Probabilistic Regularization Methods for Low–level Vision

Jose L. Marroquin and Mariano Rivera

CIMAT
Apartado Postal 402
Guanajuato, Gto. 36000
Mexico

Contact author: Jose L. Marroquin
email: jlm@cimat.mx
phone: (52-473)27155, ext. 49534
fax: (52-473)25749

1 Introduction

Current research in Computational Vision follows two main paradigms: in the first one, it is considered that the first task that a visual system has to solve consists in reconstructing, from the set of images that constitute the sensory input, a set of fields that represent, on one hand, the physical properties of the three–dimensional surfaces around the viewer, and on the other, the boundaries between patches that "belong together" in some sense, and thus, that may correspond to the outlines of plausible physical objects in the scene. This process, which is usually called Early or Low–level Vision, is supposed to be performed in natural systems by a set of loosely coupled neural networks (computational modules), each one of which specializes in the reconstruction of a particular field. Thus, specific modules have been proposed for the computation of: brightness edges; depth from stereo, shading and motion; color, lightness and albedo; velocity and optical flow; spatial and spatio–temporal interpolation and approximation, etc.

In the second view, it is noted that many of the problems that have to be solved using vision do not need a complete reconstruction of the three–dimensional world; for a given task, it may be possible to feed the raw sensory data to a network (such as a multi–layer perceptron) which directly generates the desired control commands. The plausibility of this approach is illustrated, for example in (Pomerlau, 1992), where such a network is used for an autonomous navigation task. In this case, however, it is also necessary to determine a set of fields defined on the same lattice as the observations: these fields represent the weights that indicate the relative importance of each pixel value for the subnetwork of the corresponding hidden unit.

In both cases, the determination of the corresponding fields exhibits an important common characteristic: due to the loss of information inherent to the imaging and sensory transduction processes and, in the second case, to the fact that one usually has a limited number of available "examples" to train the network, the values of the fields are constrained by the data, but not determined in a unique and stable way (i.e., the reconstruction problems are mathematically ill–posed). This means that the networks that implement the solutions must incorporate in their structure prior knowledge about the reconstructed fields.

For the sake of clarity, this paper is focused in the reconstruction (multi–module) paradigm (although most of the results may be extended to the action–oriented case as well). The general problem that we consider is then the following:

Suppose that we are given sensory measurements in the form of a set of observed fields g at the nodes of a regular lattice L (usually a square lattice is assumed, although other arrangements are possible). From these measurements, one wishes to reconstruct a field $f = \{f_i, i \in L\}$, given the "direct" equations that model g in terms of f and some noise process n:

$$\phi(g, f, n) = 0 \tag{1}$$

The simplest instance of this problem is image filtering: here, g consists of a single field (the noisy observed image); f is the desired reconstructed image, and the observation model is:

$$g - f - n = 0 \tag{2}$$

Another example is the recovery of depth from stereoscopic pairs of images. Here, the observations $g = (g_L, g_R)$ are the grey levels measured in the left and right retinas, respectively, and f is the associated disparity between pairs of corresponding points (if this "correspondence problem" is solved, and if the geometry of the sensors is known, the actual recovery of depth is a matter of simple geometric computations). If the sites of the lattice are identified by a two–dimensional index $i = (i_x, i_y)$, and assuming horizontal epipolar lines, a simplified direct equation is:

$$g_L(i_x, i_y) - g_R(i_x + f_i, i_y) - n_i = 0$$

for each $i \in L$.

Another example is that of image segmentation; here, the input lattice is partitioned into a set of non–overlapping regions $\{R_1, ...R_M\}$, so that the spatial variation of the observed images is represented by a parametric model $\Psi(i, \theta_k)$ inside region R_k:

$$g_i = \sum_{k=1}^{M} \Psi(i, \theta_k) f_{ik} + n_i \tag{3}$$

where f_{ik} is the indicator variable of region R_k: $f_{ik} = 1$ iff $i \in R_k$ and $\{\theta_1, ..., \theta_M\}$ are the parameter vectors.

In the first example, the field f is underconstrained, because the noise field is not known. In the second one, even in the absence of noise, the field f is not uniquely determined, because there may be many points in the right image with the same grey level of a given point in the left one. Finally, in the third example, non–uniqueness arises because of measurement noise and because neither the parameter vectors nor the indicator variables are known. Similar ambiguous situations arise in other early vision problems for different reasons, and in all these cases it is necessary to introduce additional prior constraints.

In this article we present systematic ways for doing this, and for embedding the solution algorithms in suitable networks.

2 Probabilistic Regularization

The classical way of finding solutions to ill–posed problems is based on regularization methods, where stability and uniqueness of the solution is enforced by the introduction of prior smoothness constraints in the solution. A more general approach — which includes the classical one as a particular case — is probabilistic, and considers f and g as realizations of random fields, so that the reconstruction of f is understood as an estimation problem. The prior knowledge about the solution is expressed in the form of a joint probability distribution for f, that specifies the desired dependencies between values at neighboring sites. In this way, one may specify not only global smoothness constraints (as in standard regularization), but also piecewise smoothness, as well as constraints on the shape of the discontinuities.

The basic tool in this approach is Bayes rule, which specifies the way in which prior information (i.e., the prior distribution P_f) is to be combined with the constraints generated by the observations (i.e., the conditional distribution $P_{g|f}$) to generate the posterior distribution $P_{f|g}$:

$$P_{f|g}(f;g) = \frac{P_f(f)P_{g|f}(f;g)}{P_g(g)}$$

note that since the observations g are given, $P_G(g)$ is a constant. The optimal estimator $\hat{f}^*$ is then obtained as the minimizer of the expected value (taken with respect to the posterior distribution) of an appropriate cost function $C(f, \hat{f})$.

This approach, then, requires the specification of three basic components (besides the cost function): the observation model $P_{g|f}$; the prior distribution P_f and the network that will effect the reconstruction. We will now analyze them in detail.

2.1 The Observation Model

The form of the constraints that sensor measurements impose on the reconstructed field depends upon the particular assumptions that are made about the image formation process. If the random variables $n_i, i \in L$ are assumed to be independent, identically distributed with distribution P_n, then the conditional distribution is found by solving for n in Eq. (1): $n_i = \phi^{-1}(f, g)$ and setting:

$$P_{g|f}(f;g) = \prod_{i \in L} P_n(\phi^{-1}(g,f))$$

which can be written in the general form:

$$P_{g|f}(f;g) = \exp[\sum_{i \in L} -\Phi_i(f,g)] \tag{4}$$

In most cases, the functions Φ_i are quadratic — i.e., the noise is assumed to be Gaussian — although other forms that reduce the influence of gross measurement errors have also been used (see Black and Rangarajan, 1996).

2.2 Prior Distribution

The success of the Bayesian approach depends on the specification of a probability distribution $P_f(f)$ that models the desired behavior of the solution. In particular, one would like to be able to specify a distribution in which fields where neighboring sites exhibit the appropriate dependencies are more probable than those in which these local constraints are violated. A general way of constructing such distributions is by defining an "energy" function $U(f)$, which is formed by a sum of terms that measure the violation of the local constraints. The probability distribution of the field is then given by the Gibbs measure:

$$P_f(f) = \frac{1}{Z} \exp[-U(f)] \tag{5}$$

where Z is a normalizing constant.

More precisely, if we define a neighborhood system $\{N_i, i \in L\}$, that is, a collection of subsets of sites indexed by the sites of L: $\{N_i \subset L, i \in L\}$ with the properties:

$$i \notin N_i$$

$$i \in N_j \Leftrightarrow j \in N_i$$

its *cliques* consist on either single sites or subsets of sites such that any two of them belonging to the same clique are neighbors of each other. With this definition, the energy may be written as:

$$U(f) = \sum_C V_C(f) \tag{6}$$

where C ranges over all the cliques of the neighborhood system, and each "potential function" V_C depends only on $\{f_i, i \in C\}$.

A random field F whose probability distribution is given by (5) and (6) is called a *Markov Random Field* on L (Geman and Geman, 1985, Li, 1995, Chellapa and Jain, 1993).

The potential functions represent the "user interface" of the model, since through them one may specify the desired characteristics of the sample fields. Although they may be arbitrarily specified, there are 3 basic types that are generally used, depending on the characteristics of the desired reconstruction:

1. Piecewise Constant Fields. Here, each f_i may only take a finite (usually small) number of values. These fields are mostly used in segmentation problems, in which case it is often convenient that each f_i takes the form of a binary unit vector whose elements correspond to the indicator variables in Eq. (3). The most widely used potential is the generalized Ising potential for cliques of size 2:

$$\begin{aligned} V_C(f_i, f_j) &= -\beta \text{ , if } f_i = f_j \\ &= \beta \text{ , otherwise} \end{aligned}$$

2. Globally Smooth Fields. This case corresponds to standard regularization; the potentials are obtained as the squares of finite difference approximations of differential operators. For first order differences, one obtains the "membrane" model:

$$V_C(f_i, f_j) = (f_i - f_j)^2 \tag{7}$$

   where i and j denote a pair of nearest neighbor sites in the lattice. The second order model corresponds to the bending energy of a thin plate, and the neighborhood system has cliques that consist of sets of 3 neighboring sites i, j, k lying on a horizontal or vertical straight line, and of sets of 4 sites p, q, r, s lying at the corners of a square whose side equals the lattice spacing. The corresponding potentials are:

$$V_{C_3}(f) = (-f_i + 2f_j - f_k)^2 \tag{8}$$

   and

$$V_{C_4}(f) = \frac{1}{4}(-f_r + f_s + f_p - f_q)^2 \tag{9}$$

   where (r, q) and (s, p) lie at opposite corners of the square.

   If one adopts the observation model (2), and assumes that P_n is a zero–mean Gaussian distribution, the posterior energy becomes equivalent to the discretized functional of standard regularization, and its (unique) maximizer corresponds to the MAP estimator (see below).

3. Piecewise Smooth Fields. This is a very important and general case. There are two basic approaches for the construction of the potentials:

   a) The discontinuities of the field are explicitly modeled by means of an auxiliary "line field" s (originally introduced by Geman and Geman, 1984), which is defined on a "dual" lattice whose sites are between each pair of (horizontal or vertical) neighboring sites of L; s is thus indexed by a pair of indices corresponding to sites of L. Each line element s_{ij} may take values on the set $\{0,1\}$, indicating the absence or presence of a line (discontinuity), respectively (in some models, s is allowed to take non–integer values in the interval $[0,1]$ as well: Geman and Reynolds, 1992; Black and Rangarajan, 1996).

The prior energy takes the form:

$$U(f,s) = \sum_{<i,j>} \Big[(f_i - f_j)^2 s_{ij} + \Psi(s_{ij}) \Big] + \sum_D W_D(s) \tag{10}$$

where $\Psi(s_{ij})$ is a function that assigns a penalty for the introduction of a discontinuity between pixels i and j.

The line potentials $W_D(s)$ assign penalties to different local line configurations. They are summed over the cliques D of a neighborhood system defined on the dual lattice, and they are used to favor, for example, piecewise smooth lines, and to prevent the formation of smooth patches that are too thin or too small.

b) The discontinuities are implicitly modeled by non–quadratic potentials $\rho(f_i - f_j)$, where ρ behaves like a quadratic function for small values of its argument, but grows at a smaller rate as its argument becomes large. The derivatives of these potentials are related to influence functions of robust statistical estimators, and are therefore called robust potentials .

If the term $\sum_D W_D(s)$ is omitted, it is always possible to express (10) in the form of a sum of robust potentials, simply by putting

$$\rho(f_i - f_j) = \inf_{s_{ij}} \Big[(f_i - f_j)^2 s_{ij} + \Psi(s_{ij}) \Big]$$

where the right hand side may be explicitly evaluated in many cases. If certain technical conditions on the ρ function are fulfilled, it is also possible to write a robust potential in the line field form (Charbonnier et. al., 1997). Being able to go from one representation to the other, one may add spatial interaction terms to robust potentials, or use continuation methods that have been developed for robust potentials in the line field case (see Black and Rangarajan, 1996; Blake and Zisserman, 1987).

4. Piecewise Parametric Models. In this case the smooth patches are assumed to follow a parametric model with a relatively small number of parameters; for example, in the case of the reconstruction of the velocity field (optical flow) from a sequence of images, an affine model for the velocity of the form $f_i = Ai + b$ is often used (recall that i is a 2–Dimensional index representing the spatial coordinates), as in (Black, Fleet and Yacoob, 2000). In other cases,

spline models with controlled stiffness are more appropriate (Marroquin et. al., 2000). The problem here is that not only the parameters for each model have to be determined, but also the domain of validity of each model, i.e., a field of indicator variables, as in Eq. (3). The prior constraints refer in this case to the spatial coherence of these domains, and may be enforced by Ising potentials.

A comparison between the performance of methods based on robust potentials and piecewise parametric models is shown in Fig. 1. In the images in the left column, a piecewise smooth image corrupted with uniform noise (top row) is reconstructed using robust potentials in the prior distribution (middle row) and using piecewise parametric (in this case, linear) models (bottom row). In this case, both methods produce similar results because the true gray level variation follows indeed a linear model. The advantage of robust potential methods in this case is that they will keep working even if the variation is not linear, provided it is piecewise smooth, whereas the parametric model solution will fail if the wrong model is used. In the images in the right column, a piecewise constant image corrupted by shot noise (the type of image that might correspond to a noisy classification; top row) is also reconstructed using robust and piecewise parametric models (middle and bottom rows). Note that in this case the robust potential solution reduces the dynamic range and introduces artifacts, such as the bands around the 2 circles, whereas the parametric model solution is much cleaner. Other examples of the application of these approaches to a variety of problems, as well as extensions and theoretical results may be found in: Li, 1995; Chellapa and Jain, 1993; Marroquin et. al., 2000 and Marroquin et. al., 2001).

2.3 Networks

Since the reconstruction is needed at the sites of the pixel lattice L, it is very natural to model the reconstructing network as a Cellular Automaton that consists of an array of processors or cells located also at the sites of L. The state of these processors at a given time t is denoted by $\xi^{(t)} = \{\xi_i^{(t)}, i \in L\}$. The interconnection pattern between processors is specified by the defined neighborhood system. The state of each processor changes from time to time with a rule that depends on its own state and that of its neighbors:

$$\xi_i^{(t+1)} = R(\xi_j^{(t)}, j \in N_i \cup \{i\})$$

Cellular automata (CA) may be deterministic (DCA) or stochastic (SCA), depending on the nature of the rule R.

Given this model for the architecture of a computational module, the important question is how to specify R, so that: in the deterministic case, the DCA has a fixed point and the reconstructed field f is obtained from it, and in the stochastic case, the automaton is regular and f is obtained from time averages of functions of its state.

In the case of globally smooth reconstructions, the energy function is usually convex, and the best estimator is obtained by minimizing this energy. The reconstructing networks are in this case equivalent to distributed iterative methods for matrix inversion (Bertsekas and Tsitsiklis, 1989). They may also be implemented analogically with pure resistor networks (see Marroquin, Mitter and Poggio, 1987).

In the case of piecewise smooth potentials, when these are represented in the line field form, and the term $\sum_D W_D(s)$ is not included, the energy function becomes quadratic in f for a given value of s, and therefore it may be minimized by the methods described in (Bertsekas and Tsitsiklis, 1989). On the other hand, if f is kept fixed, one may find the value of the s variables that minimizes U in closed form. By alternating these 2 steps, one gets an effective algorithm for the computation of the optimal estimator (Geman and Reynolds, 1992; Charbonnier, et. al., 1997). If the energy is represented in terms of robust potentials, often local descent schemes combined with continuation methods are most effective (Blake and Zisserman, 1987).

An important issue in all these cases is the determination of the parameters included in the energy function. In many cases these are hand–adjusted for a given class of images; it is better, however, to determine them automatically, as in Zhang, 1993 or in Chen et. al., 2000.

3 Discussion

For the case of piecewise constant fields, the best estimator is not necessarily obtained by minimizing the posterior energy (i.e., the Maximum a Posteriori or MAP estimator). It has been shown (Marroquin Mitter and Poggio, 1987) That the estimator that maximizes the posterior marginal probabilities (the MPM estimator) has better behavior, particularly for low signal to noise ratios. In both cases, the cost for the exact computation of the optimal estimators is too high, so that approximations must be made. The most precise are obtained with SCA, which mathematically correspond to regular Markov chains whose invariant measures correspond to the posterior distribution $P_{f|g}$. In this case, the law of large numbers for regular chains establishes that the average of any function of the state $Y(\xi)$, taken with respect to $P_{f|g}$ may be approximated arbitrarily well by the time average of $Y(\xi^{(t)})$, obtained by observing the evolution of the automaton. One may use this property for estimating the posterior marginals, by counting the number of times a given cell is in each state, from which the MPM estimator may be obtained. It is also possible to approximate the MAP estimator, by introducing a "temperature" parameter which goes slowly to zero (a procedure known as "Simulated Annealing"; see Geman and Geman, 1984).

The main drawback of these stochastic methods is their computational complexity, since many iterations are needed to obtain accurate results. This is specially important in the case of the estimation of piecewise parametric models, since here the most effective procedures consist of 2 steps, which are alternatively performed in an iterative manner until convergence is achieved. These steps are:

1. Estimate the best segmentation (i.e., the f indicator variables in Eq. (3)), given the model parameters.

2. Estimate the model parameters given the segmentation.

with an appropriate initialization step. Instances of these procedures are found in Marroquin et. al., 2000 and Black, Fleet and Yacoob, 2000.

To perform step 1, it is necessary to have efficient estimators for piecewise constant fields. One way to obtain them is derived from the Mean Field (MF) Theory of Statistical Physics, and is based on the assumption that the mean value $< f_i >$ of a MRF at each site i can be computed considering that the influence of the field on this site can be approximated by the influence of $\{< f_j >, j \in N_i\}$. Note that if f_i is a binary vector of indicator variables as in Eq. (3), $< f_{ik} >, k = 1, ...M$ represent the corresponding posterior marginal probabilities $\Pr(f_i k = 1 \mid g)$, from which the optimal estimators may be computed. The MF–based estimation algorithm may be implemented by a DCA with M layers, where each unit corresponds to a specific marginal probability. The update rule for each node involves the computation of the exponential of the sum of the local contributions of neighboring sites plus a normalization step (see Zhang, 1993).

A different approach is based on the idea of constructing a random field of discrete probability distributions using a Gauss–Markov model, so that the mean value of this field corresponds to the posterior marginal probabilities. Since this field is Gaussian, its mean value is found by the minimization of a quadratic form, which, because of the Markovian property, has a particularly simple structure. The network that computes the optimal estimator is represented in Fig. 2. Note that, unlike the MF network, in this case there is no need neither of exponentiation nor of normalization; as a result, one can get better results at a fraction of the computational cost (see Marroquin et. al., 2000 and Marroquin et. al., 2001).

References

*D.P. Bertsekas and J.N. Tsitsiklis, 1989 Parallel and Distributed Computation; Numerical Methods. Prentice Hall, Englewood Cliff, N.J.

*M. J. Black and A. Rangarajan, 1996 "On the Unification of Line Processes, Outlier Rejection, and Robust Statistics with Applications in Early Vision", International Journal of Computer Vision. 19, 1, p. 57-91.

M. J. Black, D.J. Fleet and Y. Yakoob, 2000 “Robustly Estimating Changes in Image Appearance”, Computer Vision and Image Understanding. 78, p. 8-31.

*A. Blake and A. Zisserman, 1987 Visual Reconstruction. MIT Press. Cambridge, Mass.

*P. Charbonnier, L. Blanc-Feraud, G. Aubert and M. Barlaud, 1998 “Deterministic Edge-Preserving Regularization in Computer Imaging”, IEEE Transactions on Image Processing, 6,p 298-311.

R. Chellapa and A. Jain (editors), 1993 Markov Random Fields: Theory and Practice. Academic Press, Boston.

W. Chen, M. Chen and J. Zhou, 2000 “Adaptively Regularized Constrained Total Least-Squares Image Restoration”, IEEE Transactions on Image Processing. 9,4, p. 588-596.

S. Geman and D. Geman, 1984 “Stochastic Relaxation, Gibbs Distributions and the Bayesian Resoration of Images”. IEEE Transactions on Pattern Analysis and Machine Intelligence. 6, p. 721–741.

D. Geman and G. Reynolds, 1992 “Constrained Restoration and the Recovery of Discontinuities”, IEEE Transactions on Image Processing. 14, p 367-383.

*S.Z. Li, 2001 Markov Random Field Modeling in Image Analysis, Springer-Verlag, New York.

*J.L. Marroquin, S. Mitter and T. Poggio, 1987 “Probabilistic Solution of Ill–Posed Problems in Computational Vision”, Journal of the American Statistical Association. 82, 397, p. 76–89.

J.L. Marroquin, S. Botello, F. Calderon and B.C. Vemuri, 2000, ”The MPM-MAP algorithm for image segmentation”, Proc. 15th. Int. Conf. In Pattern Recognition ICPR-2000, IEEE Comp. Soc., Barcelona, Spain, p. 303-308.

J.L. Marroquin, F. Velasco, M. Rivera and M. Nakamura, 2001 “Gauss-Markov Measure Field Models for Low-Level Vision”, IEEE Transactions on Pattern Analysis and Machine Intelligence. 23,4, p. 337-348.

D.A. Pomerlau, 1991 “Efficient Training of Artificial Neural Networks for Autonomous Navigation” Neural Computation 3, 88–97.

J.Zhang, 1993 “The Mean Field Theory in EM Procedures for Blind Markov Random Field Image Restoration.”, IEEE Transactions on Image Processing. 2, 1, p. 27-40.

Figures

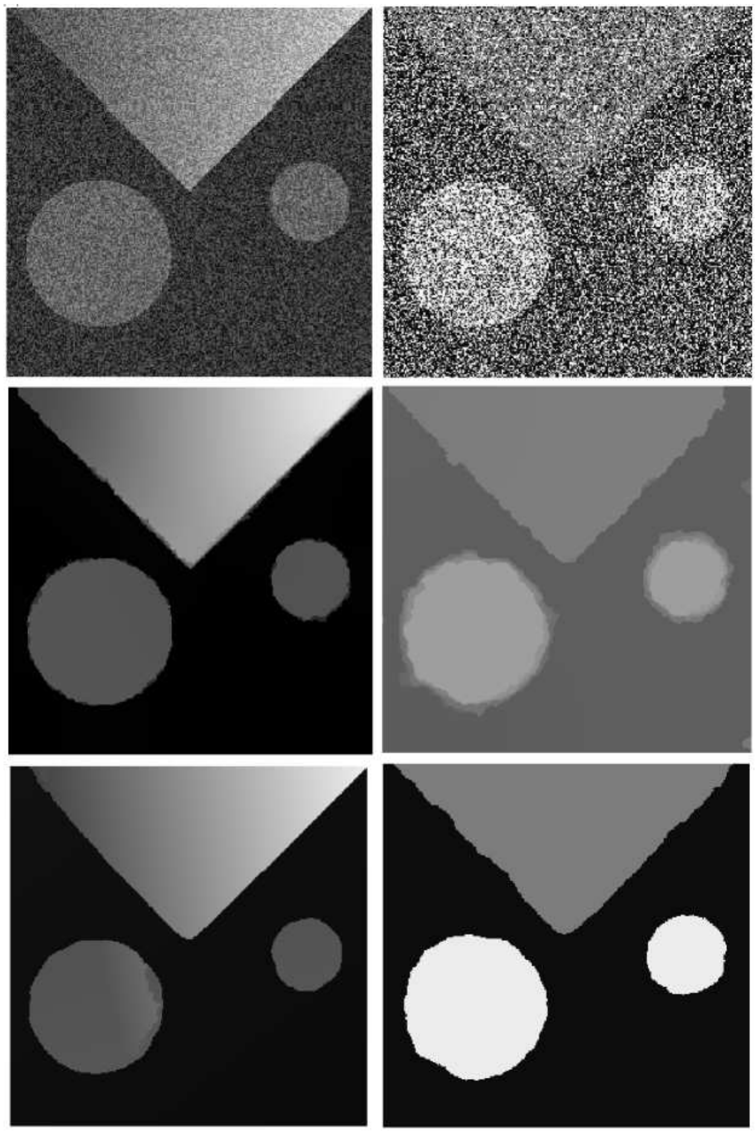

Figure 1: Left column: a piecewise smooth image corrupted with uniform noise (top row) is reconstructed using robust potentials (middle row) and piecewise planar models (bottom row). Right column: a piecewise constant image corrupted by shot noise is also reconstructed using robust and piecewise parametric models (middle and bottom rows).

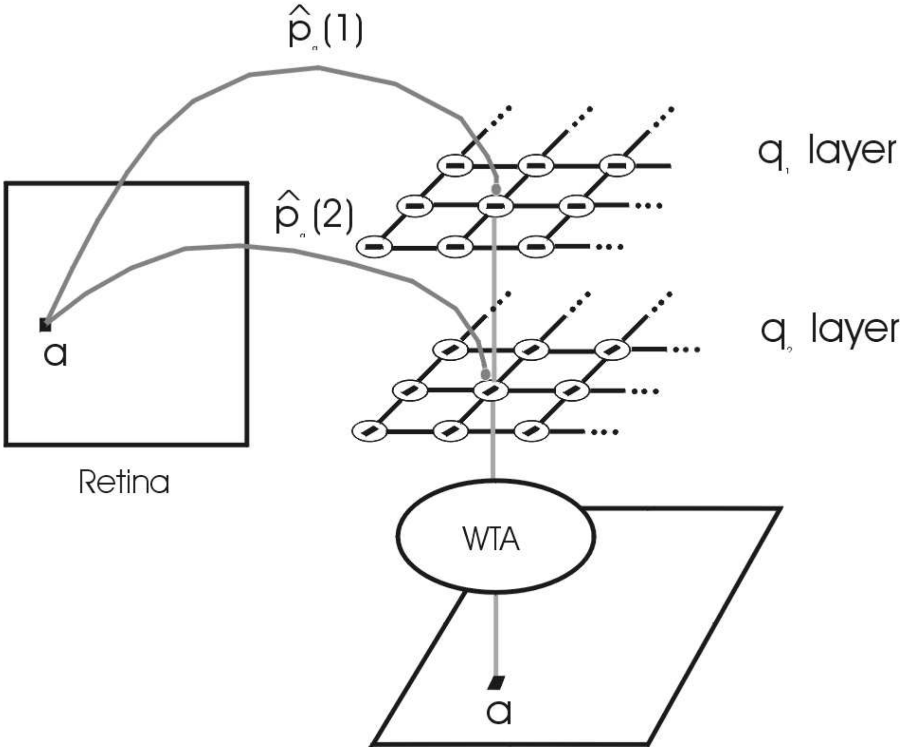

Figure 2: Network that computes the optimal estimator for a discrete-valued field. Each layer corresponds to a valid value for the field (in this example, orientation). Note that the layers are decoupled; they must, however, be synchronized for the system to work properly.