*, vol. 298, pp. 395–427, 1982.
[68] D. Terzopoulos, "Multilevel visual processes for visual surface reconstruction," *Computer Vision, Graphics, and Image Processing*, vol. 24, pp. 52–96, 1983.
[69] ——, "Multiresolution computation of visible-surface representations," Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1984.
[70] ——, "Multilevel reconstruction of visual surfaces: Variational principles and finite element representations," Artificial Intelligence Laboratory Memo 671, Massachusetts Institute of Technology, Cambridge, MA, 1984. Also in *Multiresolution Image Processing and Analysis*, A. Rosenfeld, Ed. Berlin, West Germany: Springer-Verlag, 1984, pp. 237–310.
[71] J. Duchon, "Interpolation des functions de deux variables suivant le principe de la flexion des plaques minces," *R.A.I.R.O. Analyse Numerique*, vol. 10, pp. 5–12, 1976.
[72] G. Wahba and J. Wendelberger, "Some new mathematical methods for variational objective analysis using splines and cross validation," *Monthly Weather Review*, vol. 108, pp. 1122–1143, 1980.
[73] J. H. Ahlberg, E. H. Nilson, and J. L. Walsh, "The theory of splines and their applications," in *Mathematics in Science and Engineering*, vol. 38. New York, NY: Academic Press, 1967.
[74] B. K. P. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," Project MAC Internal Rep. TR-79 and Artificial Intelligence Laboratory Tech. Rep. 232, Massachusetts Institute of Technology, Cambridge, MA, 1970.
[75] ——, "Obtaining shape from shading information," in *The Psychology of Computer Vision*, P. H. Winston, Ed. New York, NY: McGraw-Hill, 1975.
[76] B. K. P. Horn and R. W. Sjoberg, "Calculating the reflectance map," *Appl. Opt.*, vol. 18, pp. 1770–1779, 1979.
[77] B. K. P. Horn, "Hill-shading and the reflectance map," *Proc. IEEE*, vol. 69, pp. 14–47, 1981.
[78] K. Ikeuchi, "Determining surface orientations of specular surfaces by using the photometric stereo method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 3, pp. 661–669, 1981.
[79] A. Verri, F. Girosi, and V. Torre, "The mathematical properties of the 2-D motion field: from singular points to motion parameters," *J. Opt. Soc. Amer. A*, submitted.
[80] A. V. Balakrishnan, *Applied Functional Analysis*. Berlin, West Germany: Springer-Verlag, 1976.
[81] S. Geman and D. Geman, "Stochastic relaxation, Gibbs dis-

tributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, pp. 721–741, 1984.
[82] J. Marroquin, "Surface reconstruction preserving discontinuities," Artificial Intelligence Laboratory Memo 792, Massachusetts Institute of Technology, Cambridge, MA, 1984.
[83] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.*, vol. 82, pp. 76–89, 1987. Also in *Proc. Image Understanding Workshop*, L. Baumann, Ed. McLean, VA: SAI Corp., 1985.
[84] M. Bertero, T. Poggio, and V. Torre, "Ill-posed problems in early vision," Artificial Intelligence Laboratory Memo 924, Massachusetts Institute of Technology, Cambridge, MA, 1987.
[85] T. Poggio, "Integrating vision modules with coupled MRFs," Artificial Intelligence Laboratory Memo 285, Massachusetts Institute of Technology, Cambridge, MA, 1985. See also T. Poggio and staff, "MIT progress in understanding images," in *Proc. Image Understanding Workshop*, L. Bauman, Ed. San Mateo, CA: Morgan Kaufmann Publishers, 1987.
[86] T. Poggio and staff, "MIT progress in understanding images," in *Proc. Image Understanding Workshop*, L. Bauman, Ed. San Mateo, CA: Morgan Kaufmann Publishers, 1988.
[87] E. B. Gamble and T. Poggio, "Visual integration and detection of discontinuities: The key role of intensity edges," Artificial Intelligence Laboratory Memo 970 and Center for Biological Information Processing Paper 027, Massachusetts Institute of Technology, Cambridge, MA, 1987.
[88] T. Poggio *et al.*, "The vision machine," in *Proc. Image Understanding Workshop*, L. Bauman, Ed. San Mateo, CA: Morgan Kaufmann Publishers, 1988.
[89] D. Geiger and T. Poggio, "An optimal scale for edge detection," in *Proc. IJCAI*, vol. 2, pp. 745–748, 1987.

Mario Bertero was born in 1938. He received the Ph.D. degree in physics from the University of Genova in 1960, and the "libera docenza" in theoretical physics in 1968.

He held research and teaching appointments at the Universities of Genova, Bonn, and Brussels and he is now Professor of Mathematics at the University of Genova. His main research activities have been in nonrelativistic quantum scattering theory, Regge poles theory, nonlocal potentials, mathematical foundations of the optical model for nuclear reactions, deterministic and probabilistic regularization methods for inverse and ill-posed problems, Laplace transform inversion, numerical methods for the solution of inverse problems with discrete data. His current research interest is in the applications of the theory of ill-posed problems to data reduction in confocal scanning light microscopy, in confocal scanning fluorescence microscopy, in particle sizing in the micron and sub-micron region by light scattering, and in microwave thermography.

Dr. Bertero is a member of the Scientific Committee of I.A.G.A. (Institute of Global Analysis and Applications–C.N.R.–Florence) and a member of the Editorial Board of the journal *Inverse Problems*.



Tomaso A. Poggio (Associate, IEEE) was born in Genoa, Italy, on September 11, 1947. He received the Ph.D. degree in theoretical physics from the University of Genoa in 1970.

From 1971 to 1982, he was Wissenschaftlicher Assistant at the Max-Planck-Institut fur Biologische Kybernetik, Tübingen, West Germany. Since 1982, he has been a Professor at the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. In 1984, he was also appointed Professor at MIT's Whitaker College of Health Sciences and Technology, and was named the first Director of its Center for Biological Information Processing. In 1988 he has been named the Uncas and Helen Whitaker Professor. He has authored over 100 papers in areas ranging from psychophysics to biophysics, information processing in man and machine, artificial intelligence, and machine vision. He is on the editorial board of several interdisciplinary journals.

Vincent Torre was born in Johannesburg, South Africa, on July 24, 1950. He received the Ph.D. degree in theoretical physics from the University of Genoa, Italy, in 1973.

From 1974 to 1978 he worked on the electrophysiology of retinal cells at the Laboratory of Neurophysiology CNR, Pisa, Italy. From 1979 to 1981 he worked on the mechanisms of phototransduction at the Physiological Laboratory, Cambridge U.K., under the supervision of Sir Alan Hodgkin. Since 1983 he has been Associate Professor at the Department of Physics in Genoa, where he teaches system theory and leads the group on machine vision. He is the author of over 70 papers in biophysics and information processing in man and machine.