MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo 924 May 1987

ILL-POSED PROBLEMS IN EARLY VISION

M. Bertero[1], T. Poggio[2], and V. Torre[1]

Abstract: The first processing stage in computational vision, also called *early vision*, consists in decoding 2D images in terms of properties of 3D 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 non-linear case. More general stochastic regularization methods are also introduced. Specific topics in early vision and their regularization are then analyzed rigorously, characterizing existence, uniqueness and stability of solutions.

Acknowledgements. This report describes research done 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 Products Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. Some support for Tomaso Poggio is provided by a gift from the Artificial Intelligence Division of the Hughes Aircraft Corporation. NATO provided some support for Vincent Torre and Mario Bertero.

[1] Dipartimento di Fisica, Università di Genova, I-16146 Genova, Italy. [2] Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Introduction

Vision systems, either 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 matrices of light intensity values measured by an eye or a camera. For tasks such as navigation, manipulation and visual recognition, vision systems have to recover 3D properties of surfaces from the 2D images. Typical 3D properties are the distance between the surfaces and the viewer, their orientation, structure, 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 behavioural 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, computation of optical flow.* The computational modules that solve them constitute together the core of *early vision*, and provide spatial and geometrical information about the 3D 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 and also reasoning about objects. Unlike high level vision, early vision is mostly considered as a *bottom-up* set of processes that do not rely upon specific high-level information about the scene to be analysed. It is commonly argued on the basis of computational and psychophysical considerations that these different modules of early vision can be analysed independently of each other, at least 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 (another possibility is to use coupled Markov Random Field models for the integration stage, see [68]).

Even a superficial analysis of these problems reveals their common inverse nature: they can be regarded as *inverse optics* since they attempt to recover 3D properties of surfaces from the 2D images they generate. This observation characterizes the field of *early vision* as the solution of problems of inverse optics[1].

It is well known that inverse problems are very often ill-posed[2],[3] in the original sense of Hadamard[4]; that is, the solution may not exist or it is not unique or does not depend continuously on data. These problems can be formulated as discrete or continuous problems according to the type of

the available data. In order to *solve* ill-posed problems a priori information about generic properties of the solution must be used. Regularization theory is a set of techniques that have been developed for this reason. Standard regularization exploits a priori knowledge by restricting the functional space to which the "solution" belongs: the specific techniques use generalized inverses or variational formulations. An alternative possibility, that we call stochastic regularization, is based on a Bayesian approach and optimal estimation. All these reasons we introduced recently regularization techniques into computer vision. B. Horn (see [69] for a comprehensive review of his work) had approached several problems in vision from a similar point of view, using minimization techniques for their solution, without an explicit connection with regularization techniques.

The goal of this paper is to review the mathematical aspects of this approach (for a less rigorous discussion, see [1]). It is organized in two parts: Part One reviews the main mathematical results that characterize the difference between well-posed and ill-posed problems (Section 2) and between well-conditioned and ill-conditioned problems (Section 4). The notions of generalized inverses (Section 3) and of regularization methods (Section 5) are then introduced. Section 6 contains some results related to inverse nonlinear problems. Section 7 illustrates stochastic regularization.

Part Two shows that several variational principles recently introduced in early vision can be formulated as regularized solutions to ill-posed inverse problems. Four problems in early vision are studied in detail: edge detection and numerical differentiation (Section 8), optical flow (Section 9), surface interpolation (Section 10), and shape from shading (Section 11).

Part One

1. 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 non-linear problems.

In Section 2 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 4). 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 3). When the range of the operator is not closed, or when the problem is seriously ill-conditioned, regularization techniques have to be used (Section 5) 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 6. A stochastic approach to inverse problems is introduced in Section 7 and its connections with standard regularization are discussed.

2. Well-Posed and Ill-posed Problems

Hadamard[4],[5] defined a mathematical problem to be *well-posed* when:

(a) for each data 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, [6]). Continuity, however, is

a necessary but not sufficient condition for stability. A well-posed problem can be ill-conditioned (see Section 4).

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 in the computation of the characteristics of the targets from the knowledge of the sources and of the scattered radiation.

Consider a very simple example taken from classical optics. If the energy distribution u is given in the object plane of an optical instrument and if the characteristics of the instruments are known, it is possible to compute, by solving the wave equation, the energy distribution g in the image plane. In the case of Fourier optics, one finds a linear relation between u and g:

$$g(x) = \int K(x,y)u(y)dy, \tag{2.1}$$

the kernel $K(x,y)$ being the impulse response (point spread function) of the instrument. The direct problem (the computation of g given u) is clearly well-posed. The inverse problem (the computation of u given g) usually is not.

Assume that $K(x,y) = K(x-y)$, where $K(x)$ is a band-limited function. Then there exist functions u which produce a zero image (think of a function which has only Fourier components out of the band of the instrument) and therefore uniqueness does not hold. Furthermore, if $g(x)$ is the result of an experiment and therefore is affected by noise, it is not necessarily band-limited or it can have a band broader than that of the instrument. Under these circumstances the solution of Equation (2.1) does not exist.

The need to investigate problems which 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* [7],[8], 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[2],[3],[9],[10],[11]. These days, the term ill-posed is used generally (but not only) for those problems which 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.

3. 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. Notice that Equation (2.1) is a special case of Equation (3.1).

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") assure 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

$$(Lu)(x) = \int_{-\infty}^{+\infty} \frac{\sin \Omega(x-y)}{\pi(x-y)} u(y)dy. \tag{3.2}$$

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 the 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 spaces 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* .

3.1. 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 set of functions $u \in X$ that solve the variational problem

$$\|Lu - g\|_Y = \text{minimum}, \tag{3.3}$$

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 (3.3),

$$2Re(Lu - g, Lh)Y, \tag{3.4}$$

where h is an arbitrary function of X and $(\cdot,\cdot)Y$ the inner product of the Hilbert space Y. Setting (3.4) equal to zero, we obtain the Euler equation

$$L^*Lu = L^*g, \tag{3.5}$$

where L^* is the adjoint of the operator L (L^* is a mapping from Y into X). When $R(L)$ is closed, Equation (3.5) always has solutions but the solution is not unique when $N(L)$ is nontrivial. Notice that the set of solutions of Equation (3.5) coincides with the set of solutions of the equation

$$Lu = Pg, \tag{3.6}$$

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

3.2. 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 (3.5) which has minimal norm

$$\|u\|_X = \text{minimum}. \tag{3.7}$$

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 mapping L^+ from Y into X is defined by

$$u^+ = L^+ g. \tag{3.8}$$

The operator L^+ is the *generalized inverse* of L and it is continuous. Therefore, the problem of computing the generalized solution of Equation (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

$$Y = R(L) \oplus R^\perp(L), \tag{3.9}$$

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.

3.3. 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 9. Then an extension of the generalized solution goes as follows: let $p(u)$ be a norm or a seminorm on X (see Appendix A for the definition) of the following style:

$$p(u) = \|Cu\|_Z \tag{3.10}$$

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 (contains 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 which 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" which 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[7]. 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 (see Appendix B):

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

$$Lu = 0, \qquad Cu = 0 \tag{3.11}$$

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

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

(iii) The set of functions u 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. These properties hold true, for instance, in the case of interpolation problems in reproducing kernel Hilbert spaces (see Appendix C).

3.4. Generalized solutions for problems with discrete data

We conclude this section by noticing that problems with discrete data can be formulated as (3.1), g being now a n-dimensional vector in an Euclidean space. In fact, ignoring the errors in the data, a linear inverse problem with discrete data can be formulated as follows[12] :

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

$$g_i = F_i(u); i = 1, \ldots, n. \tag{3.12}$$

In particular, when the functionals F_i are continuous on X by Riesz Theorem (see Appendix A), there exist functions $\psi_1, \psi_2, ..., \psi_n$ such that

$$F_i(u) = (u, \psi_i)_X, \tag{3.13}$$

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

For example, the problem discussed in Section 2 takes this form when $g(x)$ is measured in a finite set of points only, say $x_1, x_2, ..., x_n$, and X is an L^2 space. In such a case we have

$$\psi_i(x) = K(x_i, y). \tag{3.14}$$

It is important to recognize also that interpolation problems take this form when X is a reproducing kernel Hilbert space (see Appendix C).

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 an N-dimensional Euclidean space Y and if we define an operator L from X into Y by means of the relation

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

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 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 Hilbert space (see, for instance, [12], [13]). A simple example is discussed in Appendix C.

4. 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 3, 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 Equation 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^+ (see Appendix A). Analogously, from Equation 3.1, with $u = u^+$, it follows that

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

Combining Equations 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 l.h.s. of Equation 4.3 can be arbitrarily close to the r.h.s.

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 semi-definite 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.,

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

More generally, if we indicate by $\sigma_+(L^*L)$ the positive part of the spectrum of the operator L^*L, we have

$$\alpha = \sqrt{\frac{\max \sigma_+(L^*L)}{\min \sigma_+(L^*L)}}. \tag{4.6}$$

In order to provide an example of a well-posed problem which can be extremely ill-conditioned, we consider the finite *moment problem*, i.e., the problem of determining a function $u(x)$, defined for example on $[0,1]$, given its moment up to the order $N-1$:

$$g_n = \int_0^1 x^{n-1} u(x) dx; \quad n = 1, ..., N. \tag{4.7}$$

If we take $X = L^2(0,1)$, then it is easy to recognize[12] that the operator LL^* is just the Hilbert matrix

$$[H_N(-1)]_{nm} = \frac{1}{n+m-1}; \quad n, m = 1, ..., N, \tag{4.8}$$

which is a classical example of an ill-conditioned matrix. From well-known results it follows that the condition number for the generalized solution of problem 4.7 is approximately given by $\alpha \cong exp(1.75N)$ and therefore it grows exponentially with N. Already for moderate values of N, α takes unacceptable values.

5. 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) as shown in Appendix A. 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 technique.

5.1. Tikhonov regularization

The most investigated approach to ill-posed problems is the *regularization method* of Tikhonov[14]. 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 Equation (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 $g_\varepsilon = g + n_\varepsilon$ and n_ε represents experimental errors or noise, we have

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

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 algorithms is obtained by minimizing the functional

$$\|Lu - g\|^2_{\ Y} + \lambda \|Cu\|^2_{\ Z} = \text{minimum}, \tag{5.3}$$

where C is a constraint operator, satisfying for instance the conditions stated in Section 3. In the original paper of Tikhonov, it is given by

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

where the weights $c_r(x)$ are strictly positive functions and $u^{(r)}(x)$ indicates the r^{th}-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 the Equations (3.11) have only the trivial solution $u = 0$ and that when $\lambda \to 0$, $g \in R(L)$, $R_\lambda g$ converges to ${L_C}^+ g$ [3].

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

(i) Among all u such that $\|Cu\|_Z \leq E$ find u that minimizes $\|Lu - g_\varepsilon\|_Y$ [15]. Using the method of Lagrange multipliers the solution of this problem can be reduced to the solution of Equation (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 $\|Lu - g_\varepsilon\|_Y \leq \varepsilon$, with given ε, find u that minimizes $\|Cu\|_Z$ [16],[17]. Again, the solution of the problem is equivalent to finding the unique λ such that

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

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

(iii) Among all u such that $\|Lu - g_\varepsilon\|_Y \leq \varepsilon$, $\|Cu\|_Z \leq E$, find a u of the type (5.5). This is equivalent[18],[19] to taking

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

The first method consists of finding the function u that satisfies the constraint $\|Cu\|_Z \leq E$ and best approximates the data. The second method conputes 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.

5.2. Regularization and filtering

The regularized solution (5.5), (5.6) takes a very simple form in the case where L is compact and $C = I$ the identity operator in X. Then, using the singular value decomposition of L (see Appendix A) we obtain

$$u_\lambda = \sum_k \frac{\alpha_k^2}{\alpha_k^2 + \lambda} \frac{1}{\alpha_k} (g, v_k)_Y u_k, \tag{5.10}$$

and therefore the regularized solution is essentially a "filtered" version of the non-regularized (generalized) solution of Equation (3.1),

$$u^+ = \sum_k \frac{1}{\alpha_k} (g, v_k)_Y u_k. \tag{5.11}$$

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

$$u_\lambda = \sum_k \Phi_k(\lambda) \frac{1}{\alpha_k} (g, v_k)_Y u_k \tag{5.12}$$

satisfying the conditions: (i')$\Phi_k(\lambda) \leq 1$; (ii')$\Phi_k(\lambda) \to 1$, for any k, when $\lambda \to 0$; (iii')$\Phi_k \frac{(\lambda)}{\alpha_k}$ bounded for any k and any $\lambda > 0$. Such a procedure is often used in the inversion of compact operators as well as in the inversion of convolution operators[2](see Section 8).

5.3. Smoothing and interpolation

As already remarked, regularization algorithms can be used also for ill--ditioned problems. A well-known example is the smoothing of a function whose values, specified on a finite set of points, are affected by errors[20]. 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

$$u(x_i) = g_i \quad ; \quad i = 1, \ldots, n \tag{5.13}$$

and minimizes the seminorm[13]

$$p(u) = \int_0^1 |u''(x)|^2 dx. \tag{5.14}$$

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 mininization of the seminorm (5.14), but condition (5.13) is replaced by

$$\sum_{i=1}^n |u(x_i) - g_i|^2 \leq \varepsilon^2 \tag{5.15}$$

(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.

5.4. Cross validation and generalized cross validation

We conclude this section with a short description of the *cross validation method*[21],[22]. 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 3.

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

$$F_\lambda^{[k]}[u] = \frac{1}{n}\frac{\sum_{i=1}^{n}}{i \neq k} |(Lu)_i - g_i|^2 + \lambda \|u\|_X^2. \tag{5.16}$$

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

$$V_o(\lambda) = \frac{1}{n}\sum_{k=1}^{n} |(Lu_\lambda^{[k]})_k - g_k|^2 \tag{5.17}$$

and the cross validation method consists in determining the value of λ, say $\overline{\lambda}$, which minimizes (5.17). The computation of the minimum is based on the relation

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

where u_λ is the minimizer of the functional

$$F_\lambda[u] = \frac{1}{n}\sum_{k=1}^{n} |(Lu)_i - g_i|^2 + \lambda \|u\|_X^2, \tag{5.19}$$

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

$$A(\lambda) = LL^*(LL^* + \lambda I)^{-1}, \tag{5.20}$$

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

It has been shown[23] 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

$$V(\lambda) = (\frac{1}{n}Tr[I - A(\lambda)])^{-2}(\frac{1}{n}\|(I - A(\lambda))\vec{g}\|^2), \tag{5.21}$$

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.

6. 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 of this equation exists if and only if g is in the range of the operator A.

6.1. Linearization

The simplest way of treating Equation (6.1) is to try to linearize the problem. This is the case of a *differentiable operator*[24]. 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 $u \in X$,

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

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 Equation (6.1) is known and if the operator A is differentiable at u_o, then Equation (6.1) can be approximated by the linear equation

$$\partial g_o = L_o \partial u_o, \tag{6.4}$$

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

The procedure can be iterated. By means of the solution ∂u_o of Equation (6.4), one gets a new approximation, $u_1 = u_o + \partial u_o$, of the true solution u. Then one considers the linear equation $\partial g_1 = L_1 \partial u_1$, where $L_1 = A'(u_1)$, $\partial g_1 = g - A(u_1)$, and $\partial u_1 = u - u_1$. By solving this equation one gets a new approximation $u_2 = u_1 + \partial u_1$ and so on. It is easily recognized, by writing Equation (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[24]. They include the continuity of the inverse of the derivative of the operator A. In several inverse problems this condition is not satisfied. It has been suggested[25] 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.

6.2. 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[26] and an abstract approach is given by Morozov[27].

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 Equation (6.1) by solving the variational problem

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

analogous to the problem (3.3). When a solution of such a problem exists for any $g \in Y$, one says that Equation (6.1) is strictly normally solvable. A sufficient condition for strict normal solvability is that the range of A is weakly closed in Y [28]. Notice that this condition may be stronger than the condition of closure of the range which applies to the case of linear operators (Section 3). 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 3.1; i.e., by minimizing a norm or seminorm such as (3.10). 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 be easily 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 [26] for a special case where uniqueness holds true). However, it is not difficult to prove the existence of at least one local minimum. Here we give the proof under conditions which are satisfied in the case of the problem of shape from shading (Section 11).

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 goes as follows:

Let M_λ be the lower bound of $\Phi_\lambda[u]$ and let $\{u_n\}$ be a minimizing sequence such that

$$M_\lambda \leq \Phi_\lambda[u_n] \leq M_\lambda + 1/n. \tag{6.9}$$

It follows that:

$$\|Cu_n\|_Z \leq \left(\frac{1}{\lambda}\Phi_\lambda[u_n]\right)^{1/2} < \left(\frac{M_\lambda + 1}{\lambda}\right)^{1/2}; \tag{6.10}$$

and therefore, the sequence $\{v_n = Cu_n\}$ is bounded. Since C^{-1} is compact, we can extract from $\{u_n = C^{-1}v_n\}$ a subsequence strongly convergent and such that the corresponding subsequence of the v_n is weakly convergent. Without loss of generality, we can assume that these conditions are satisfied by the sequence $\{u_n\}$ itself. Then, let u_λ be the strong limit of $\{u_n\}$ and v_λ be the weak limit of v_n; it follows that $u_\lambda = C^{-1}v_\lambda$.

Since Cu_n weakly converges to Cu_λ, from the lower weak semicontinuity of the norm we have

$$\|Cu_\lambda\|_Z \leq \liminf \|Cu_n\|_Z. \tag{6.11}$$

On the other hand, from the strong convergence of u_n to u_λ and from the continuity of $A(u)$ we have

$$\|A(u_\lambda) - g\|_Y = \lim \|A(u_n) - g\|_Y, \tag{6.12}$$

and, by combining Equations (6.9), (6.11), and (6.12), we get

$$M_\lambda \leq \Phi[u_\lambda] \leq \liminf \Phi_\lambda[u_n] = \lim \Phi_\lambda[u_n] = M_\lambda. \tag{6.13}$$

It follows that

$$\Phi[u_\lambda] = M_\lambda, \tag{6.14}$$

and the existence of the minimum point is proven.

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, Equation (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 derivative;

(c) the derivative of A at u, $A'(u)$, is invertible;

then, by a rather easy generalization of the theorems contained in [26], one can prove that if g_ε is noisy data, with $\|g - g_\varepsilon\|_Y \leq \varepsilon$, and if in the functional (6.8), with g replaced by g_ε, we choose the regularization parameter λ in such a way that $\lambda = \gamma\varepsilon^2$, where γ is an arbitrary constant, then any minimum point of such a functional converges to u when $\varepsilon \mapsto 0$; therefore, for sufficiently small values of ε, there exists only one minimum point.

7. 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[29],[30],[31].

Here we consider a Bayesian approach that has the advantage of showing the connection between Markov Random Field models and standard regularization as developed in this paper. In particular, we will be able to see in which sense standard regularization is a special case of MRF models and is itself equivalent to Wiener filtering.

The first step is to write Equation (3.1) in this form,

$$g = Lu + w, \tag{7.1}$$

where w is a function representing the effect of the noise on the data. Notice that in this representation no assumption is implicit about the structure of the noise (additive noise, signal dependent noise, etc.).

The second step is to assume that there exist *stochastic processes* $\underline{u}, \underline{g}, \underline{w}$, related by

$$\underline{g} = L\underline{u} + \underline{w}, \tag{7.2}$$

and that the functions g, u, w appearing in Equation (7.1) are values of the processes $\underline{u}$, $\underline{g}$, $\underline{w}$, associated with a specific outcome of a given experiment (we use here the nomenclature introduced in [32]).

For simplicity, we also assume that the processes $\underline{u}$, $\underline{g}$, $\underline{w}$ have zero mean. This assumption is in fact not restrictive because, if it is not true, it is always possible to introduce processes satisfying this condition just by subtracting the means from the original ones. Thanks to the linearity of L, relation (7.2) holds true also for the new processes.

When the mean is zero, the autocorrelation and the autocovariance of a process $\underline{u}$ coincide. If the values of the zero mean process $\underline{u}$ are functions of a variable x (eventually multi-dimensional), the *autocovariance function* of $\underline{u}$ is

$$C_{\underline{u}}(x, x') = E\{\underline{u}(x)\underline{u}(x')\}, \tag{7.3}$$

where E indicates the expected value. As in the previous sections, we assume that the functions u, the values of the process $\underline{u}$, belong to a Hilbert space X (for instance, a space of square integrable functions) and that the functions g, w, values of the processes $\underline{g}$, $\underline{w}$ respectively, belong to the same (possibly different from X) Hilbert space Y. The appropriate description of stochastic

processes with values in Hilbert spaces is given in terms of *weak random variables* or cylinder set measures[33].

Then, the autocovariance function of the process $\underline{u}$ can be considered as the kernel of an operator $R_{\underline{u}}$ defined on the space X:

$$(R_{\underline{u}}\varphi)(x) = \int C_{\underline{u}}(x,x')\phi(x')dx', \phi \in X. \tag{7.4}$$

The operator $R_{\underline{u}}$ is called the *covariance operator* (or the covariance) for the process $\underline{u}$. It can also be defined for weak random variables with values in an abstract Hilbert space X [33] .

Coming back now to our basic equations (7.1), (7.2), the inverse problem consists in estimating a value of $\underline{u}$, given an observed value g of $\underline{g}$ and given *a priori* probabilistic knowledge on the processes $\underline{u}$ and $\underline{w}$.

We take a Bayesian approach and write the *a posteriori* probability density as

$$P(u/g) = const P(u)P(g/u) \tag{7.5}$$

where $P(u)$ is the *a priori* probability density of process $\underline{u}$ and $P(g/u)$ is the conditional probability density of the data $\underline{g}$ given $\underline{u}$.

We consider now the special case of $\underline{u}$ being a gaussian process (or equivalently the linear transformation – such as a derivative – of a gaussian process). In this case, the *a priori* probability distribution of $\underline{u}$ is

$$P(u) = const \cdot \exp[-\frac{1}{2}(u, R_{\underline{u}}^{-1}u)_X]. \tag{7.6}$$

Let us assume that the noise process $\underline{w}$ is additive, white and gaussian with variance σ^2. Then the *a priori* probability $P(g/u)$ can be written as

$$P(g/u) = const \cdot \exp[-\frac{1}{2\sigma^2}\|g - Lu\|_Y^2]. \tag{7.7}$$

Depending on the optimality criterion there are now several ways of obtaining the best estimate of u given the data g. A commonly used estimate is the *Maximum A Posteriori* (MAP) estimate

$$P(u_{\text{best}}/g) = \max\{P(u/g) | u \in X\}. \tag{7.8}$$

From Equations (7.5) – (7.7) we have

$$P(u/g) =$$

$$const \cdot exp[-\frac{1}{2\sigma^2}(\|Lu - g\|_Y^2 + \sigma^2(u, R_{\underline{u}}^{-1}u)_X)]. \tag{7.9}$$

If we put

$$R_{\underline{u}} = (C^*C)^{-1}, \tag{7.10}$$

then from Equations (7.8) and (7.9) we have

$$M(u_{\text{best}}) = \min\{M(u) | u \in X\}, \tag{7.11}$$

where

$$M(u) = \|Lu - g\|_Y^2 + \sigma^2 \|Cu\|_X^2. \tag{7.12}$$

It follows that $u_{\text{best}} = F_0 g$ where F_0 is given by (5.6) with $\lambda = \sigma^2$. If we put $R_{\underline{w}} = \sigma^2 I$, where I is the identity operator in Y ($R_{\underline{w}}$ is the covariance operator of white noise), then F_0 can also be written in the following form:

$$F_o = R_{\underline{u}} L^* (L R_{\underline{u}} L^* + R_{\underline{w}})^{-1}, \tag{7.13}$$

with $R_{\underline{u}}$ given by Equation (7.10).

The operator F_o is sometimes called *Wiener filter* (or Wiener-Kolmogoroff filter) and is quite similar to the operator (5.6)[31]. In other words, the regularizing operator (5.6) is equivalent to a Wiener filter in the case of *white noise*, provided that the constraint operator C is related to the covariance operator $R_{\underline{u}}$ by the relation (7.10).

Most of our previous assumptions can be relaxed in a more general probabilistic scheme based on the formalism of Markov Random Fields (MRF) defined on finite lattices. In particular, the noise may not be additive, the operator L may not be linear and $P(u)$ does not need to be gaussian.

MRF models have been formulated for several problems in early vision. Under simplifying assumptions they reduce to the discrete equivalent of standard regularization. Though they are computationally expensive they may represent a powerful extension of the methods described in this paper[34],[35],[36].

Part Two

The initial stage of machine vision, now called early vision, consists of distinct but interrelated problems like "edge detection," "computation of optical flow," "structure from motion," "structure from stereo matching," etc. From a theoretical point of view, these problems can be considered as independent, at least to a first approximation. The integration of various outputs is performed at a higher stage, where geometrical reasoning and much *a priori* information will be used. These different modules may reflect processing occurring in our brain, where simultaneously we compute different information from images: we can extract rapid changes in image brightness (edge detection); we can recover the shape of an object from its shading (shape from shading); we can understand the motion of objects from the changing images (computation of optical flow); we recover the 3D structure of a scene from a pair of images (structure from stereo); and we are able to have a dense description of 3D surfaces from sparse features (visual surface interpolation).

Several of these problems have been recently solved with variational techniques, in particular by Horn, Grimson and Hildreth. We will show that many of these results and several new ones – in particular existence and uniqueness of solutions – are direct consequences of mathematical results presented in Part One.

Part Two is divided in four sections, each dealing with a main topic of early vision. Section 8 presents the ill-posed nature of numerical differentiation. In Section 9 we show how recently obtained mathematical results on optical flow [37],[38],[39],[45] are straightforward consequences of regularization theory. Section 10 discusses a recent approach to surface interpolation, illustrating how variational principles[39],[40],[41],[42],[43],[44] can be viewed as regularized solutions to discrete ill-posed problems. Section 11 reviews recent variational approaches to shape from shading, in the framework of regularization theory.

8. Edge detection and numerical differentiation

Recently standard regularization techniques have been applied to a classical problem of early vision – edge detection. Edge detection, intended as the process that attempts to detect and localize changes of intensity in the image (this definition does not encompass all the meanings of edge detection) is a problem of numerical differentiation[46]. As we will see in the next section, differentiation is a common operation in early vision and is not restricted to edge detection. Differentiation is ill-posed because the solution does not depend continuously 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 (in L_2 norm) "noise" term $\epsilon \sin \Omega x$. $f(x)$ 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. Differentiation can also be seen as the recovery of the solution u to the inverse problem $g = Lu$ where

$$L : X \rightarrow Y \qquad (Lu)(x) = \int_{-\infty}^{x} u(y)dy. \tag{8.1}$$

Thus u is the derivative of the data g. The operator (8.1) is not bounded in $L^2(-\infty, \infty)$, and the range of L is not closed. Therefore, the inverse problem is ill-posed .

8.1. Regularization of differentiation

As shown in Section 3, it is possible to restore well-posedness by redefining the solution space X and the data space Y. Let us redefine the solution space X as the subset X' of square integrable functions $f(x)$ in $(-\infty, +\infty)$ such that

$$\int_{-\infty}^{+\infty} (1 + \frac{1}{\omega^2}) \, |F(\omega)|^2 \, d\omega \tag{8.2}$$

exists, where $F(\omega)$ is the Fourier transform of $f(x)$. The new data space Y' is simply the range of L. It is easy now to see that the inverse problem $L : X' \rightarrow Y'$ when the operator L is defined as in Equation 8.1 is well-posed.

Differentiation can be transformed into a well-posed problem by using Tikhonov's regularizing operators. For equations of the convolution type, such as (2.1) with $K(x, y) = K(x - y)$, the regularizing operators correspond

to convolving $g(x)$ with a filter $f(x,\lambda)$ (where $\lambda > 0$ is the regularization parameter) whose Fourier transform $\tilde{F}(\omega,\lambda)$ satisfies the following conditions:

(i) $0 \leq \tilde{F}(\omega,\lambda) \leq 1$ for $\lambda \geq 0$ and all ω;

(ii) $\tilde{F}(\omega,\lambda)$ is an even function with respect to ω and $\in L_2(-\infty,+\infty)$;

(iii) $\tilde{F}(\omega,\lambda)j\omega$ belongs to $L_2(-\infty,+\infty)$ for any $\lambda > 0$;

(iv) $\lim_{\lambda \to 0} \tilde{F}(\omega,\lambda) = 1$.

This regularizing filter is equivalent to a smooth low pass filter. Three main types of filtering have been used in edge detection. We will list their main properties below.

8.2. Band-limited filters

Band-limited filters are an obvious choice for regularizing differentiation, since the simplest way of avoiding 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[47],[48] and have already been used in edge detection[49]. . These filters satisfy all conditions of Tikhonov needed to regularize differentiation, if we take the inverse of the band-width as the regularization parameter.

8.3. Support-limited filters

All real filters have a finite extension and are support-limited. A class of support-limited filters that has been used in edge detection[50] is the so-called difference of boxes (DOB). These filters are Haar functions [51] which form a basis for square integrable functions on a bounded interval. However, these filters do not satisfy Condition (iii) above, 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 for ω going to infinity tends to zero as ω^{-1}.

It is possible, however, to introduce smooth support-limited filters whose Fourier transform tends to 0 as desired as $\omega \to \infty$. If the function $f(x,\lambda)$ has, for instance, continuous derivative up to order p and the $(p+1)^{\text{th}}$ derivative is integrable, then $\left|\tilde{F}(\omega,\lambda)\right|$ tends to zero as $|\omega|^{-(p+1)}$. Furthermore, if $f(x,\lambda)$

is C^∞, then $\tilde{F}(\omega, \lambda)$ tends to zero more rapidly than any inverse power of ω. An example is provided by the function

$$f(x,\lambda) = \begin{cases} C_\lambda \exp\left(\frac{1}{1-(x/\lambda)^2}\right), & |x| \leq \lambda \\ 0, & |x| > \lambda \end{cases} \tag{8.3}$$

where C_λ is a constant such that $\tilde{F}(0,\lambda) = 1$. Therefore, the best support-limited filter for edge detection and numerical differentiation is not the DOB but the filter (8.3), which is often used in digital signal processing when aliasing needs to be reduced.

8.4. Filters with minimal uncertainty

The Gaussian function minimizes the product of spread in the space and in the frequency domain[32] and can be viewed as a filter with minimal uncertainty. Filtering with a Gaussian function regularizes differentiation, because the Gaussian function $f(x,\lambda) = \exp(-x^2/\lambda^2)$ satisfies all conditions of Tikhonov. Moreover, filtering with a Gaussian transforms a continuous and bounded function into an entire function.

8.5. Interpolation and approximation

Numerical differentiation can also be regularized in a different way. It is possible to interpolate or approximate the data with an analytic function and subsequently compute the analytical derivative of the interpolating or approximating function.

For instance in 1D the "image" model is $y_i = f(x_i) + \epsilon_i$, where y_i is the data and ϵ_i represent errors in the measurements. We want to estimate f' from an interpolating or approximating function f. We can choose a regularizing functional $\|Cf\|^2 = \int (f''(x))^2 dx$, where f'' is the second derivative of f. This choice corresponds to a constraint of smoothness on the intensity profile. Its physical justification is that the (noiseless) image is indeed very smooth because of the imaging process: the image is a bandlimited function and therefore has bounded derivatives. We can decide that the data are noiseless and therefore we want to interpolate the data. This procedure is equivalent to interpolating the data with cubic splines and differentiation can be safely obtained by the analytical differentiation of the interpolating spline. If the data are noisy and we want to take errors in the measurements into account, we can look for an approximating function minimizing

$$\sum_i (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx. \tag{8.4}$$

In [52] it was shown (a) that the solution $f(x)$ can be obtained by convolving the data y_i (assumed on a regular grid and satisfying appropriate boundary conditions) with a convolution filter R, and (b) that the filter R is a cubic spline with a shape very close to a Gaussian and a size controlled by the regularization parameter λ. Differentiation can then be accomplished by convolution of the data with the appropriate derivative of this filter. The optimal value of λ can be determined for instance by cross validation[21],[22] and other techniques. This corresponds to finding the optimal scale of the filter[46].

These results can be directly extended to two dimensions. The resulting filters, which are spline filters for discrete data and Butterworth-like filters for continuous data (they are eventually indistinguishable in practice) are very similar to the derivatives-of-a-gaussian extensively used in recent years [46],[53],[54],[55]. If the regularizing functional $\|Cf\|$ is

$$\int\int \left(\nabla^2 \mathrm{grad} f\right)^2 dx\ dy, \tag{8.5}$$

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

$$R_2(x,y) = \frac{1}{2}\int_0^\infty \frac{J_0(\omega z)}{\lambda\omega^6 + 1}\omega d\omega, \tag{8.6}$$

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

$$\int\int \left(\nabla^2 f(x,y)\right)^2 dx\ dy, \tag{8.7}$$

the filter becomes

$$R_2(x,y) = \frac{1}{2}\int_0^{+\infty} \frac{J_0(\omega^2)}{\lambda\omega^4 + 1} d\omega. \tag{8.8}$$

Therefore numerical differentiation can be regularized in a number of ways which 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.

9. Computation of optical flow

A major aim of early vision is the segmentation of the visible scene in regions corresponding to distinct rigid objects. Motion is an important source of information for this goal. The imaging device projects the 3D velocity field of viewed objects into the image as a 2D field. When a moving scene is viewed by a camera it is possible to recover directly the *optical flow*. The optical flow[39] is commonly defined as the distribution of apparent velocities of movement of brightness patterns in an image. Optical flow and the 2D motion field are related and their relationship has been carefully analysed in [45].

In this section we discuss two different approaches to the computation of optical flow. Horn and Schunck[39] 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 assumption 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:

$$\frac{dE}{dt}(x,y,t) = 0. \tag{9.1}$$

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 $E(x,y,t)$. The normal component is not determined and it must be recovered (see [45] for an analysis of the validity of the underlying assumptions).

Hildreth[37],[38] 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 8). Hildreth[37],[38] 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

$$v^{\perp} = \frac{\partial}{\partial t}\nabla^2 E\big|_{\Gamma}, \tag{9.2}$$

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

$$v^{\perp} = \frac{\partial}{\partial t}\frac{\partial^2 E}{\partial n^2}\Big|_{\Gamma}, \tag{9.3}$$

where $\partial^2/\partial n^2$ is the second derivative along the direction of the gradient[46].

In this section we will discuss the ill-posed nature of the recovery of optical flow as proposed by these authors.

9.1. 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. 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

$$\vec{V}(s) = v^{\top}(s)\vec{t} + v^{\perp}(s)\vec{n}. \tag{9.4}$$

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

$$\vec{t} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \quad \vec{n} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}, \tag{9.5}$$

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

$$\vec{g}(s) = g(s)\vec{n}. \tag{9.6}$$

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

$$\|\vec{V}\|_X^2 = \int_\Gamma \vec{V}(s) \cdot \vec{V}(s) ds =$$

$$\int_\Gamma |v^{\top}(s)|^2 ds + \int_\Gamma |v^{\perp}(s)|^2 ds. \tag{9.7}$$

The projection operator is

$$L\vec{V}(s) = v^{\perp}(s)\vec{n}, \tag{9.8}$$

and the set of the solutions of the equation

$$L\vec{V} = \vec{g}, \tag{9.9}$$

with $\vec{g}$ defined by Equation (9.6), is the set of the velocity fields $\vec{V}$ given by

$$\vec{V}(s) = \psi(s)\vec{t} + g(s)\vec{n}, \tag{9.10}$$

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

$$\vec{V}^{+} = \vec{g} \tag{9.11}$$

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

9.2. 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

$$C\vec{V} = \dot{\vec{V}} \tag{9.12}$$

where the dot means derivation with respect to s. Then the C-generalized solution is the velocity field of the form (9.10) which minimizes the functional

$$\|C\vec{V}\|_X^2 = \int_\Gamma \dot{\vec{V}} \cdot \dot{\vec{V}} ds. \tag{9.13}$$

It is easy to show that existence and uniqueness of the C-generalized solution can be derived from the general result given in Section 3.3.

First, consider the question of uniqueness. We know that the C-generalized solution is unique if and only if the intersection of $N(C)$ and $N(L)$ is the null element (Condition (i) of Section 3.3). 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 (9.13).

The existence of the solution follows from the fact that the operator (9.12) satisfies Conditions (ii) and (iii) of Section 3.3. Condition (ii) is a rather general property of differential operators (see Appendix B for comments), 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, $s_1, s_2, \ldots, s_n$, where the tangent has both right and left limit. Then a solution $\vec{V}(s)$ of the form (9.10) 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 assure the continuity of $\vec{V}(s)$. We can derive these conditions from the equations

$$\vec{V}_+(s_i) = \vec{V}_-(s_i); \quad i = 1, \ldots, n, \tag{9.14}$$

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

$$\begin{aligned} \psi_+(s_i) &= (\sin\phi_i)^{-1}\left[g_+(s_i)\cos\phi_i - g_-(s_i)\right] \\ \psi_-(s_i) &= (\sin\phi_i)^{-1}\left[g_+(s_i) - g_-(s_i)\cos\phi_i\right], \end{aligned} \tag{9.15}$$

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 Equation (9.15) the right and left limit of ψ. All these conditions characterize the subset $D(C)$ which contains the unique solution which minimizes the seminorm (9.13). 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[37],[38]: the first is a translation of an arbitrary contour and the second is an 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

$$\dot{\vec{t}} = \dot{\theta}\vec{n} \quad , \quad \dot{\vec{n}} = -\dot{\theta}\vec{t} \tag{9.16}$$

where $\dot{\theta}$ is just the curvature, one can derive from Equation (9.10)

$$\dot{\vec{V}}(s) = \left[\dot{\psi}(s) - \dot{\theta}(s)g(s)\right]\vec{t} + \left[\dot{g}(s) + \dot{\theta}(s)\psi(s)\right]\vec{n} \tag{9.17}$$

and therefore, when ψ satisfies the conditions (9.15)

$$\begin{aligned} \|C\vec{V}\|_X^2 &= \int_\Gamma \{|\dot{g}(s)|^2 + |\dot{\theta}(s)g(s)|^2\}ds + \\ &+ \int_\Gamma \{|\dot{\psi}(s)|^2 + |\dot{\theta}(s)\psi(s)|^2 + 2\dot{\theta}(s)\dot{g}(s)\psi(s) - 2\dot{\theta}(s)g(s)\dot{\psi}(s)\}ds \end{aligned} \tag{9.18}$$

This is a functional of ψ, which is an arbitrary function except for being differentiable and satisfying conditions (9.15). Then, by annihilating the first variation of this functional, it follows that, on each regular arc, the function ψ which minimizes the functional is solution of the differential equation

$$-\ddot{\psi}(s) + |\dot{\theta}(s)|^2\psi(s) + 2\dot{\theta}(s)\dot{g}(s) + \ddot{\theta}(s)g(s) = 0. \tag{9.19}$$

In the case of a closed contour, the C-generalized solution is given by the unique solution of Equation (9.19) satisfying the conditions (9.15). If the contour is regular everywhere, then one has to add boundary conditions such as

$$\psi(0) = \psi(l) \quad , \quad \dot{\psi}(0) = \dot{\psi}(l). \tag{9.20}$$

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 (9.18)

$$\dot{\psi}(0) = \dot{\theta}(0)g(0) \quad , \quad \dot{\psi}(l) = \dot{\theta}(l)g(l). \tag{9.21}$$

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

$$\psi(0) = v^{\top}(0) \quad , \quad \psi(l) = v^{\top}(l), \tag{9.22}$$

where $v^{\top}(0)$ and $v^{\top}(l)$ are the measured values.

If the motion of the contour is a pure translation and $\vec{a} = \{a_1, a_2\}$ is the constant velocity field, the noise-free data are given by

$$g(s) = -a_1 \sin\theta + a_2 \cos\theta. \tag{9.24}$$

Then, if we put

$$\psi(s) = a_1 \cos\theta + a_2 \sin\theta, \tag{9.25}$$

taking into account that $\dot{\psi} = \dot{\theta}g$, $\dot{g} = -\dot{\theta}\psi$, it is easy to verify that ψ satisfies Equation (9.19). In the case of an open contour, also the boundary conditions (9.21) are satisfied (the boundary conditions (9.15) are obvious since the velocity field is continuous).

In the case of a rigid polygon, since an arbitrary rigid motion consists of a translation plus a 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, Equation (9.19) becomes $\ddot{\psi}(s) = 0$ and therefore $\psi(s)$ is a linear function of s. The boundary conditions (9.20) (plus the boundary condition (9.22) 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[37],[38]

$$\Phi_\lambda[\vec{V}] = \|L\vec{V} - \vec{g}\|_X^2 + \lambda\|C\vec{V}\|_X^2. \tag{9.26}$$

If we denote by $\vec{V}_\lambda$ the minimum of the functional (9.26) and if we put

$$\vec{V}_\lambda = \psi_\lambda(s)\vec{t} + \phi_\lambda(s)\vec{n} \tag{9.27}$$

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

$$-\ddot{\psi}_\lambda(s) + 2\dot{\theta}(s)\dot{\phi}_\lambda(s) + |\dot{\theta}(s)|^2\psi_\lambda(s) + \ddot{\theta}(s)\phi_\lambda(s) = 0 \tag{9.28a}$$

$$-\lambda\left[\ddot{\phi}_\lambda(s) + 2\dot{\theta}(s)\dot{\psi}_\lambda(s) + |\dot{\theta}(s)|^2\phi_\lambda(s) + \ddot{\theta}(s)\psi_\lambda(s)\right] + \phi_\lambda(s) = g(s) \tag{9.28b}$$

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 5.

In practice, the most economical method for the computation of $\vec{V}_\lambda$ is perhaps the conjugate gradient method. Regularizing properties of this method[56],[57] can also be used in order to avoid the minimization of (9.26).

In the previous treatment we have neglected the errors in the determination of the contour which imply an approximate knowledge of the operator L (Equation (9.8)). However, if the equation $L\vec{V} = \vec{g} + \partial\vec{g}$, where $\partial\vec{g}$ is the error on the data, is replaced by the equation $(L + \partial L)\vec{V} = \vec{g} + \partial\vec{g}$, where ∂L 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 $\partial\vec{g}$ and in the other case it is $\partial\vec{g} - (\partial L)\vec{V}$). This point of view assumes that the errors in the determination of the contour have been included in the errors on the data.

9.3. Two-dimensional optical flow

As we already recalled at the beginning of this section, Horn and Schunck[39] attempted to recover the optical flow in the entire image and not just on a one-dimensional contour. Their basic equation is Equation (9.1), which, written explicitly, provides the relationship

$$\vec{\nabla}E \cdot \vec{V} = -\partial_t E \tag{9.29}$$

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 tangent and normal. We denote by $\vec{t}$ and $\vec{n}$ the tangent and normal to the level line at the point $\{x,y\}$

$$\vec{t} = |\vec{\nabla}E|^{-1} \begin{pmatrix} \partial_y E \\ -\partial_x E \end{pmatrix} \quad , \quad \vec{n} = |\vec{\nabla}E|^{-1} \begin{pmatrix} \partial_x E \\ \partial_y E \end{pmatrix}. \tag{9.30}$$

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

$$\vec{V}(x,y) = v^{\top}(x,y)\vec{t} + v^{\perp}(x,y)\vec{n}. \tag{9.31}$$

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

$$(L\vec{V})(x,y) = v^{\perp}(x,y)\vec{n}. \tag{9.32}$$

The data will be given by

$$\vec{g}(x,y) = g(x,y)\vec{n}, \tag{9.33}$$

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

$$\vec{V}(x,y) = \psi(x,y)\vec{t} + g(x,y)\vec{n} \tag{9.34}$$

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

9.4. 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 [39]Horn and Schunck (1981) can be formulated in this framework.

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

$$C\vec{V} = \begin{pmatrix} \partial_x \vec{V} \\ \partial_y \vec{V} \end{pmatrix}, \tag{9.35}$$

with the associated seminorm

$$\|C\vec{V}\|_Z^2 = \int_\Omega \{\partial_x \vec{V} \cdot \partial_x \vec{V} + \partial_y \vec{V} \cdot \partial_y \vec{V}\}. \tag{9.36}$$

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[39].

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 3.2. 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 $\vec{n} \cdot \partial_x \vec{n} = \vec{n} \cdot \partial_y \vec{n} = 0$, $\vec{t} \cdot \partial_x \vec{t} = \vec{t} \cdot \partial_y \vec{t} = 0$, $\partial_x \vec{n} \cdot \partial_x \vec{t} = \partial_y \vec{n} \cdot \partial_y \vec{t} = 0$ we obtain

$$\|C\vec{V}\|_Z^2 = \int_\Omega \left\{ |\vec{\nabla} g|^2 + \left(|\partial_x \vec{n}|^2 + |\partial_y \vec{n}|^2 \right) |g|^2 \right\} dxdy +$$

$$\int_\Omega \left\{ |\vec{\nabla} \psi|^2 + \left(|\partial_x \vec{t}|^2 + |\partial_y \vec{t}|^2 \right) |\psi|^2 + 2 \left[\left(\vec{n} \cdot \partial_x \vec{t} \right) \partial_x g + \left(\vec{n} \cdot \partial_y \vec{t} \right) \partial_y g \right] \psi + \right.$$

$$\left. + 2g \left[\left(\vec{t} \cdot \partial_x \vec{t} \right) \partial_x \psi + \left(\vec{t} \cdot \partial_y \vec{n} \right) \partial_y \psi \right] \right\} dxdy. \tag{9.37}$$

In order to find the Euler equation of the functional (9.37) one has to consider a variation of $\psi, \psi \mapsto \psi + h$ 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 Equation (9.37) 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 which minimizes the functional (9.37) is the unique solution of the boundary value problem:

$$\nabla \psi + \left(\left| \partial_x \vec{t} \right|^2 + \left| \partial_y \vec{t} \right|^2 \right) \psi +$$

$$+ 2 \left[\left(\vec{n} \cdot \partial_x \vec{t} \right) \partial_x g + \left(\vec{n} \cdot \partial_y \vec{t} \right) \partial_y g \right] + \left(\vec{n} \cdot \Delta \vec{t} \right) g = 0 \tag{9.38}$$

$$\left. \frac{\partial \psi}{\partial \nu} \right|_{\partial\Omega} = \left(\vec{n} \cdot \frac{\partial \vec{t}}{\partial \nu} \right) g \Big|_{\partial\Omega}, \tag{9.39}$$

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 (9.19) with the boundary conditions (9.21). The boundary condition (9.39) 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 a 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 such a case the C-generalized solution is ill-posed and that one must introduce regularized approximations. These can be obtained by minimizing the analogue of the functional (9.26), and this is precisely the method used in [39].

10. Surface reconstruction

Most algorithms able to recover depth from pairs of stereo images provide sparse depth values; that is, depth is obtained only for special points in the viewed scene. Since a global description of the 3D structure of the viewed scene is desirable, it is useful to consider the problem of recovering a mathematical representation of a visible surface $f(x,y)$ from the sparse data[40],[41].

10.1. Surface interpolation

The original data are a finite set of depth values $z_i = f(x_i, y_i)$, $i = 1, \ldots, n$ (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[40],[41] proposed to find f such that it minimizes the seminorm

$$\|Cf\|^2 = \int \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right] dx dy. \qquad (10.1)$$

Uniqueness of solution is guaranteed by the existence of at least four non-coplanar points $z_i = f(x_i, y_i)$ [40],[41].

This procedure can be seen as an application of generalized inverses in the case of discrete data (see Section 3.4): 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 3.4.

The null space $N(C)$ is composed of the set of functions $f(x,y) = ax + by + c$ with a, b, c constants. These functions consist of all planar surfaces defined in Ω. The null space $N(L)$ has been defined in Section 3.4 and consists of the set of functions such that $f(x_i, y_i) = 0$ for $i = 1, \ldots, n$. 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 non-coplanar points, as required in [40], [41].

10.2. 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 $g_i = f(t_i) + \varepsilon_i$, $i = 1, \ldots, N$ 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 also smooth[39],[40],[41],[42],[43],[44]. This approach can be seen as an application of regularization theory. In Part ne we have already shown that interpolation is an ill-posed problem which can be solved by the use of a generalized inverse. We will present now an approach to interpolation directly originating from regularization theory[58],[59], which clarifies the relationship between splines, regularization theory, and gives a different framework to the results on visual interpolation[40],[41],[42],[43],[44],[45].

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:

$$g_i = F_i(f) + \varepsilon_i = f(t_i) + \varepsilon_i \quad i = 1, \ldots, n, \tag{10.2}$$

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

$$\sum_{i=1}^{n} \Big(f(t_i) - g_i \Big)^2 + \lambda J_m(f), \tag{10.3}$$

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

$$J_m(f) = \int\int_{-\infty}^{+\infty} \sum_{\nu=0}^{m} \binom{m}{\nu} \left(\frac{\partial^m f}{\partial x^\nu \partial y^{m-\nu}} \right)^2 dx dy, \tag{10.4}$$

(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[21],[22]. If $m = 2$ we have the functional (10.1). 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 [58] 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" $t_1, \ldots, t_n$ 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 $\sum_{\nu=1}^{n} \alpha_\nu \phi_\nu(t_i) = 0 (i = 1, \ldots, n)$ 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 [40] and [41]. Moreover, the solution has an explicit representation[58] as:

$$f_{n,m,\lambda}(t) = \sum_{j=1}^{m} c_j E_m(t, t_j) + \sum_{\nu=1}^{n} d_\nu \phi_\nu(t), \tag{10.5}$$

where

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

with

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

and

$$\theta_m = 1/2^{2m-1}\pi\left[(m-1)!\right]^2. \tag{10.7}$$

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

$$\begin{aligned} (K + \rho)C + Td &= g \\ T^\top C &= 0 \end{aligned} \tag{10.8}$$

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 $g = (g_1, \ldots, g_n)$.

10.3. 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 $3D$ 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

$$\int\int \left(\partial^4 f / \partial x^2 \partial y^2\right)^2 dx dy. \tag{10.9}$$

In this case different kinds of doubly cubic splines can be used, according to the available data[60]. The algorithms are then convolution algorithms (see section 8.4).

11. 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 initally formulated in [61] and [62] as the solution of five ordinary differential equations called the characteristic strip equations. Of considerable use in this problem has been the reflectance map $R(p,q)$ [63],[64] 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

$$p = \frac{\partial z}{\partial x} \quad \text{and} \quad q = \frac{\partial z}{\partial y} \tag{11.1}$$

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

$$\vec{n} = \frac{1}{\sqrt{1+p^2+q^2}}\{-p,-q,1\}. \tag{11.2}$$

The reflectance map can be computed from the bidirectional reflectance-distribution function and the light source arrangement[63].

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

$$E(x,y) = R\Big(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\Big) = R(p,q) \tag{11.3}$$

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[65]. In this approach, several equations of the type of Equation (11.3) 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

$$R(p,q) = \vec{n} \cdot \vec{s}. \tag{11.4}$$

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

$$f = \frac{2p}{1+\sqrt{1+p^2+q^2}} \quad g = \frac{2q}{1+\sqrt{1+p^2+q^2}}, \tag{11.5}$$

the reflectance map becomes

$$R(f,g) = \frac{4-(f^2+g^2)}{4+(f^2+g^2)}x$$

$$\left(-\frac{4f}{4-(f^2+g^2)}, -\frac{4g}{4-(f^2+g^2)}, 1\right)\cdot\vec{s} \tag{11.6}$$

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

11.1. 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

$$\int_\Omega \left(E(x,y) - \vec{n}\cdot\vec{s}\right)^2 dxdy + \lambda\int_\Omega (n_x^2+n_y^2)dx\ dy, \tag{11.7}$$

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:

$$\int_\Omega |(E(x,y) - R(f,g)|^2 dxdy + \lambda\int_\Omega (f_x^2+f_y^2+g_x^2+g_y^2)dx\ dy, \tag{11.8}$$

with $R(f,g)$ given by Equation (11.6). The variational problem is not quadratic in the unknown $\{f,g\}$ and the results of nonlinear inverse problems have to be used.

11.2. Regularization of shape from shading

We give an application of the result stated in Section 6 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:

$$\|u\|_X^2 = \int_\Omega p^2(x,y)dxdy + \int_\Omega q^2(x,y)dxdy. \tag{11.9}$$

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 (11.2), (11.4) we define a nonlinear operator $A: X\to Y$ as follows:

$$(Au)(x,y) = \frac{s_3 - ps_1 - qs_2}{\sqrt{1+p^2+q^2}}. \tag{11.10}$$

Because $\vec{n}$ and $\vec{s}$ are unit vectors, it is obvious that $|(Au)(xy)| \leq 1$ 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

$$|s_3 - ps_1 - qs_2| \leq \sqrt{1 + p^2 + q^2}, \sqrt{1 + p_n^2 + q_n^2} \geq 1, \tag{11.11}$$

it follows that

$$|Au - Au_n| \leq |s_1||p - p_n| + |s_2||q - q_n| + \left|\sqrt{1 + p^2 + q^2} - \sqrt{1 + p_n^2 + q_n^2}\right|. \tag{11.12}$$

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

$$|Au - Au_n|^2 \leq \left(|p - p_n|^2 + |q - q_n|^2\right). \tag{11.13}$$

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

Finally, we consider the constraint operator C defined by

$$\|Cu\|_Z^2 = \int_\Omega \left[C_o(p^2 + q^2) + C_1(p_x^2 + p_y^2 + q_x^2 + q_y^2)\right] dxdy \tag{11.14}$$

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

$$\int_\Omega |(Au)(x, y) - g(x, y)|2dxdy + \lambda\|Cu\|_Z^2, \tag{11.15}$$

where the first term in Equation (11.15) is Equation (11.10) and the constraint operator C is defined in Equation (11.14). Because the operator A is continuous and the constraint operator has a compact inverse, the results presented in Section 6 indicate the existence of at least a local minimum of the functional (11.15). Furthermore, if $g \in R(A)$, the u_λ converges to an exact solution when $\lambda \mapsto 0$.

12. Discussion

The review of early vision presented in Part Two shows that certain regularization techniques can be useful for a correct and sound "solution" of many vision problems. The key idea of all regularization techniques is to introduce *a priori* knowledge — or *constraints* — which have to satisfy the solution. Therefore we will have different solutions according to the assumptions we have made, that is our *a priori* knowledge about the world.

Physical plausibility of the solution is the most important criterion in selecting 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 plays 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. Standard regularization, however, offers a restricted universe of possible constraints since only certain *a priori* assumptions can be translated into the language of Tikhonov stabilizers.

In our example of the computation of motion, the constraint of smoothness is justified by the observation that the projection of three-dimensional objects in motion onto the image plane tends, in a probabilistic sense, to yield smoother velocity fields [37]. In the case of edge detection the constraint of a small norm for the derivative of image intensity is directly justified by the bandlimiting properties of the optics. In the case of motion, however, and more dramatically in the case of surface reconstruction, the constraint of smoothness is not always correct. This suggests – as we will discuss later – that more general stabilizing functionals are needed to deal with the general problem of discontinuities.

A method for checking physical plausibility of a variational principle is to use the Euler-Lagrange equation associated with the variational problem. In the computation of optical flow, the following *sufficient and necessary* condition has been obtained [66] (see also Section 9) for the solution of the variational principle Equation (9.26), to be the correct physical solution:

$$\vec{t} \cdot \frac{\partial^2 \vec{V}}{\partial s^2} = 0, \tag{12.1}$$

where $\vec{t}$ is the tangent vector to the contour and $\vec{V}$ is the true velocity field. The equation is satisfied by a uniform translation or expansion and by rotation only if the contour is polygonal. Therefore, the smoothness principle will give correct results when (a) motion can be approximated locally by pure

translation, rotation or expansion, or (b) objects have images consisting of connected straight lines. In other situations, the smoothness principle will not yield the correct velocity field, but may yield one that is qualitatively similar and close to human perception[37],[38]. In the corresponding case for edge detection (intended as numerical differentiation), the solution is correct *if and only if* the intensity profile is a polynomial spline of appropriate degree.

From a more biological point of view, a careful comparison of the various "regularization " solutions with human perception promises to be a very interesting area of research, as suggested by Hildreth's work on the computation of motion. For some classes of motions and contours, the solution of Equations (9.19) and (9.22) is not the physically correct velocity field. In these cases, however, the human visual system also appears to derive a similar, incorrect velocity field[37],[38].

It may be useful to remember again that Tikhonov stabilizers do not represent the only way to regularize ill-posed problems. Different or additional constraints such as shape (monotonicity, convexity) have been proposed.

The most obvious way to solve inverse ill-posed problems without requiring smoothness of the solution is to use Markov Random Fields as proposed in [36]. In this approach, discontinuities can be preserved introducing appropriate line processes[34],[35] and appropriate potential terms. A possible problem in this approach is the critical dependency of the solution on the parameters coupling the different Markov Random Fields.

12.1. Stereo matching

Not all inverse problems of early vision can be solved using the regularizing techniques introduced in Part One. For example, stereopsis, 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

$$P_R f(x) \mapsto R(x + f(x)). \tag{12.2}$$

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

$$L(x) = P_R d(x). \tag{12.3}$$

The operator in Equation (12.3) 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

$$\|L(x) - R(x + d(x))\|. \tag{12.4}$$

A sufficient condition for the solution of (12.4) 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

$$\|L(x) - R(x + d(x))\| + \lambda \|d'(x)\|. \tag{12.5}$$

The second term in (12.5) 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, Equation (12.5) does not provide a physically plausible solution. Equation (12.5) requires $d(x)$ to be continuous and differentiable. Equation (12.5) is, however, valid if the disparity gradient is strictly less than 2 (Julesz' definition): in this case there are no occlusions and Equation (12.5) provides a physically plausible solution.

Another problem with Equation (12.5) 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.

12.2. Pseudoinverses versus regularized solutions

It may be useful to summarize some mathematical conclusions on the relationship between pseudoinverses and regularized solutions that may be relevant in early vision.

(1) When we are dealing with operators defined on Hilbert spaces X, Y of the type

$$L : X \mapsto Y, \tag{12.6}$$

we have three main cases:

(a) if L is injective, linear, and continuous, and the range of L is given by $R(L) = Y$, the inverse problem is trivially well-posed because the inverse operator is continuous;

(b) if L is not injective, and $R(L)$ is closed, then by the method of pseudosolutions, the inverse problem becomes well-posed in the sense that the generalized inverse is continuous;

(c) if $R(L)$ is not closed, the use of pseudosolutions by itself does not guarantee the existence and continuity of the inverse solution in the case of noisy data; then, we have to apply regularization methods. This case is the normal one in early vision problems, because noise is always present in the data.

(2) When we are considering operators defined on finite dimensional spaces R^n and R^m, of the type

$$L : R^n \mapsto R^m \tag{12.7}$$

we have again three main cases:

(a) If p denotes the rank of the matrix associated with the operator L and $p = n = m$, then L is injective and $R(L) = R^m$. A solution always exist and is unique and the inverse problem is well defined.

(b) If $p < n$, then uniqueness does not hold, but it can be restored by considering the generalized Moore-Penrose inverse.

(c) If $p < m$, then existence does not hold for arbitrary data but it can be restored again by considering the Moore-Penrose inverse.

For operators in finite dimensional spaces, the inverse or generalized inverse is always continuous. Therefore with finite dimensional spaces, the use of generalized Moore-Penrose inverses is sufficient to guarantee well-posedness and the use of regularization methods is not strictly required. It is useful to remember again that well-posed problems can be ill-conditioned, and in such a case it is necessary to use regularization methods as in the case of ill-posed problems. This is the case when the input data are very noisy and when differentiation on the input data is required, as in the recovery of optical flow or edge detection. Finally, it is useful to observe that:

(3) The distinction between interpolation and approximation of discrete data is associated with the use of pseudoinverse solutions or regularization

methods. Regularization methods are intrinsically approximating solutions, while pseudosolutions can be seen as interpolating solutions through the original data.

12.3. Learning regularization algorithms from examples

12.4. Limits of regularization methods in vision

We believe that regularizing methods provide a useful and mathematically sound framework for many problems in early vision. Not all problems in early vision, however, can be easily treated in this framework (e.g. stereo matching). Moreover it is often desirable to have solutions that preserve discontinuities. In this case Markov Random Fields formulations[36] are likely to represent more powerful tools.

Beyond early vision, efficient vision systems are likely to use more complex and elaborate *a priori* information and to be equipped with reasoning capabilities which encompass early vision and regularization methods. It is possible, though as yet unclear, that between *early vision* and *high vision* different sources of low-level informations are integrated into a unified representation of surface properties, such as the $2\frac{1}{2}D$ sketch. At this point MRF models could be quite useful in providing a flexible (though complex) tool for sophisticated integration.

13. Appendices

13.1. Appendix A.

For the convenience of the reader, we summarize in this appendix some basic ideas of functional analysis which are used in the paper.

All the questions of existence, uniqueness, and continuity of the solution have a precise meaning (and a precise answer, when the answer is known) if we carefully define the sets X and Y to which the functions u, g (Equation (3.1)) belong. In particular, continuity requires that we define what we mean by the "vicinity" of two functions in X or Y; i.e., we must introduce a metric in both spaces.

Among various possible choices of metrics, the one corresponding to a Hilbert space is the most simple and interesting, since a Hilbert space is the space most similar to the usual Euclidean space.

A Hilbert space X is a linear space of functions, satisfying the following conditions:

(a) For any pair of functions $u, v \in X$ (with * being complex conjugate, a complex (real) valued function $(u,v)_X$, called the scalar product in X, is defined such that:

(i) $(u,u)_X > 0, = 0$, if and only if $u = 0$;

(ii) $(u,v)_X = (v,u)_X^*$;

(iii) $(\lambda u + \mu v, z)_X = \lambda(u,z)_X + u(v,z)_X$ for arbitrary complex (real) numbers λ, μ;

(b) X is complete; i.e., the Cauchy criterion for the convergence of sequences holds true;

(c) X is separable; i.e., there exists a countable orthonormal basis.

The classical example of a Hilbert space is provided by any space of square integrable functions (L^2-spaces), the scalar product being defined by

$$(u,v)_X = \int u(x)v^*(x)dx. \tag{A1}$$

A norm can be introduced in X by means of the relation

$$\|u\|_X = (u,u)_X^{1/2}, \tag{A2}$$

and it satisfies the properties:

(i') $\|u\|_X > 0, = 0$ if and only if $u = 0$;

(ii′) $\|\lambda u\|_X = |\lambda|\|u\|_X$ for any complex (real) λ;

(iii′) $\|u + v\|_X \leq \|u\|_X + \|v\|_X$ (triangle inequality).

Then the distance $\varrho(u, v)$ from u to v is defined by

$$\varrho(u, v) = \|u - v\|_X. \tag{A3}$$

A real-valued functional $p(u)$, defined on X, is called a *seminorm* on X, if it satisfies properties (ii′) and (iii′) of the norm. Then it follows that $p(0) = 0$ and $p(u) \geq 0$ but the condition $p(u) = 0$ does not necessarily imply $u = 0$. The set of the functions such that $p(u) = 0$ is a linear subspace called the null space of p, i.e., it contains the zero of X, and, if it contains u, v, then it contains also $\lambda u + \mu v$, for any complex (real) λ, μ. An example of a seminorm, which is not a norm, is the following:

$$p(u) = \left(\int_0^1 |u'(x)|^2 dx\right)^{1/2}. \tag{A4}$$

The null space of $p(u)$ contains all the constant functions.

An operator from a Hilbert space X into a Hilbert space Y is defined by a mapping which transforms functions of a subset of X (the domain of the operator, denoted by $D(L)$) into functions of a subset of Y (the range of the operator, denoted by $R(L)$). When the domain is a linear subspace and the mapping is linear, the operator is called linear. We use the notation $L : X \to Y$ for denoting a linear operator from X into Y.

If $D(L) = X$ and if, for any sequence u_n converging to u, the sequence Lu_n converges to Lu, then L is a continuous operator. A linear operator L is continuous if and only if it is bounded, i.e., there exists a constant C such that, for any $u \in X$

$$\|Lu\|_Y \leq C\|u\|_X. \tag{A5}$$

The quantity

$$\|L\| = \sup_{u \in X} \frac{\|Lu\|_Y}{\|u\|_X} \tag{A6}$$

is called the norm of the operator L.

A linear continuous operator $L : X \to Y$ always admits an adjoint operator $L^* : Y \to X$, defined by

$$(Lu, v)_Y = (u, L^* v)_X \tag{A7}$$

for any $u \in X$, $v \in Y$. The operator L^* is also linear and bounded and $\|L^*\| = \|L\|$.

Using the definition of the adjoint operator, we can write Equation (A6) in the following form

$$\|L\| = \left\{ \sup_{u \in X} \frac{(L^* Lu, u)_X}{(u, u)_X} \right\}^{1/2}; \tag{A8}$$

therefore, $\|L\|$ is the square root of the supremum of the spectrum of $L^* L$. Here we have used a result which is an extension of the well-known variational property of the maximum eigenvalue of a matrix.

A particular class of linear operators is provided by the linear functionals, which correspond to the case where Y is the space of the real (complex) numbers. Therefore, a functional is an operator which associates numbers to elements of X. Linear continuous functionals are characterized by the Representation Theorem of F. Riesz: let $F(u)$ be a continuous functional on X, then there exists a unique function $\phi \in X$ such that

$$F(u) = (u, \phi)_X \tag{A9}$$

for any $u \in X$.

A linear operator $L : X \to Y$ is said to be compact (or also completely continuous) when it is bounded and transforms any bounded set of X into a precompact set of Y (a precompact set is a set whose closure is compact; i.e., it has the Bolzano-Weierstrass property). Compact operators are interesting since they are the most similar to matrices.

In order to establish the compactness of an operator, one needs compactness criteria in functional spaces. The basic result is the Ascoli-Arzelà Theorem, whose proof is the paradigm of all the proofs of compactness. Ascoli-Arzelà's theorem states that a sequence of continuous functions $u_n(x)$ is precompact if:

(i) The functions $u_n(x)$ are uniformly bounded; i.e., there exists a constant C such that

$$|u_n(x)| \leq C \tag{A10}$$

for any n and any x.

(ii) The functions $u_n(x)$ are uniformly continuous; i.e., for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|u_n(x) - u_n(x')| \leq \varepsilon \tag{A11}$$

for any x, x' with $|x - x'| \leq \delta$ and any n.

A very simple example of a compact operator in a L^2-space is provided by an integral operator

$$(Lu)(x) = \int K(x, y) u(y) dy, \tag{A12}$$

whose kernel $K(x, y)$ is square integrable

$$\int dx \int dy |K(x, y)|^2 \leq +\infty. \tag{A13}$$

The most striking property of compact operators in Hilbert spaces is that they have a singular value decomposition (SVD) similar to that of a matrix. The singular system of a compact operator is defined as the set of the solutions of the coupled equations

$$Lu_k = \alpha_k v_k \quad , \quad L^* v_k = \alpha_k u_k, \tag{A14}$$

where the α_k (the singular values) are positive numbers and the u_k, v_k (the singular functions) are functions in X and Y respectively.

When L is compact, its singular system always exists and has the following properties: the α_k have finite multiplicity and tend to zero when $k \to \infty$ (we exclude here the case of a finite rank operator); the u_k form an orthonormal basis in the orthogonal complement of $N(L)$ and the v_k form an orthonormal basis in the orthogonal complement of $N(L^*)$, i.e. the closure of $R(L)$.

It is easy to verify that a function $g \in Y$ is in the range of L if and only if (Picard conditions)

$$g \in N(L^*)^{\perp} \quad , \quad \sum_k \frac{|(g, v_k)_Y|^2}{\alpha_k^2} < +\infty. \tag{A15}$$

Since $\alpha_k \to 0$ when L is not a finite rank operator, it is clear that an arbitrary function orthogonal to $N(L^*)$ does not always satisfy the second condition in Equation (A15) and therefore $R(L)$ is not closed.

We conclude that the problem (3.1) with L compact is ill-posed.

On the other hand, the range of a finite rank compact operator is closed since its dimension is finite.

Another example of continuous operators with closed range is provided by the projection operators; i.e., linear operators P such that:

(i) $P^* = P$;

(ii) $P^2 = P$.

It follows that $\|P\| = 1$. Furthermore, $N(P)$ is the set of all the functions $u = (I - P)v$, where v is an arbitrary function of X and $R(P)$ is the set of all the functions u such that $u = Pu$. Therefore $R(P)$ is closed. A projection operator is compact if and only if the dimension of $R(P)$ is finite.

13.2. Appendix B

In this Appendix we will outline the proof of the main result stated in Section 3.3; namely, that if the constraint operator C satisfies conditions (i) – (iii), there exists a unique C-generalized solution u_C^+ for any g such that $Pg \in LD(C)$. However, we first show that conditions (ii) and (iii) are satisfied by seminorms defined in terms of differential operators.

Consider in $L^2(0,1)$ the following seminorm:

$$p(u) = \left(\int_0^1 \left| u^{(k)}(x) \right|^2 dx \right)^{1/2}, \qquad (B1)$$

where $u^{(k)} = d^k u/dx^k$. The domain of the operator $Cu = u^{(k)}$ is the set of all functions u which have square integrable derivatives at least up to order k. Therefore, C is not everywhere defined in $L^2(0,1)$ and is not continuous (bounded). However, a differential operator such as C is a typical example of a closed operator, i.e., of an operator satisfying the following conditions[33]: if $\{u_n\} \subset D(C)$ is a sequence convergent to u and such that $\{Cu_n\}$ is a sequence convergent to $v \in Z$, then $u \in D(C)$ and $Cu = v$. Furthermore, in our specific case, $R(C) = L^2(0,1)$ since, given an arbitrary function $v \in L^2(0,1)$, there always exists a function $u \in D(C)$ such that $u^{(k)} = v$. Therefore, C satisfies condition (ii). This condition implies that C has a bounded generalized inverse C^+ since C^+ is just the inverse of the restriction of C to $N(C)^\perp$ (the inverse of a closed operator is also closed and an everywhere-defined closed operator is bounded).

As concerns Condition (ii), notice that $N(C)$ is the set of all the polynomials of degree $\leq k-1$. Therefore, $N(C)$ is a k-dimensional closed subspace. It follows that, whenever L is a linear, continuous operator defined on X with its range in an arbitrary Hilbert space Y, $LN(C)$ is a k-dimensional closed subspace in Y.

Finally, as concerns Condition (i), it is satisfied whenever $N(L)$ contains no polynomials of degree $\leq k-1$.

The proof of the result stated in Section 3.3 is based on the fact that conditions (i) – (iii) imply the following one: there exists a constant β^2 such that, for any $u \in D(C)$

$$\|Lu\|_Y^2 + \|Cu\|_Z^2 \geq \beta^2 \|u\|_X^2. \qquad (B2)$$

This condition is called by Morozov the completion condition[27]. Suppose we define on $D(C)$ the scalar product

$$(u,v)_O = (Lu, Lv)_Y + (Cu, Cv)_Z. \qquad (B3)$$

Then from Condition (i) it follows that $\|u\|_0 = 0$ implies $\|u\| = 0$, and $D(C)$ is complete in the topology induced by this scalar product, i.e., $D(C)$ becomes a Hilbert space. Condition (B2) indeed implies that any Cauchy sequence in the topology of $D(C)$ is also a Cauchy sequence in X.

In order to prove inequality (B2), we introduce the space $W = Y \oplus Z$ and define the operator $B : X \mapsto W$ as follows:

$$Bu = \{Au, Cu\}, \quad u \in D(C). \tag{B4}$$

Then, since B is closed, inequality (B2) is equivalent to stating that B has an inverse B^{-1} and that $R(B)$ is closed in W.

The existence of B^{-1} is an easy consequence of Condition (i). In order to prove that $R(B)$ is closed, let $\{u_n\} \subset D(C)$ be a sequence such that $\{Bu_n\}$ is convergent in W; we must prove that the limit belongs to $R(B)$. Since $\{Bu_n\}$ is convergent in W, it follows that $\{Au_n\}$ is convergent in Y and $\{Cu_n\}$ is convergent in Z. Put $u_n = u_n^{(0)} + u_n^{(1)}$, with $u_N^{(0)} \in N(C)$ and $u_n^{(1)} \in N(C)^{\perp}$. As already remarked, the restriction of C to $N(C)^{\perp}$ has a bounded inverse and therefore there exists a constant γ^2 such that

$$\|Cu_n\|_Z^2 = \|Cu_n^{(1)}\|_Z^2 \geq \gamma^2 \|u_n^{(1)}\|_Z^2. \tag{B5}$$

This implies that $\{u_n^{(1)}\}$ is a Cauchy sequence and therefore it is convergent. Let $u^{(1)}$ be the limit. Since the operator C is closed, $u^{(1)} \in D(C)$ and $\{Cu_n\}$ converges to $Cu^{(1)}$. Now we have $Lu_n = Lu_n^{(0)} + Lu_n^{(1)}$ and both $\{Lu_n\}$ and $\{Lu_n^{(1)}\}$ are convergent. It follows that $\{Lu_n^{(0)}\}$ is also convergent, and, thanks to the closure of $LN(C)$, there exists $u^{(0)} \in N(C)$ such that $Lu_n^{(0)}$ converges to $Lu^{(0)}$. By combining all the results we have that $\{Bu_n\}$ converges to $B(u^{(0)} + u^{(1)})$ and therefore $R(B)$ is closed.

Now, starting from the completion condition (B2), the proof of the existence of the C-generalized solution for any g such that $Pg \in LD(C)$ can be done as in[27]. The proof is just an easy extension of the proof of the classical result that any closed and convex set has a unique element of minimal norm. Notice that in [27], the C-generalized solution is called the solution of the basic problem.

When the operator C has a bounded inverse C^{-1}, conditions (i) - (iii) are obviously satisfied. In such a case, the C-generalized L_C^+ is given in [3]:

$$L_C^+ = C^{-1}(LC^{-1})^+, \tag{B6}$$

where $(LC^{-1})^+$ is the generalized inverse of the operator $LC^{-1} : Z \to Y$.

An example of an operator C satisfying this assumption is the following: take $X = L^2(0,1)$ and $Z = L^2(0,1) \oplus L^2(0,1)$ and define C by

$$Cu = \{u, u'\} \tag{B7}$$

with domain the set of the absolutely continuous functions with square integrable first derivative. Then,

$$\|Cu\|_Z^2 = \int_0^1 |u(x)|^2 dx + \int_0^1 |u'(x)|^2 dx, \tag{B8}$$

and we have a functional of the type (5.4).

13.3. Appendix C

In this appendix we show that the problem of linear interpolation is equivalent to the computation of a generalized solution in a suitable space.

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

$$(u,v)_X = u(0)v(0) + \int_0^1 u'(x)v'(x)dx. \qquad (C1)$$

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

$$u(x) = u(0) + \int_0^x u'(x')dx' \qquad (C2)$$

it follows that

$$u(x) = (u, Q_x)_X, \qquad (C3)$$

where

$$Q_x(x') = 1 + \min\{x, x'\}. \qquad (C4)$$

Clearly $Q_x \in X$ for any x, and therefore 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

$$Q(x,x') = Q_x(x') = Q_{x'}(x), \qquad (C5)$$

and its name is due to the relation

$$(Q_x, Q_{x'})_X = Q(x,x'). \qquad (C6)$$

Assume now that a function $u \in X$ is specified at the points $x_1, x_2, \ldots, x_N$ ($x_n \in [0,1]$) and let $g_1, g_2, \ldots, g_N$ be its values. The interpolation problem (i.e., find $u \in X$ such that $u(x_n) = g_n$ for $n = 1, \ldots, N$) can be formulated, thanks to (C3), as the problem of determining $u \in X$ such that

$$(u, Q_{x_n}) = g_n \quad ; \quad n = 1, \ldots, N \qquad (C7)$$

and therefore it takes the form (3.12), (3.13). If we recall that the generalized solution is orthogonal to $N(L)$ (Section 3) and that $N(L)$ is the orthogonal complement of the subspace spanned by the functions

$$\phi_n(x) = Q(x_n, x) \qquad (C8)$$

(L is defined as in Equation (3.14)), we conclude that the generalized solution must be a linear combination of the functions ϕ_n

$$u^+(x) = \sum_{n=1}^{N} c_n Q(x_n, x). \qquad (C9)$$

From Equation (C4) it follows that $u^+(x)$ is just the linear interpolation of the data g_n.

Interpolation by means of splines of degree $m = 2k - 1 \quad (k \geq 1)$ can be obtained along similar lines by a suitable definition of the reproducing kernel Hilbert space X [12] . Interpolation by means of natural splines of the same degree[13] can be formulated as the problem of determining, in the same space, a C-generalized solution which minimizes the seminorm (B1).

References

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

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

[3] Bertero, M., "Regularization methods for linear inverse problems," in *Inverse Problems*, C.G. Talenti, ed. Springer, Berlin, 1986.

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

[5] Hadamard, J., *Lectures on the Cauchy Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, 1923.

[6] Courant, R., and D. Hilbert, *Methods of Mathematical Physics Vol. 2*, Interscience, London, 1962.

[7] Nashed, M. Z., ed., *Generalized Inverses and Applications*, Academic Press, New York, 1976.

[8] Groetsch, C.W., *Generalized Inverses of Linear Operators*, Dekker, New York, 1977.

[9] Lavrentiev, M., *Some Improperly Posed Problems of Mathematical Physics*, Springer-Verlag, Berlin, 1967.

[10] Payne, L.E., "Improperly posed problems in partial differential equations," *SIAM Regional Conference Series in Applied Math SIAM*, 1975.

[11] Groetsch, C.W., "The theory of Tikhonov regularization for Fredholm equations of the first kind," *Research Notes in Math 105*, Putnam, Boston, 1984.

[12] Bertero, M., C. DeMol, and E.R. Pike, "Linear inverse problems with discrete data: I-general formulations and singular system analysis," *Inverse Problems 1*, 301-330, 1985.

[13] Greville, T.N.E., ed., *Theory and Application of Spline Functions*, Academic Press, New York, 1969.

[14] Tikhonov, A.N., "Solution of incorrectly formulated problems and the regularization method," *Soviet Math. Dokl. 4*, 1035–1038, 1963.

[15] Ivanov, V.K., "On linear problems which are not well-posed," *Soviet Math. Dokl. 3*, 981–983, 1962.

[16] Ivanov, V.K., "The approximate solution of operator equations of the first kind," *USSR Comp. Math. Math. Phys. 6*, 197-205, 1966.

[17] Morozov, V.A., "On the solution of functional equations by the method of regularization ," *Soviet Math. Dokl. 7*, 414-417, 1966.

[18] Miller, K., "Least squares methods for ill-posed problems with a prescribed bound," *SIAM J. Math. Anal. 1*, 52-74, 1970.

[19] Franklin, J.N., "On Tikhonov's method for ill-posed problems," *Math. Comp. 28*, 889–907, 1974.

[20] Reinsch, C.H., "Smoothing by spline functions," *Numer. Math. 10*, 177–183, 1967.

[21] Wahba, G., "Practical approximate solutions to linear operator equations when the data are noisy," *SIAM J. Numer. Anal. 14*, 1977.

[22] Wahba, G., "Ill-posed problems: numerical and statistical methods for mildly, moderately and severely ill-posed problems with noisy data," *Technical Report 595*, Univ. of Wisconsin, Madison, WI, 1980.

[23] Craven, P., and G. Wahba, "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math. 31*, 377–403, 1979.

[24] Kantorovich, L.V., and G.P. Akilov, *Functional Analysis in Normed Spaces*, Pergamon Press, 1964.

[25] Roger, A., "Newton-Kantorovich algorithm applied to an electro-magnetic inverse problem," *IEEE Trans. Antennas Propagat. AP-20*, 232–238, 1981.

[26] Tikhonov, A.N., "Solution of nonlinear integral equations of the first kind," *Soviet Math. Dokl. 5*, 835-838, 1964.

[27] Morozov, V.A., *Methods for Solving Incorrectly Posed Problems*, Springer-Verlag, New York, 1984.

[28] Nashed, M.Z., "Generalized inverses, normal solvability, and iteration for singular operator equations," in *Nonlinear Functional Analysis and Applications* , L.B. Rall, ed., Academic Press, New York, 1971.

[29] Strand, O.N., and E.R. Westwater, "Minimum-RMS estimation of the numerical solution of a Fredholm integral equation of the first kind," *SIAM J. Numer. Anal. 5*, 1968.

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

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

[32] Papoulis, A., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1965.

[33] Balakrishnan, A.V., *Applied Functional Analysis*, Springer-Verlag, Berlin, 1976.

[34] Geman, S. and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. PAMI*, in press, 1984.

[35] Marroquin, J. "Surface reconstruction preserving discontinuities," *Artificial Intelligence Laboratory Memo 792*, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[36] Marroquin, J., S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *Proc. Image Understanding Workshop*, L. Baumann, ed., SAIC, McLean, VA, 1985.

[37] Hildreth, E.C., *The Measurement of Visual Motion*, Massachusetts Institute of Technology Press, Cambridge, MA, 1984.

[38] Hildreth, E.C., "Computation of the velocity field," *Proc. R. Soc. Lond. B. 221*, 189–220, 1984.

[39] Horn, B.K.P., and B.G. Schunck, "Determining optical flow," *Artificial Intelligence 17*, 185–203, 1981.

[40] Grimson, W.E.L., *From Images to Surfaces: A Study of the Human Early Visual System*, Massachusetts Institute of Technology Press, Cambridge, MA, 1981.

[41] Grimson, W.E.L., "A computational theory of visual surface interpolation," *Phil. Trans. R. Soc. Lond. B 298*, 395–427, 1982.

[42] Terzopoulos, D., "Multilevel computational processes for visual surface reconstruction," *Computer Vision , Graphics, and Image Processing 24*, 52–96, 1983.

[43] Terzopoulos, D., "Multiresolution computation of visible-surface representations," Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984.

[44] Terzopoulos, D., "Multilevel reconstruction of visual surfaces: variational principles and finite element representations," *Artificial Intelligence Laboratory Memo 671*, Massachusetts Institute of Technology. Also printed in

Multiresolution Image Processing and Analysis, A. Rosenfeld, ed., Springer-Verlag, New York, 237-310, 1984.

[45] Verri, A. and T. Poggio, "Motion field and optical flow: qualitative properties," *Artificial Intelligence Laboratory Memo 917*, Massachusetts Institute of Technology , Cambridge, MA, 1986.

[46] Torre, V. and T. Poggio, "On edge detection," *Artificial Intelligence Laboratory Memo 768*, Massachusetts Institute of Technology, Cambridge, MA, 1984. Also printed in *IEEE Trans. PAMI 8*, 147–163, 1986.

[47] Frieden, B.R., "Linear and circular prolate functions," in *Progress in Optics IX*, E. Wolf, ed., North-Holland, Amsterdam, 312–408, 1971.

[48] Landau, H.J., and H.O. Pollack, "Prolate spherical wave functions, Fourier analysis and uncertainty – II," *Bell Syst. Tech. J. 40*, 65–84, 1961.

[49] Shanmugan, K.F., F.M. Dickey, and J.A. Green, "An optimal frequency domain filter for edge detection in digital pictures," *IEEE Trans. PAMI 44*, 99–149, 1965.

[50] Hermuth, H.H., *Transmission of Information by Orthogonal Functions*, Springer-Verlag, Berlin, 1972.

[51] Herskovitz, A., and T.O. Binford, "On boundary detection," */AIL Memo 183*, Massachusetts Institute of Technology , Cambridge, MA, 1980.

[52] Poggio, T., H. Voorhees, and A. Yuille, "Regularizing edge detection," *Artificial Intelligence Laboratory Memo 776*, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[53] Marr, D. and T. Poggio, "A theory of human stereo vision ," *Proc. Roy. Soc. Lond. B 204*, 301-328, 1979. An earlier version appeared as *Artificial Intelligence Laboratory Memo 451*, Massachusetts Institute of Technology , Cambridge, MA, 1977.

[54] Marr, D., and E.C. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B 207*, 187–217, 1980.

[55] Canny, J F., "Finding edges and lines in images," *Artificial Intelligence Laboratory Technical Report TR-720*, Massachusetts Institute of Technology, Cambridge, MA, 1983.

[56] Kammerer, W.J., and M.Z. Nashed, "On the convergence of the conjugate gradient method for singular linear operator equations," *SIAM J. Numer. Anal. 9*, 165–181, 1972.

[57] Bertero, M., P. Brianzi, M. Defrise, and C. DeMol, "Iterative inversion of experimental data in weighted spaces," *Proc. URSI International Symposium on Electromagnetic Theory*, Budapest, Hungary, 1986.

[58] Duchon, J., "Interpolation des formations de deux variables suivant le principe de la flexion des plaques minces," R.A.I.R.O., *Analyse Numerique 10*, 5–12, 1976.

[59] Wahba, G., and J. Wendelberger, "Some new mathematical methods for variational objective analysis using splines and cross validation," *Monthly Weather Review 108*, 1122–1143, 1980.

[60] Ahlberg, J.H., E.H. Nilson, and J.L. Walsh, "The theory of splines and their applications," in *Mathematics in Science and Engineering 38*, Academic Press, New York, 1967.

[61] Horn, B.K.P., "Shape from shading: a method for obtaining the shape of a smooth opaque object from one view," *Project MAC Internal Report TR-79* and Artificial Intelligence Laboratory Technical Report TR-232, Massachusetts Institute of Technology , Cambridge, MA, 1970.

[62] Horn, B.K.P., "Obtaining shape from shading information," in *The Psychology of Computer Vision*, P.H. Winston, ed., McGraw-Hill, New York, 1975.

[63] Horn, B.K.P. and R.W. Sjoberg, "Calculating the reflectance map," *Applied Optics 18*, 1770–1779, 1979.

[64] Horn, B.K.P., "Hill-shading and the reflectance map," *Proc. of the IEEE 69*, 14–47, 1981.

[65] Ikeuchi, K. "Determining surface orientations of specular surfaces by using the photometric stereo method," *IEEE Trans. PAMI 3*, 661–669, 1981.

[66] Yuille, A., "The smoothest velocity field and token matching schemes," *Artificial Intelligence Laboratory Memo 724*, Massachusetts Institute of Technology, Cambridge, MA, 1983.

[67] Mayhew, J.E.W. and J.P. Frisby, "Psychophysical and computational studies towards a theory of human stereopsis," *Artificial Intelligence 17*, 349–385, 1981.

[68] Poggio, T. "Integrating vision modules with coupled MRFs," *Artificial Intelligence Laboratory Working Paper 285*, Massachusetts Institute of Technology, Cambridge, MA, 1985. See also Poggio, T. and staff "MIT progress in understanding images," in *Proceedings of the Image Understanding Workshop*, L. Bauman, ed., SAIC, McLean, VA, 1987.

[69] Horn, B.K.P., *Robot Vision*, MIT Press–McGraw-Hill, Cambridge–New York, 1986.