A NONLINEAR DIFFUSION EQUATION AS A FAST AND OPTIMAL SOLVER OF EDGE DETECTION PROBLEMS.

Ilya Pollak[†], *Alan S. Willsky*[†], *Hamid Krim*[‡]

† Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
ipollak@mit.edu

‡ Electrical and Computer Engineering Department
North Carolina State University
Raleigh, NC 27695-7914

ABSTRACT

A nonlinear diffusion process known to be effective for image segmentation is analyzed in 1-D. It is shown that it optimally solves certain edge detection problems. A fast implementation of the algorithm is introduced.

1. INTRODUCTION.

The recent years have seen a great number of exciting developments in the field of nonlinear diffusion filtering of images. Many theories have been proposed that result in edge-preserving scale spaces possessing various interesting properties (see [1, 5, 4, 7, 6] and many other references in [9]). One striking feature uniting many of these frameworks–including our own [6]–is that they are deterministic. Usually, one starts with a set of "common-sense" principles which an image smoothing operation should satisfy. Examples of these are the axioms in [1] and the observation in [5] that, in order to achieve edge preservation, very little smoothing should be done at points with high gradient. From these principles, a nonlinear scale space is derived, and then it is analyzed–again, deterministically. Note, however, that since the objective of these techniques is usually restoration or segmentation of images in the presence of noise, a natural question to ask would be:

> Do the nonlinear diffusion techniques solve standard estimation or detection problems? (*)

Affirmative answer would help us understand which technique is suited best for a particular application, and aid in designing new algorithms. It would also put the tools of the classical detection and estimation theory at our disposal for the analysis of these techniques. There has been a shortage of published attempts to address these issues–most likely, because the complex nature of the nonlinear partial differential equations (PDEs) considered and of the images of interest make this analysis prohibitively complicated. Most notable exceptions are [8, 11] which establish a qualitative relation between between the Perona-Malik equation [5] and gradient descent procedures for estimating random fields modeled by Gibbs distributions. In [2], concepts from robust statistics are used to modify the Perona-Malik equation.

The goal of this paper is to move forward the discussion of question (*). We consider a very simple nonlinear diffusion (a variant of those in [6]) which provides a multiscale sequence of segmentations of its initial condition. We describe an efficient implementation of the diffusion, requiring $O(N \log N)$ computations in the worst case, where N is the size of the input signal. We apply our algorithm to 1-D signals, and describe detection problems which are solved optimally by this diffusion. All results are stated without proof due to space constraints. The proofs will be published in a longer paper, currently in preparation.

2. BACKGROUND AND NOTATION.

A family of systems of ordinary differential equations, called Stabilized Inverse Diffusion Equations (SIDEs), was proposed in [6] for restoration, enhancement, and segmentation of signals and images. The (discretized) image to be processed is taken to be the initial condition for the equation, and the evolution of the equation provides a fine-to-coarse family of segmentations (i.e., piecewise-constant approximations) of the image. This family is indexed by the "scale" (or "time") variable t, which assumes values from 0 to ∞. Initially ($t = 0$), the finest possible segmentation is assumed: each pixel is a separate region. In the course of evolution, two neighboring regions are merged whenever the difference between their intensity values becomes equal to zero (as shown in [6], this will occur in finite time for every pair of regions). The intensity value u_i inside the i-th region evolves according to

$$\dot{u}_i = \frac{1}{m_i} \sum_{j \in A_i} F(u_j - u_i) p_{ij}, \tag{1}$$

where $\dot{u}_i$ is the time derivative of u_i; m_i is the number of pixels in the i-th region; A_i is the set of the indices of all the neighbors of region i; p_{ij} is the length of the boundary between regions i and j; F is a function which is monotonically decreasing and continuous everywhere except at zero; it is an odd function and non-negative for positive values of the argument.

This work was supported in part by AFOSR grant F49620-98-1-0349, ONR grant N00014-91-J-1004, and by subcontract GC123919NGD from Boston University under the AFOSR Multidisciplinary Research Program on Reduced Signature Target Recognition.

The usefulness of SIDEs for image segmentation was shown in [6]; in particular, it was experimentally demonstrated that SIDEs are robust to noise outliers and blurring. They are considerably faster than other image processing algorithms based on evolution equations, since region merging reduces the dimensionality of the system during evolution.

In this paper, we consider a special case of (1) in 1-D, which results if one drops the "monotonically decreasing" requirement on F, and takes $F(v) = \text{sgn}(v)$ instead. Specifically, we are interested in the evolution of the following equation:

$$\begin{aligned} \dot{u}_1 &= \frac{\text{sgn}(u_2 - u_1)}{m_1}, \quad \dot{u}_N = \frac{\text{sgn}(u_{N-1} - u_N)}{m_N}, \\ \dot{u}_n &= \frac{1}{m_n}(\text{sgn}(u_{n+1} - u_n) - \text{sgn}(u_n - u_{n-1})), \quad (2) \\ &\quad \text{for } n = 2, \ldots, N-1, \end{aligned}$$

with the initial condition

$$\mathbf{u}(0) = \mathbf{u}_0, \quad (3)$$

where, as we explain below in Section 5, $\mathbf{u}_0$ is either the signal to be processed or a sequence of logarithms of likelihood ratios. Both here and in the rest of the paper, N stands for the number of samples in the signals under consideration. Boldface letters denote these signals, whose entries are always denoted by the same letter with subscripts 1 through N: $\mathbf{u} = (u_1, \ldots u_N)^T$. Just as in [6], initially $m_n = 1$, for $n = 1, \ldots, N$. As soon as u_i becomes equal to u_{i+1}, these values stay equal forever, and their equations are replaced with

$$\dot{u}_i = \dot{u}_{i+1} = \frac{(\text{sgn}(u_{i+2} - u_{i+1}) - \text{sgn}(u_i - u_{i-1}))}{m_i + m_{i+1}}. \quad (4)$$

We apply the SIDE (2,3) to binary classification problems. Given an observation $\mathbf{y}$, the goal is to label each sample as coming from one of two classes, i.e. to produce a binary signal $\mathbf{h}$ whose entries are zeros and ones. We call any such binary signal $\mathbf{h}$ a *hypothesis*. We denote the set of all N-dimensional hypotheses by $\{0,1\}^N$. If $h_i \neq h_{i+1}$, we say that an *edge* is hypothesized at location i, and we call $\text{sgn}(h_{i+1} - h_i)$ the *sign of the edge*.

A *statistic* is simply a function $\phi : I\!R^N \times \{0,1\}^N \to I\!R$. The *optimal* hypothesis $\mathbf{h}^*(\mathbf{u})$ for a signal $\mathbf{u} \in I\!R^N$ with respect to ϕ is

$$\mathbf{h}^*(\mathbf{u}) \stackrel{\text{def}}{=} \arg \max_{\mathbf{h} \in \{0,1\}^N} \phi(\mathbf{u}, \mathbf{h}).$$

Sometimes it is necessary to choose the best hypothesis among those whose number of edges does not exceed some constant ν:

$$\mathbf{h}^*_{\leq \nu}(\mathbf{u}) \stackrel{\text{def}}{=} \arg \max_{\mathbf{h} \in \{0,1\}^N, \, \mathbf{h} \text{ has } \nu \text{ or fewer edges}} \phi(\mathbf{u}, \mathbf{h}).$$

Note that a hypothesis is uniquely defined by the set of its edges and the sign of one of the edges. Therefore, binary classification problems can also be viewed as edge detection problems. For the problems considered in this paper, the optimal edge locations will typically be level crossings of some signal. A signal $\mathbf{u}$ is said to have an *α-crossing* at location i if $(u_i - \alpha)(u_j - \alpha) < 0$, where $j = min\{n\text{: } n > i, u_n \neq \alpha\}$. We define the hypothesis *generated* by a set of α-crossings $\{g_1, \ldots, g_\nu\}$ of $\mathbf{u}$ as the hypothesis whose edges are at $g_1, \ldots, g_\nu$ and for which the sign of the edge at g_1 is equal to $\text{sgn}(\alpha - u_{g_1})$.

3. SIDE AS AN OPTIMIZER OF A STATISTIC.

The usefulness of the SIDE (2,3) in solving edge detection problems comes from its ability to maximize certain statistics.

Proposition 1 *Fix the initial condition $\mathbf{u}_0$ of the SIDE (2), and let $\mathbf{u}(t)$ be the corresponding solution. Suppose that a statistic ϕ satisfies two conditions:*

1) $\frac{d}{dt}\left\{\phi(\mathbf{u}(t), \mathbf{h}) - \mathbf{h}^T\mathbf{u}(t))\right\} = 0$;

2) *there exists $\alpha \in I\!R$ such that, $\forall t \geq 0$, the optimal hypothesis $\mathbf{h}^*(\mathbf{u}(t))$ is generated by the set of all α-crossings of $\mathbf{u}(t)$.*

Let $\nu_\alpha(t)$ be the number of α-crossings of $\mathbf{u}(t)$. Then

$$\mathbf{h}^*_{\leq \nu_\alpha(t)}(\mathbf{u}_0) = \mathbf{h}^*(\mathbf{u}(t)). \qquad \blacksquare$$

This proposition says that, if the SIDE is evolved until $\nu_\alpha(t)$ α-crossings remain, then these α-crossings are the optimal edges, where "optimality" means maximizing the statistic $\phi(\mathbf{u}_0, \mathbf{h})$ subject to the constraint that the hypothesis have $\nu_\alpha(t)$ or fewer edges. It can be verified that $\nu_\alpha(t)$ is a non-increasing function of time, with $\nu_\alpha(\infty) = 0$. Unfortunately, $\nu_\alpha(t)$ is not guaranteed to assume every integer value between $\nu_\alpha(0)$ and 0. It can be shown, however, that even if for some integer $\nu < \nu_\alpha(0)$ there is no t such that $\nu_\alpha(t) = \nu$, we can still find $\mathbf{h}^*_{\leq \nu}(\mathbf{u}_0)$ using the set of α-crossings of the solution to the SIDE. If $\nu \geq \nu_\alpha(\mathbf{u}_0)$, then, from the definitions of $\mathbf{h}^*_{\leq \nu}(\mathbf{u}_0)$ and $\mathbf{h}^*(\mathbf{u}_0)$, we immediately have:

Proposition 2 *Suppose that ϕ is a statistic which satisfies the two conditions of Proposition 1. If $\nu \geq \nu_\alpha(\mathbf{u}_0)$, then*

$$\mathbf{h}^*_{\leq \nu}(\mathbf{u}_0) = \mathbf{h}^*(\mathbf{u}_0). \qquad \blacksquare$$

4. IMPLEMENTATION.

Our goal is to get $\mathbf{h}^*_{\leq \nu}(\mathbf{u}_0)$ for a given number ν and a given statistic ϕ satisfying the conditions of Proposition 1. Rather than computing $\mathbf{u}(t)$ for all t, we compute the evolution of the set of its α-crossings, since, as explained in the previous section, this set is all we need for determining $\mathbf{h}^*_{\leq \nu}(\mathbf{u}_0)$. We do this by using a different region merging method than that described in [6] and reviewed in Section 2. We re-define a *region* as the set of samples between two α-crossings of $\mathbf{u}_0$. We now give a summary of the algorithm without proof.

1. **Initialize.** Let A be the set of all α-crossings of $\mathbf{u}_0$, ordered from left to right, and let $\bar{\nu} = \nu_\alpha(0)$. If $\bar{\nu} \leq \nu$, stop: by Proposition 2, $\mathbf{h}^*_{\leq \nu}(\mathbf{u}_0)$ is generated by A.
2. **Compute the energies.** Denote the elements of the set A by $g_1, \ldots, g_{\bar{\nu}}$, and form $\bar{\nu} + 1$ regions: $(1, g_1), (g_1 + 1, g_2), \ldots$, $(g_{\bar{\nu}+1}, N)$, where notation (i, j) means a region consisting of samples i through j. Let ρ_{ij} be defined by: $\rho_{ij} = 1$ if $i = 1$ or $j = N$, and $\rho_{ij} = 2$ otherwise. Define the *energy* of (i, j) as $E_{ij} = \frac{1}{\rho_{ij}} \left|\sum_{n=i}^{j}(u_{0,n} - \alpha)\right|$.
3. **Remove the region with minimal energy.** Let (i_m, j_m) be the region for which E_{ij} is the smallest (if there are several regions with the smallest energy, choose any one). Re-define A and $\bar{\nu}$ via

$$A \leftarrow A \backslash \{i_m, j_m\}, \quad \bar{\nu} \leftarrow \text{ the size of the new } A.$$

If $\bar{\nu} = \nu + 1$, let $B = A$. If $\bar{\nu} > \nu$, go to step 2.

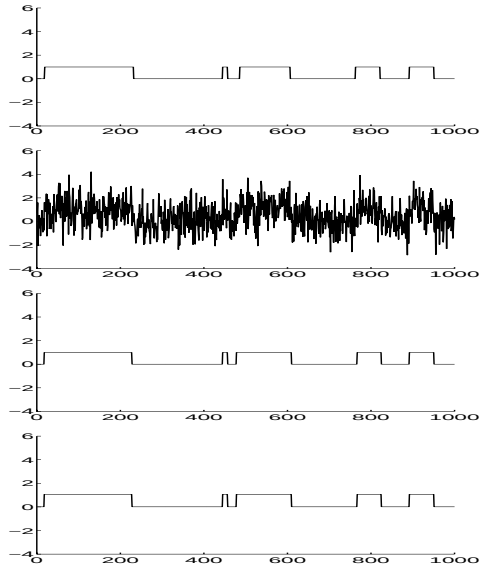

Figure 1: Edge detection for a binary signal in Gaussian noise.

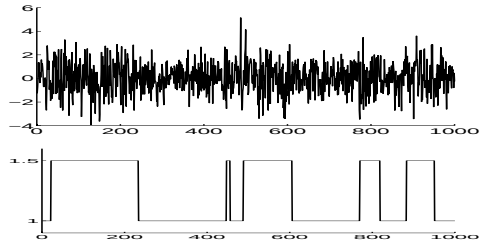

Figure 2: Detection of changes in variance of Gaussian noise.

4. **Post-processing.** If $\bar{\nu} = \nu$, stop: $\mathbf{h}^*_{\leq\nu}(\mathbf{u}_0)$ is generated by A. Otherwise, let b_1 and $b_{\nu+1}$ be the first and last elements of the set B, respectively. Let $\mathbf{h}^1$, $\mathbf{h}^2$, and $\mathbf{h}^3$ be the hypotheses generated by A, $B\backslash\{b_1\}$, and $B\backslash\{b_{\nu+1}\}$, respectively. Then

$$\mathbf{h}^*_{\leq\nu}(\mathbf{u}_0) = \arg \max_{\mathbf{h}\in\{\mathbf{h}^1,\mathbf{h}^2,\mathbf{h}^3\}} \phi(\mathbf{u}_0, \mathbf{h}).$$

Using fast sorting algorithms [3], it is possible to make this implementation run in $O(\sum_{\bar{\nu}=\nu}^{\nu_\alpha(0)+1} \log \bar{\nu} + N)$ time, which is $O(N)$ in the best case and $O(N \log N)$ in the worst case. By contrast, the brute-force method of testing every hypothesis with $1, \ldots, \nu$ edges has polynomial complexity $O(N^{\min(\nu, N-\nu)})$.

5. EDGE DETECTION PROBLEMS OPTIMALLY SOLVED BY THE SIDE.

In this section, we exhibit detection problems whose solution is equivalent to maximizing a statistic satisfying the conditions of Proposition 1, for any initial condition of the SIDE. These problems can therefore be solved by the SIDE.

5.1. Two Distributions with Known Parameters.

Let $\mathbf{y}$ be an observation of a sequence of N independent random variables. Suppose that each random variable has probability density function (pdf) either $f(y, \theta_0)$ or $f(y, \theta_1)$, where θ_0 and θ_1 are known. It is also known that the number of changes between the two pdf's does not exceed ν; however, it is not known where these changes occur.

To obtain the maximum likelihood hypothesis [10], we have to maximize the log likelihood function

$$\sum_{i:h_i=1} \log f(y_i, \theta_1) + \sum_{i:h_i=0} \log f(y_i, \theta_0),$$

where the hypothesis $\mathbf{h}$ is such that the sample y_i is hypothesized to be from the pdf $f(y, \theta_{h_i})$. Note that by making the definitions

$$u_{0,i} = \log f(y_i, \theta_1) - \log f(y_i, \theta_0), \tag{5}$$

we see that the log likelihood is equal to

$$\mathbf{h}^T \mathbf{u}_0 + \sum_{i=1}^{N} \log f(y_i, \theta_0).$$

The second term is independent of $\mathbf{h}$, and therefore maximizing this function is equivalent to maximizing

$$\phi(\mathbf{u}_0, \mathbf{h}) \stackrel{\text{def}}{=} \mathbf{h}^T \mathbf{u}_0, \tag{6}$$

which obviously satisfies the first condition of Proposition 1, for any $\mathbf{u}(t)$. It can also be easily verified that, for any $\mathbf{u}_0 \in I\!R^N$, the hypothesis $\mathbf{h}^*(\mathbf{u}_0)$, optimal with respect to ϕ, is generated by the zero-crossings of $\mathbf{u}_0$. Thus, the SIDE can be employed for finding the maximum likelihood hypothesis $\mathbf{h}^*_{\leq\nu}(\mathbf{u}_0)$, where $\mathbf{u}_0$ is related to the observation $\mathbf{y}$ through (5).

Example 1 *Changes in mean in a Gaussian random vector.*
In this example, $f(y, \theta_j)$ is the Gaussian density with mean θ_j and variance 1. We took $\theta_0 = 0$ and $\theta_1 = 1$. We assumed that the right number of jumps, 10, is known, and so the stopping rule for SIDE was $\nu_\alpha(t) \leq 10$. Figure 1, from top down, depicts the pure mean sequence with ten changes in mean, a corresponding observation $\mathbf{y}$, and the edges detected by the SIDE (the bottom plot will be explained in the next subsection). Note that the result is extremely accurate, despite the fact that the data is very noisy. The computations took 0.6 seconds on Sparc 10, thanks to the fast implementation described in Section 4. ■

Example 2 *Changes in variance in a Gaussian random vector.*
Now $f(y, \theta_j)$ is a zero-mean Gaussian density with standard deviation θ_j; $\theta_0 = 1$ and $\theta_1 = 1.5$. The changes between the two are at the same locations as the jumps in the previous example (see the top plot of Figure 1). The top plot of Figure 2 shows an observation $\mathbf{y}$. Again, we assume that the number of changes is known. The bottom plot of Figure 2 shows the changes detected by the SIDE, depicted as a binary sequence of θ_0's and θ_1's. In addition to being very accurate, the computations took just 0.6 seconds. ■

5.2. Two Gaussian Distributions with Unknown Means.

Suppose that $f(y, \theta_j)$ is the Gaussian density with mean θ_j and variance σ^2. Let $\mathbf{h}$ be a hypothesis, and let $\mathbf{Y}$ be a sequence of N random variables which are conditionally independent given $\mathbf{h}$,

with the i-th random variable Y_i having conditional pdf $f(y, \theta_{h_i})$. Let ν be an upper bound on the number of edges in $\mathbf{h}$. Let K be the number of zeros in $\mathbf{h}$, and define $\sigma_1 = \frac{\sigma}{\theta_1 - \theta_0}\sqrt{N}$. Let the prior knowledge be as follows:

θ_0 and $\mathbf{h}$ are unknown;

σ, σ_1, and ν are known;

K is a random variable with the following discrete Gaussian probability mass function:

$$\mathrm{pr}(K = k) = C \exp\left\{ -\frac{1}{2}\left(\frac{k - \frac{N}{2}}{\sigma_1} \right)^2 \right\}, \quad k = 1, \ldots, N-1,$$

where C is a normalization constant.

Given an observation $\mathbf{y}$ of $\mathbf{Y}$, we seek the best hypothesis in the generalized likelihood ratio sense [10]: the maximum likelihood estimates of the hypothesis and θ_0 are calculated for each value of K, and these estimates are then used in a multiple hypothesis testing procedure to estimate K. In other words, we seek

$$(\hat{\mathbf{h}}, \hat{\theta}_0, \hat{K}) = \arg \max_{\mathbf{h}, \theta_0, k} (\log f_1(\mathbf{y}|\mathbf{h}, \theta_0, k) + \mathrm{pr}(K = k)),$$

where f_1 is the conditional pdf of $\mathbf{Y}$. After simplifying this formula, we obtain that $\hat{\mathbf{h}}$ must maximize

$$\phi(\mathbf{y}, \mathbf{h}) \stackrel{\mathrm{def}}{=} \mathbf{h}^T \mathbf{y} - \frac{N-k}{N} \sum_{i=1}^{N} y_i.$$

Note that $\frac{d}{dt}\left\{ \phi(\mathbf{u}(t), \mathbf{h}) - \mathbf{h}^T \mathbf{u}(t)) \right\} = \frac{N-k}{N} \sum_{i=1}^{N} \dot{u}_i(t) = 0$, as verified by summing up equations (2). Therefore, ϕ satisfies the first condition of Proposition 1, for any solution $\mathbf{u}(t)$ of (2). It can also be easily shown that the second condition is satisfied for any such $\mathbf{u}(t)$, with $\alpha = \frac{1}{N}\sum_{i=1}^{N} u_{0,i}$. Thus, to find $\hat{\mathbf{h}}$, we have to evolve the SIDE whose initial condition is the observed signal: $\mathbf{u}_0 = \mathbf{y}$. We stress here that, even though our model assumes the knowledge of σ and σ_1, they are never used in computing $\hat{\mathbf{h}}$, hence the title of this subsection is justified. The only parameter required by the SIDE is ν.

The experimental result for the data of Example 1 is shown in the bottom plot of Figure 1. Note that the result is still very good and is very close to the maximum likelihood result with known parameters (the two differ in only two pixels out of the thousand).

6. CONCLUSIONS AND CURRENT RESEARCH.

We analyzed a nonlinear diffusion and showed that it produces maximum likelihood solutions for certain edge detection problems. We also presented its fast implementation. We are currently investigating whether similar statements can be made for signals whose samples come from more than two probability distributions.

Even though all analysis done in this paper concerned 1-D signals, we believe that our technique will be most useful in image processing. This is evidenced by Figure 3 which shows the results of running an algorithm, similar to the one described in Section 4, in 2-D. The data on the left is a very blurry and noisy synthetic aperture radar image of two textures: forest and grass. The algorithm was stopped when two regions remained, and the resulting boundary (shown superimposed onto the logarithm of the original image) is extremely accurate. The logarithm of a similarly blurry and noisy ultrasound image of a thyroid is shown on the right, with the boundary detected by the SIDE.

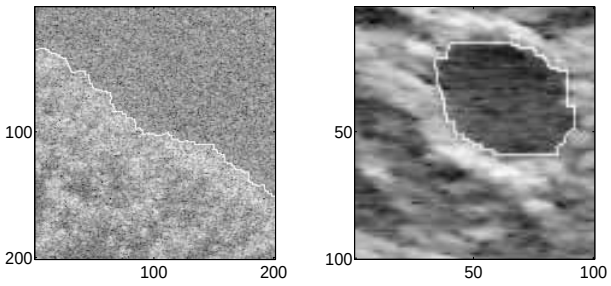

Figure 3: Edge detection in 2-D.

7. ACKNOWLEDGMENTS

We would like to thank Michèle Basseville and Igor Nikiforov for stimulating discussions on change detection.

8. REFERENCES

[1] L. Alvarez, P.L. Lions, and J.M. Morel. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Num. Anal.*, 29, 1992.

[2] M.J. Black, G. Sapiro, D.H. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Trans. on Image Processing*, 7(3), 1998.

[3] T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

[4] S. Osher and L.I. Rudin. Feature-oriented image enhancement using shock filters. *SIAM J. Num. Anal.*, 27(4), 1990.

[5] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. on PAMI*, 12(7), 1990.

[6] I. Pollak, A. S. Willsky, and H. Krim. Image segmentation and edge enhancement with stabilized inverse diffusion equations. Technical Report LIDS-P-2368, Laboratory for Information and Decision Systems, MIT, 1996.

[7] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 1992.

[8] P.C. Teo, G. Sapiro, and B. Wandell. Anisotropic smoothing of posterior probabilities. In *Proc. ICIP*, Santa Barbara, CA, 1997.

[9] B.M. ter Haar Romeny, editor. *Geometry-Driven Diffusion in Computer Vision*. Kluwer Academic Publishers, 1994.

[10] H. van Trees. *Detection, estimation, and modulation theory*, volume 1. Wiley, 1968.

[11] S.C. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. on PAMI*, 19(11), 1997.