HIDDEN MARKOV MEASURE FIELD MODELS FOR IMAGE SEGMENTATION

José L. Marroquín, Edgar Arce and Salvador Botello

Comunicación Técnica No I-02-05/12-04-2002
(CC/CIMAT)

Hidden Markov Measure Field Models for Image Segmentation

Jose L. Marroquin, Edgar Arce and Salvador Botello

Center for Research in Mathematics (CIMAT)
Apartado Postal 402
Guanajuato, Gto. 36000
Mexico

Abstract

Parametric image segmentation consists in finding a label field, that defines a partition of an image into a set of non–overlapping regions, and the parameters of the models that describe the variation of some property within each region. A new Bayesian formulation for the solution of this problem is presented, based on the key idea of using a doubly stochastic prior model for the label field, which allows one to find exact optimal estimators, for both this field and the model parameters, by the minimization of a differentiable function. An efficient minimization algorithm, and comparisons with existing methods on synthetic images are presented, as well as examples of realistic applications to the segmentation of Magnetic Resonance volumes, to motion segmentation, and to edge-preserving filtering.

index terms: Markov Random Fields, Bayesian methods, image segmentation, motion segmentation, edge–preserving filtering.

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

1 Introduction

After the seminal work by Besag [1, 2] and Geman and Geman [3], probabilistic methods, and in particular, Markov Random Field (MRF) models, have been used with great success for the solution of a number of important problems in image analysis; there is a vast amount of published works on the subject, that include applications in image restoration [3, 4, 5, 6], segmentation [7, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16], edge–preserving filtering [17, 18, 19, 20], reconstruction in inverse problems [21, 22], etc. (see also [23], [24], and references contained therein). There are 2 main reasons for this success: discrete MRF's provide a systematic way —firmly rooted in Bayesian estimation theory — for including prior constraints about the shape and average size of homogeneous regions in an image; since these macroscopic properties result from local interactions, a wide variety of behaviors may be obtained, simply by varying a few parameters in the definition of local potentials in the MRF model. The second reason is that, even when exact optimal estimators cannot be precisely computed, it is possible to design reasonable approximate algorithms that work well in many cases, although sometimes with high computational costs.

A particular problem, that has been approached with these kind of models, is image segmentation: it consists in finding a partition of an image into a set of non–overlapping regions $\{R_1, \ldots, R_M\}$, so that the variation of some property (such as intensity, depth, velocity, color, etc.) within each region R_k is either constant, or follows a simple model Φ_k. What makes this problem specially difficult is the fact that one has to estimate both the parameters that characterize each model Φ_k , and the corresponding regions of validity R_k at the same time. To solve it, prior MRF models have been used in conjunction with iterative procedures —in particular, the Expectation Maximization (EM) algorithm [25], which are reasonably effective, but entail a high computational cost. The goal of this paper is to present a new class of probabilistic models that permits the characterization of the solution to complex segmentation problems in terms of the minimization of a differentiable energy function, for which efficient algorithms can be devised. We will show that these models, which are also rigorously based on Bayesian estimation theory, represent a significant improvement over classical MRF's, both in terms of the accuracy of the solutions and of computational complexity, and are also versatile and generally applicable. The plan of our presentation is the following: in section 2, we review the classical MRF formulation of parametric segmentation problems, introduce our new model, and present efficient estimation algorithms. In section 3, we compare experimentally the performance of the new scheme with that of classical ones, and discuss the problem of control parameter selection. In section 4, we present 3 examples of applications, to illustrate the versatility of our approach: to the segmentation of brain Magnetic Resonance (MR) volumes; to motion segmentation and to edge–preserving filtering. Finally, some conclusions are drawn in section 5.

2 Hidden Markov Field Models for Image Segmentation

2.1 Classical MRF Models

The probabilistic models that have been used in most cases to formulate segmentation problems, fit the general description given in Fig. 1. To understand it, we introduce the following notation: let L represent the pixel (or voxel, in 3-D problems) lattice, where images I are observed. The model assumes that there are M regions, $\{R_1, \ldots, R_M\}$, such that $L = \bigcup_{k=1}^{M} R_k; R_i \cap R_j = \emptyset, i \neq j$, so that the observation at pixel $r \in L$ is given by:

$$I(r) = \sum_{k=1}^{M} \Phi(r, \theta_k) b_k(r) + n(r) \tag{1}$$

where $n(r)$ is a white noise field with known distribution P_n (e.g., $\{n(r), r \in L\}$ are 0–mean, independent, identically distributed Gaussian random variables with standard deviation σ); $\Phi(\cdot, \cdot)$ is a parametric model; θ_k is the parameter vector that corresponds to region R_k, and $b_k(r)$ is the corresponding indicator function: $b_k(r) = 1 \Leftrightarrow r \in R_k$; note that $b(r)$ satisfies the constraints:

$$\sum_{k=1}^{M} b_k(r) = 1 \ , \ b_k(r) \in \{0, 1\} \ , \text{for all } r \in L \tag{2}$$

Associated with b, there is a label field f, with $f(r) \in Z_M = \{1, \ldots, M\}$, that indicates to which region does r belong, i.e., $b_k(r) = \delta(f(r) - k)$. In this model, the field f is assumed to be a sample from a MRF, i.e., a sample from the state space Z_M^N (where N is the cardinality of L), obtained with a Gibbsian distribution:

$$P_f(f) = \frac{1}{Z_f} \exp[-\sum_C V_C(f)] \tag{3}$$

where Z_f is a normalizing constant and the sum in the exponent ranges over the cliques of a given neighborhood system on L, and $\{V_C(f)\}$ are "potential functions", each one of which depends only on the value of f at the sites that belong to the clique C (see [3, 23] for details). These potential functions, together with the the neighborhood system selected, control the appearance of the sample field f, and hence, the properties of the estimated segmentation. A potential that is often used is the generalized Ising model, which considers cliques of size 2 (e.g., pairs of sites that are less than 2 units appart), and potentials of the form:

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

where β is a parameter that controls the granularity of the field. Since the field f is not directly observable, it is often called a hidden MRF model.

2.1.1 Estimation Algorithms

The segmentation problem consists in finding an optimal estimator for both the field f and the parameter vector $\theta = (\theta_1, \ldots, \theta_M)$, given the observations I. To obtain it, using Bayesian estimation theory, one follows the steps [6]:

1. Find the likelihood of the observations $P(I|f, \theta)$.
2. Using the prior distribution $P_f(f)$ (and $P_\theta(\theta)$, if available), find the posterior distribution $P(f, \theta|I)$, using Bayes rule.
3. Define an appropriate cost function $C(\hat{f}, \hat{\theta}, f, \theta)$, that associates a cost to estimators $\hat{f}, \hat{\theta}$, given that the true values are f, θ.
4. Find the optimal estimators f^*, θ^* by minimization of:

$$Q(\hat{f}, \hat{\theta}) = E[C(\hat{f}, \hat{\theta}, f, \theta)|I] \tag{4}$$

We now analyze them in detail.

The likelihood of the observations is obtained from the observation model (1) and the noise distribution P_n (assumed known):

$$P(I|f, \theta) = \prod_{r \in L} v_{f(r)}(r) \tag{5}$$

where each M–vector $v(r)$ is defined by:

$$v_k(r) = P\left(I(r)|f(r) = k, \theta\right) = P_n\left(I(r) - \Phi(r, \theta_k)\right) \tag{6}$$

For example, for Gaussian noise, we have:

$$v_k(r) = \sqrt{\frac{\gamma}{\pi}} \exp[-\gamma |I(r) - \Phi(r, \theta_k)|^2] \tag{7}$$

where γ is a parameter that depends on the noise variance.

Using (7), (3) and Bayes rule, one finds the posterior distribution as:

$$P(f, \theta|I) = \frac{1}{Z_p} \exp[-U(f, \theta)]$$

where Z_p is a normalizing constant, and

$$U(f, \theta) = -\sum_{r \in L} \log v_{f(r)}(r) + \sum_C V_C(f) - \log P_\theta(\theta) \tag{8}$$

where a non–informative (constant) prior P_θ may be used, if there are no prior constraints on θ.

The minimization of Q (Eq. (4)) is usually performed by 2–step procedures, which may be generically called Segmentation/Model Estimation (SM) algorithms, in which, the best segmentation, given the current estimate for the model parameters θ is found in step S, and the best estimate for θ, given the current estimate for the segmentation, is found in step M. It has been found that finding a "soft" or "fuzzy" segmentation in the S step, increases the robustness of SM procedures with respect to the initial estimate for θ, which must be given. Thus, it is more convenient to work with the indicator vector field b instead of f, where one imposes the constraints: $\sum_k b_k(r) = 1$ and $b_k(r) \geq 0$, for all r, instead of the "hard" constraints (2).

The general form for SM algorithms is thus:

1: Find an initial estimate $\bar{\theta}$ for θ;

2: Repeat until convergence:

S Step: Set $\bar{b} = \arg\min_{\hat{b}} Q(\hat{b}, \bar{\theta})$

M Step: Set $\bar{\theta} = \arg\min_{\hat{\theta}} Q(\bar{b}, \hat{\theta})$

The cost functions that are normally used (although, in most cases they are not explicitly defined) are separable; expressed in terms of b they are of the form:

$$C(\hat{b}, \hat{\theta}, b, \theta) = C_b(\hat{b}, b) + C_\theta(\hat{\theta}, \theta)$$

Depending on the form of C_b and C_θ, one gets different SM algorithms. For example, setting

$$\begin{aligned} C_\theta(\hat{\theta}, \theta) &= 0 \text{ , if } \hat{\theta} = \theta \\ &= 1 \text{ , otherwise} \end{aligned}$$

and

$$C_b(\hat{b}, b) = \sum_{r \in L} |\hat{b}(r) - b(r)|^2$$

one gets the well known EM algorithm [25] (although this is not its usual derivation): in this case one has that

$$\begin{aligned} Q_b(\hat{b}) &= E[C_b(\hat{b}, b)|I] = \textstyle\sum_{r \in L} \sum_{k=1}^{M} E[(\hat{b}_k(r) - b_k(r))^2|I] \\ &= \textstyle\sum_{r \in L} \sum_{k=1}^{M} \left(\hat{b}_k^2(r) - 2(\hat{b}_k(r) E[b_k(r)|I]) + E[b_k^2(r)|I] \right) \end{aligned}$$

By setting $\partial Q_b / \partial \hat{b}_k(r) = 0$, one gets that $\bar{b}$ in the S step is given by $\bar{b}_k(r) = E[b_k(r)|I] = P(f(r) = k|I)$, i.e., $\bar{b}(r)$ corresponds to the posterior marginal probability distribution

for $f(r)$. $\bar{\theta}$ in the M step is simply the Maximum a Posteriori (MAP) estimate for θ, given $b = \bar{b}$, which is found by minimizing (8), appropriately modified, so that it is expressed in terms of b, i.e., by minimization, with respect to θ of:

$$U(\bar{b}, \theta) = \sum_{r \in L} \sum_{k=1}^{M} \bar{b}_k(r) \log v_k(r) - \log P_\theta(\theta)$$

Other choices for C_b (with the same choice for C_θ) give other SM algorithms, in which a hard segmentation is computed in the S step; for example, using

$$C_b(\hat{b}, b) = \sum_{r \in L} \Big[1 - \delta(\hat{b}(r) - b(r))\Big]$$

where $\delta(x)$ equals 1 if and only if x is the zero vector (i.e., C_b corresponds to the number of segmentation errors), one gets the Maximizer of the Posterior Marginals (MPM estimator) in the S step [6]. Using

$$\begin{aligned} C_b(\hat{b}, b) &= 1 \text{ , if } \hat{b} = b \\ &= 0 \text{ , otherwise} \end{aligned}$$

one gets the MAP estimator for b in the S step [13, 16], etc.

The problem with this class of algorithms is that it is not possible to perform the exact minimization of $Q(\hat{b}, \bar{\theta})$ with respect to $\hat{b}$ in the S step, i.e., neither the posterior marginals, nor the optimal MPM or MAP estimators can be exactly computed, since they involve either the computation of a sum with M^N terms (for the posterior marginals) or the solution of a combinatorial optimization problem with $N = |L|$ variables (for the MAP estimator). Hence, one must resort to approximations; the most precise are based on stochastic, Markov Chain Monte Carlo (MCMC) algorithms [3, 26], and are computationally very expensive; fast approximations (e.g., the ICM algorithm [2]) are highly sensitive to noise. Approximations based on Mean Field theory [27, 28, 19, 29], are faster than MCMC, but still relatively expensive, and also sensitive to noise (see section 3). A recent algorithm for the MPM estimator, presented in [30], based on a Gaussian approximation for the posterior marginals, is fast and resistant to noise; however, since the MPM estimator corresponds to a hard segmentation, the corresponding MPM-MAP procedure is very sensitive to the initial estimate for θ.

2.2 Hidden Markov Measure Field Models

The difficulties mentioned above may be solved if one uses a different probabilistic model for the generation of the label field f. Instead of the 1-step procedure described in Fig. 1, we propose to use a 2-step probabilistic model, with an additional hidden

field p. This model is presented in Fig. 2: on a first step, a Markov random vector field p is generated with distribution $P(p) = \frac{1}{K}\exp[-\sum_C W_C(p)]$, where K is a normalizing constant, C are the cliques of a given neighborhood system, and W_C are given potential functions, and where each vector $p(r)$ takes values on the M–vertex simplex S_M:

$$S_M = \{u \in \Re^M : \sum_{k=1}^{M} u_k = 1 \ , \ u_k \geq 0, k = 1, \ldots, M\} \tag{9}$$

Hence, $p(r)$ may be interpreted as a discrete probability measure on Z_M (the label state space). On a second step, the label field f is generated in such a way that each $f(r)$ is an independent sample from the distribution $p(r)$, so that

$$P(f|p) = \prod_{r \in L} p_{f(r)}(r) \tag{10}$$

Note that the prior for f is:

$$P_f(f) = \int_{S_M^N} P(f|p) dP(p)$$

which is not Gibbsian. To see this, note that for f to be a MRF with respect to any neighborhood system $\{N_r, r \in L\}$, one should have that for any 2 sites r, t, with $t \notin N_r$, $f(r)$ should be independent of $f(t)$, given $f(s), s \in N_r$. Now, the probabilistic dependencies of f are induced by the corresponding dependencies of the p field, which are a consequence of its Markovian structure, and conditioning on $f(s), s \in N_r$ does not alter these dependencies, since

$$\begin{aligned} P(p|f(s), s \in N_r) &= \frac{P(f(s), s \in N_r|p)P(p)}{P(f(s), s \in N_r)} \\ &= \frac{\prod_{s \in N_r} p_{f(s)}(s) P(p)}{\int_{S_M^N} \prod_{s \in N_r} p_{f(s)}(s) dP(p)} = \frac{1}{Z_d} e^{-U(p)} \end{aligned}$$

where Z_d is a normalizing constant and

$$U(p) = \sum_C W_C(p) - \sum_{s \in N_r} \log p_{f(s)}(s)$$

Thus, knowledge of $f(s), s \in N_r$ does not provide complete knowledge of $p(s), s \in N_r$, but only biases $P(p)$ introducing new potentials — that correspond to cliques of size 1 — on $U(p)$, which means that the probabilistic dependencies of p are not altered; this, in turn, implies that the long range dependencies on f are preserved as well. Therefore, even though p is Markovian, f is not a MRF of any order, so that this class of models is different from the classical ones. As in the classical case, however, the

potential functions (for the p field in this case) may be used to enforce the appropriate prior constraints on the label field. The spatial coherence of regions $\{R_1, \ldots, R_M\}$, for instance, may be enforced by requiring that each vector $p(r)$ is similar to its spatial neighbors. A simple quadratic potential that expresses this condition is:

$$W_{rs}(p(r), p(s)) = \lambda |p(r) - p(s)|^2 = \lambda \sum_{k=1}^{M} (p_k(r) - p_k(s))^2 \tag{11}$$

where λ is a positive parameter, and $< r, s >$ are neighboring sites in L. Other potentials may be defined to enforce more complex constraints (see section 4), but here we concentrate on this simple one.

The posterior distribution $P(p, \theta|I)$ is obtained from Bayes rule as:

$$P(p, \theta|I) = \frac{1}{Z} P(I|p, \theta) P_p(p) P_\theta(\theta) \tag{12}$$

where Z is a normalizing constant. The conditional distribution $P(I|p, \theta)$ is obtained as:

$$P(I|p, \theta) = \prod_{r \in L} P(I(r)|p, \theta)$$

The conditional distribution $P(I(r)|p, \theta)$ may be obtained by first computing the joint conditional distribution $P(I(r), f(r)|p, \theta) = P(I(r)|f(r), p, \theta) P(f(r)|p, \theta)$, and then marginalizing over $f(r)$:

$$P(I(r)|p, \theta) = \sum_{k=1}^{M} P(I(r)|f(r) = k, p, \theta) P(f(r) = k|p, \theta)$$

Using the fact that $P(I(r)|f(r), p, \theta) = P(I(r)|f(r), \theta)$ and that $P(f(r) = k|p, \theta) = p_k(r)$, one obtains:

$$P(I(r)|p, \theta) = \sum_{k=1}^{M} v_k(r) p_k(r) = v(r) \cdot p(r) \tag{13}$$

where $v_k(r)$ is given by (7).

From (13) and (12) one finally gets:

$$P(p, \theta|I) = \frac{1}{Z} \exp[-U(p, \theta)] \tag{14}$$

with

$$U(p, \theta) = -\sum_{r \in L} \log(v(r) \cdot p(r)) + \sum_{C} W_C(p) - \log P_\theta(\theta) \tag{15}$$

where we consider cliques of size 2 and potentials given by (11). To obtain the optimal estimator f^* for the label field, we use the following 2-step procedure:

1: Find the MAP estimators p^*, θ^* for p, θ:

$$p^*, \theta^* = \arg \max_{p \in S_M^N, \theta} P(p, \theta | I) \tag{16}$$

2: Find f^* as the maximizer of $P(f | p = p^*, \theta^*, I)$

The first step is equivalent to the minimization of $U(p, \theta)$, given by (15), subject to the constraints:

$$p(r) \in S_M \text{ , for all } r \in L \tag{17}$$

with S_M defined by (9), while the second step consists simply on finding the mode for each discrete measure $p^*(r)$ in a decoupled way:

$$f^* = \arg \max_k p_k^*(r) \tag{18}$$

The computational burden, thus, lies on the first step; since (15) is differentiable, however, this minimization may be carried out very efficiently, as we now show.

2.3 Energy Minimization Algorithm

The minimization of (15) may be effected using any general purpose constrained optimization technique; we have found, however, that due to the simplicity of the constraints (17), and the structure provided by the Markovianity of the p field, a multiscale gradient projection Newtonian descent (GPND) [31, 32] gives best results. This method is based on the idea of moving, at each iteration, in a direction d such that $\nabla U \cdot d < 0$ (so that it is a descent direction), and that the new point lies in the feasible region. This is achieved by choosing d as the projection of the negative gradient onto the tangent subspace defined by the set of active constraints (see [32], pp 331-339). The convergence may be accelerated if one considers each element $p_k(r)$ (or $\theta_j(r)$) as the position of a particle of unit mass, subject to a force equal to $-\partial U / \partial p_k(r)$ (resp. $-\partial U / \partial \theta_j$). The equations of motion for these particles may be obtained from Newton's second law:

$$\ddot{\theta} = -\nabla_\theta U - 2\alpha \dot{\theta}$$

$$\ddot{p} = -\nabla_p U - 2\alpha \dot{p}$$

where α is the friction coefficient. The discretization of these equations gives an iterative gradient descent algorithm with inertia; to satisfy the constraints (17), each new

particle position $p_k(r)$ must be projected back into S_M, to get the complete iteration as:

$$\begin{aligned} \theta^{(t+h)} &= \frac{2}{\alpha h+1}\theta^{(t)} + \frac{\alpha h-1}{\alpha h+1}\theta^{(t-h)} - \frac{h^2}{\alpha h+1}\nabla_p U(p^{(t)},\theta^{(t)}) \\ \tilde{p} &= \frac{2}{\alpha h+1}p^{(t)} + \frac{\alpha h-1}{\alpha h+1}p^{(t-h)} - \frac{h^2}{\alpha h+1}\nabla_p U(p^{(t)},\theta^{(t)}) \\ p^{(t+h)}(r) &= \Pi_{S_M}(\tilde{p}(r)) \text{ , for all } r \in L \end{aligned} \tag{19}$$

where the operator $\Pi_{S_M}(u)$ finds the closest point in S_M to a vector $u \in \Re^M$. To find this projection, we consider the following observations: a simplex $S_{M,A}$, with K vertices, defined by

$$S_{M,A} = \{u \in \Re^M : \sum_{k=1}^{M} u_k = 1 \ ; \ u_k \geq 0, k \in A \ ; \ u_k = 0, k \notin A\}$$

where $A \subseteq Z_M$ is a set of indices and $|A| = K$, is contained in the hyperplane:

$$H_A = \{u \in \Re^M : \sum_{j \in A} u_j = 1 \text{ and } u_j = 0 \text{ , for } j \notin A\}$$

The orthogonal projection $x = \Pi_A(u)$ of a point $u \in \Re^M$ onto the hyperplane H_A satisfies $u_j - x_j = c$, where c is a constant, for $j \in A$, and $x_j = 0$, for $j \notin A$. The constant c may be found by noting that $\sum_{j\in A} x_j = \sum_{j\in A} u_j - cK = 1$, so that x may be found by the formula:

$$\begin{aligned} x_k &= u_k - \frac{\sum_{i\in A} u_i - 1}{K} \text{ , if } k \in A \\ &= 0 \text{ , if } k \notin A \end{aligned} \tag{20}$$

Now, if x is the closest point in $S_{M,A}$ to $\tilde{p}$, and x is in the interior of $S_{M,A}$, then x must be equal to $\Pi_A(\tilde{p})$. On the other hand, if x is not in the interior of $S_{M,A}$, then it must lie on a simplex $S_{M,A'}$, with $|A'| < |A|$ on the boundary of $S_{M,A}$. This active simplex corresponds precisely to $A' = \{k \in A : x_k \geq 0\}$, where $x = \Pi_A(\tilde{p})$. This observation suggests that the closest point in S_M to a given point $\tilde{p}$ may be found by recursively projecting $\tilde{p}$ into H_A, and then updating A if necessary, so that it corresponds to the active subsimplex in the boundary of S_M. This is done simply by excluding from A those indices that correspond to negative components of x. This gives the following algorithm for finding $x = \Pi_{S_M}(\tilde{p}(r))$:

1: set $x = \tilde{p}(r)$ and $A = Z_M$;

2: while $x \notin S_M$ do:

a: set $x = \Pi_A(\tilde{p})$ using (20) ;

b: set $A = \{k : x_k \geq 0\}$;

Note that this algorithm will converge at most in M iterations.

The minimization of (15) may be further accelerated using a multiscale approach: one may get descriptions of the observed image I at increasingly coarser scales $\{I = I_0, \ldots, I_K\}$, by recursively smoothing and subsampling it (the standard Gaussian pyramid [34]). At each scale k, one may then obtain a corresponding likelihood field $v^{(k)}$, replacing I by I_k in (7), and minimize the corresponding energy $U_k(p^{(k)}, \theta^{(k)})$. At scale K, this may be done efficiently, because of the reduced number of variables. Once the minimizers $p^{*(k)}, \theta^{*(k)}$ are found, they are transmitted as starting points for the minimization at scale $k-1$, until scale 0 (the original image) is reached. Care must be exercised when transmitting a solution $p^{*(k)}$ to a finer scale, since the interpolation process that is involved should guarantee that the interpolated p field is also in S_M^N. If one considers interpolation methods of the form:

$$p^{(k-1)}(r) = \sum_{s \in N^k(r)} w_{rs} p^{*k}(s)$$

where $N^k(r)$ denotes the set of sites in the coarse grid k on which the interpolation for site r in grid $k-1$ is based, then the constraint $p^{(k-1)} \in S_M^N$ will hold if $p^{(k)} \in S_M^N$ and $\sum_{s \in N^k(r)} w_{rs} = 1$, for all r. This is the case, for example, of bilinear (or trilinear, in the case of 3-D data) interpolation, which is the method we use; note that the θ variables are not interpolated, but only transmitted. Finally, the λ parameter that controls the strength of the interacion between neighboring pixels, and hence, the granularity of the solution, must also be adjusted; since the inter–pixel distance is duplicated when going from a fine to a coarse scale, the corresponding λ parameter should be halved, so that $\lambda_K = 2^{-K}\lambda_0$. This multiscale optimization procedure is the one we use in all the experiments reported below.

3 Experimental Performance

In this section we use synthetic images to compare the experimental performance of the new approach presented here (labeled HMMF) with that of classical MRF models; we compare with MPM estimators, first, because they are known to perform better than MAP estimators, particularly for high noise levels [6], and second, because they are based on the estimation of the posterior marginals, which are the basis for EM procedures, whose performance is also interesting to compare. For the classical case, we use the generalized Ising model, and 3 methods for computing the posterior marginals for the f field: a stochastic MCMC algorithm (the Gibbs Sampler [3]); the Mean Field

(MF) approximation [27] and the Gaussian approximation reported in [30] (labeled GMMF).

In the first set of experiments, the task is to perform intensity–based segmentation from noisy data, when the regions corresponding to each class have known constant intensity (i.e., $\Phi(r, \theta_k) = \theta_k$, assumed known), and the purpose is to compare the robustness of each method with repect to noise. We assumed 5 classes, with the class distribution shown in Fig. 3, and with $\theta_k = k$, $k = 1, \ldots, 5$. The observed images are obtained by adding white Gaussian noise with increasing variance. As a performance measure, we choose the average number of segmentation errors. The results are summarized in Fig. 4. As one can see, for low noise levels, all methods give similar results; as the noise level increases, the performance of the MF and MCMC approximations break down (for $\sigma = 1.5$ and $\sigma = 2$, respectively), while GMMF and HMMF degrade more gracefully, with HMMF giving the best results. In all cases, the control parameters for each method were hand–adjusted to get the best possible performance.

We also compared the performances of HMMF and GMMF using the same 8–class segmentation problem presented in [30]. Suppose that one is given, for each pixel of a 256×256 region L, observations of 3 features (for example, RGB values), and one knows a priori that there are 8 possible classes (colors), whose (known) mean values $\mu_{m,k}, m = 1, ...3, k = 1, ...8$ correspond to the vertices of the unit cube. The true spatial class distribution $c(r)$ is assumed to be as shown in Fig. (5-a); the observations $(g_1(r), g_2(r), g_3(r), r \in L)$ are constructed using the model:

$$g_m(r) = \mu_{m,c(r)} + n_m(r)$$

where $n_m(r), m = 1, ...3$ are Gaussian random variables with zero mean and variance $\sigma = 4$ (which corresponds to a SNR$\approx$ 0.25). The normalized likelihood field is thus computed as:

$$\widehat{p}_r(k) = \frac{1}{Z} \exp\left[-\frac{1}{2\sigma^2} \sum_{m=1}^{3} (g_m(r) - \mu_{m,k})^2\right]$$

where Z is a normalizing constant. The maximum likelihood estimator (MLE) $\widehat{c}$ of the field c is defined as:

$$\widehat{c}(r) = \arg\max_k \widehat{p}_r(k)$$

The results are shown in Fig. 5-b and c. As one can see, the performance of the 2 methods is similar, which is consistent with the results of Fig. 4.

In a second set of experiments, we test the relative robustness of the EM algorithm and our proposed procedure, with respect to initialization. To do this, we take again a synthetic, piecewise constant image with 3 classes, but this time we assume that the intensities $(\theta_1, \theta_2, \theta_3)$ are not known. We then generate 20 random starting points, with

uniform distribution on the dynamic range of the observed image, and note whether the corresponding algorithm converged to a neighborhood (a ball of radius 0.1) of the true values of θ, in which case, the run was labeled as a "success". Fig 6 illustrates this procedure. We tested the EM algorithm, using MF and MCMC to compute the posterior marginals, and the direct HMMF method presented here. The results are summarized in table 1, which also includes the corresponding average processing times. As one can see, the EM algorithm, even with an accurate approximation for the marginals (obtained with MCMC) is quite sensitive to initialization, giving a maximum success rate of 60%. This rate falls down to 0 for high noise levels, if the MF approximation is used. HMMF, on the other hand, is much more robust (giving 100% success rate in both cases), and, since it does not need to iterate, alternating between E and M steps, achieves this at a fraction of the computational time (all times refer to a PC–based workstation running at 1.8GHz).

HMMF has an additional advantage: if the exact number of models is not known in advance, one may initialize the procedure with a relatively large number of models, and the superfluous models will be automatically eliminated, in the sense that if the parameter vectors for 2 models j, k become almost equal, the p distributions will exhibit only one dominant mode in the corresponding support region, corresponding to either one of these models.

A final word must be said about the setting of the control parameters for these methods. In all cases (i.e., for EM/MCMC, EM/MF and HMMF), there are 2 control parameters: one that corresponds to the noise variance and the regularization parameter that controls the granularity of the reconstructed regions. Ideally, these parameters should be estimated —or at least fine–tuned — from the data, and in principle, some of the procedures that have been proposed to do this in the classical case [12, 35, 2, 22, 27, 28, 9, 36] may be extended to the case of HMMF as well. The problem is that these procedures have, in general, a very high computational cost, which we are trying to avoid in this case. The development of efficient hyperparameter estimation methods for HMMF is thus an important open problem, which we are currently investigating; however, HMMF is not too sensitive to the precise setting of these parameters: Fig. 7 shows the level curves for the error surface (i.e., average number of segmentation errors) obtained by a systematic variation of the noise and granularity parameters (γ in Eq. (7) and λ in Eq. (15)), for the experiment of Fig. 4, for different values of the Signal to Noise Ratio (SNR), defined in this case as that average separation between adjacent class intensities divided by the noise Std. deviation. As one can see, there is a large region around the optimal setting where the error surface is very flat, and the overlap between these flat regions for different values of the SNR is also quite large. This means that it is possible to calibrate the method for a particular problem class, selecting "good" values for the control parameters for a test image that belongs to the class, and use these values for the whole class of problems, getting acceptable results.

This is the approach we follow for the applications described in the next section.

4 Applications

4.1 Segmentation of Brain Magnetic Resonance Images

Magnetic resonance (MR) images of the brain provide a means for imaging tissue at very high resolutions, and the assignment of each voxel to a specific class (i.e., White Matter (WM), Gray Matter (GM) or Cerebro Spinal Fluid (CSF)) is important for visualization (as in surgical planning); for solving inverse problems (e.g., in electric tomography); for relative volume quantification, which is important for the diagnosis and prognosis of certain illnesses, etc. The main difficulties found in the automatic segmentation of brain MR volumes are due to 2 reasons: one is the presence of noise in the data, which cause voxel-by-voxel classification methods to produce granular or fragmented regions that violate anatomical constraints [37, 38], and the other is that image intensities are, in general, non–constant for each tissue class, due to irregularities in the magnetic fields, varying magnetic properties of biological tissues, operating conditions of the MR equipment, etc. For these reasons, a precise segmentation method should include an appropriate model for spatial interactions —to control the spurious granularity due to noise — and also the simultaneous estimation of smoothly varying intensity models for each class. This makes this problem an ideal candidate for the probabilistic segmentation methods that we have described in the previous sections.

Bayesian estimation, with prior MRF models for the label field, combined with SM methods (such as EM) for the estimation of smooth intensity models, have in fact been used by a number of researchers [13, 14, 15, 16], with the problems and limitations discussed in section 2. In most of these works, the smooth intensity models Φ are assumed to be of the form:

$$\Phi(r, \theta, k) = \mu_k \beta(r, \theta)$$

where $(\mu_1, \ldots, \mu_M)$ are the mean intensities for each tissue class, and $\beta(r, \theta)$ is a multiplicative bias field that is supposed to affect all tissue classes in the same way. If one wants a model that depends linearly on the parameters, however, it is necessary to perform a logarithmic transformation on the image intensities, which alters the noise distribution in a complex way, so that the Gaussian assumption is no longer valid, and also alters the image histogram, making the separation more difficult. For these reasons, and also because this simple model does not take into account the spatial variation of magnetic properties of specific tissues, we prefer to follow [40] and use a

more flexible, spline–based model of the form:

$$\Phi(r, \theta, k) = \sum_{j=1}^{J} \theta_{kj} N_j(r) \tag{21}$$

where $\{N_j, j = 1, \ldots, J\}$ are quadratic tensor product B–spline basis functions [41], translated to a node of a regular sub-grid of the voxel lattice, which we call the *spline sub-grid*: $N_j(r) = B^2((r - n_j) \cdot d)$, where n_j denotes the coordinate vector (in voxels) of the j^{th} node of the spline sub-grid and $d = (1/\Delta_x, 1/\Delta_y, 1/\Delta_z)^T$ is a scaling vector, with $\Delta_x, \Delta_y, \Delta_z$ denoting the distance between neighboring nodes on the spline sub–grid for each direction. Since inter–slice intensity variations in MRI are usually larger than intra–slice ones, we use a value of 32 voxels for Δ_x and Δ_y, and of 1 voxel for Δ_z. $B^2(x, y, z)$ is given by:

$$B^2(x, y, z) = b^2(x) b^2(y) b^2(z)$$

with

$$\begin{aligned} b^2(x) &= \frac{1}{2}(1.5 - 2x^2) \ , \ |x| \in [0, 0.5] \\ &= \frac{1}{2}(x^2 - 3|x| + 2.25) \ , \ |x| \in [0.5, 1.5] \\ &= 0 \ , \ |x| > 1.5 \end{aligned}$$

To further control the rigidity of the models, we impose a "membrane" Gibbsian prior on θ, of the form:

$$P_\theta(\theta) = \frac{1}{Z_\theta} \exp[-\sum_{k=1}^{M} \sum_{<u,v>} \eta_{rs} (\theta_{ku} - \theta_{kv})^2] \tag{22}$$

where the second sum is taken over nearest neighbor pairs of nodes $< u, v >$ in the spline sub-grid. For η_{uv} we used 0.1 in the $x - y$ direction and 0.01 in the z direction.

To validate this application of the HMMF procedure, we use the Brainweb MRI simulator [42, 43], which allows one to generate high quality simulated MRI volumes from known (ground truth) anatomical models, for different levels of noise and spatial inhomogeneities. Fig. 8 shows a sample slice of the simulated MRI, the anatomical model, the HMMF segmentation and the reconstructed intensity $\Phi(r, \theta^*_{f^*(r)})$. Fig 9 shows a comparison between HMMF results and the best —to our knowledge — published results on the same data, namely, the procedure presented in [44], which uses an EM/MF approach. The performance index ξ used for the comparison is:

$$\xi_k = \frac{2V_{GPk}}{V_{Pk} + V_{Gk}}$$

where V_{GPk} denotes the total number of voxels that were correctly assigned to class k by a given procedure; V_{Pk} is the total (correct + incorrect) number of voxels assigned to class k by this procedure and V_{Gk} denotes the total number of voxels belonging to class k in the anatomical model (ground truth). Note that ξ_k is always between 0 and 1, with 1 corresponding to a perfect segmentation. As one can see, the performance of HMMF is practically insensitive to the presence of spatial inhomogeneities, indicating they are adequately modeled. The values for the control parameters were: $\lambda = 0.01$ in the $x - y$ direction; $\lambda = 0.001$ in the z direction and $\gamma = 1$. It is important to note that the same values were used for the complete set of experiments (i.e., for different values of the noise intensity and of the spatial inhomogeneities). The initial values for p and θ were: $p_k(r) = 1/M$, for all k, r, and θ selected in such a way that each model $\Phi(r, \theta_k)$ corresponded to a constant intensity, with these intensities corresponding to the minimimum, middle and maximum intensities of the MR volume. It is worth noting that the procedure presented in [44] is more complex, and includes the estimation of prior probabilities for each class and for each voxel from statistical studies, and the registration of the volume under study with a reference brain, in order to get an appropriate mapping of these probabilities. With our procedure, we get a competitive performance without the inclusion of these probabilities, which makes the method easier to generalize to other cases (i.e., segmentation of MRI of other organs).

We have also applied the HMMF procedure to real brain MR volumes; a sample of the results appears in Fig. 10. In all cases, the separation of the brain parenchyma from non–brain tissue was performed automatically using non–rigid registration with an atlas [40].

4.2 Motion Segmentation

We now present an application example where the parameter vector θ enters into the energy function in a highly non–linear way, and show that the HMMF method still gives very good results. This example is the segmentation of objects moving according to different velocity models from an image sequence. This is an important problem in computational vision [45][46][47]: useful descriptions of complex scenes are usually composed of several moving objects and simple parametric descriptions of their motions. What makes this problem difficult is that one has to find both the model parameters and the corresponding objects (i.e., the region where the model is applicable) at the same time. Although other approaches are possible (e.g., [48]), the most successful follow the Bayesian paradigm discussed in section 2.1 [49, 50, 51, 52, 53, 36]. In this case, the models Φ are vector–valued (since they represent velocities in 2-D), and the observation model is:

$$I_1(r) = I_2(r + \Phi(r, \theta_{f(r)})) + n(r) \tag{23}$$

where I_1, I_2 represent 2 succesive frames from the sequence, and $f(r)$ indicates the active model in pixel r, as before. Note that since θ enters as an argument of the intensity I_2, one would need to solve a highly non–linear optimization problem in the M step in the EM procedure, if this approach is used. To avoid this problem and lower the computational cost, often a globally smooth optic flow is pre–computed, and then segmented [53, 47], in which case the energy function in the M step becomes quadratic. This has a disadvantage, however; since a regularization method must be used to compute the optic flow, the final segmentation results are likely to lose small details and localization of the boundaries between regions. With the HMMF approach one may work directly with the image intensities, since the added non–linearity represents only a marginal increase in the computational complexity of the procedure. The motion models Φ that have been used in most cases, correspond either to pure translation (constant) or affine (planar) models. Constant models are easy to fit, but their application is restricted to simple cases; affine models, on the other hand, have some problems: ideally, one would like the support regions for each model to be spatially localized and compact, since they correspond to a moving object; an affine model supported in a particular region, however, may also have low fitting error in places that are far away, producing spurious granularity in the segmentations. Also, if in the course of the segmentation procedure the support region for an affine model includes 2 objects moving at different velocities, the slopes of the planes that represent each velocity component may become very high, making impossible for the procedure to recover and converge to the desired values. These problems are avoided if one uses the spline models (21) with a Gibbsian prior ("membrane splines"), as described above, for each component of the velocity, since these models extrapolate as constants outside their support region, and hence are less prone to produce spurious interactions with other regions, and are numerically more stable. Besides, they provide a way (via the η parameter in (22)) to control the rigidity of the model, and hence, the character of the reconstructed optic flow, which may go from piecewise constant (for high values of η) to piecewise smooth (for low values). The smooth patches approximate well affine flows. This is illustrated in the synthetic example of Fig. 11, where we show how a piecewise affine flow is reconstruced using piecewise constant ($\eta = 100$) and piecewise smooth ($\eta = 1$) models. Note that membrane spline models are more general than affine ones, and can accurately model more complex (e.g., projective) transformations that often occur in real image sequences.

The initial values for θ for the minimization of (15), are not very critical in the piecewise constant (pure translation) case (see section 3). We have found that it is not necessary in this case to precompute the optic flow; we took 15×15 pixel windows, randomly placed in image I_1, and found, for each window W_k, the parameter vector θ_k

as:

$$\theta_k = \arg\min_{\theta} \sum_{r \in W_k} (I_1(r) - \Phi(I_2(r+\theta))^2$$

which may be very efficiently done using the Gauss–Newton algorithm [54]. Once the minimum of (15) is found for the piecewise constant case, the piecewise smooth segmentation is initialized with constant membrane spline models which are set equal to the optimal translations. The performance of this scheme, using 2 frames of a real motion sequence, is illustrated in Fig. 12. For the piecewise constant case, we used 8 models, which were automatically reduced to 5 for the piecewise smooth case. Since the vertical component of the motion was very small in this case, we show only the horizontal component of the reconstructed flow in panels (e) and (f). The values for the control parameters were $\gamma = 0.0001$ and $\lambda = 1$ in both cases, and the distance between nodes of the membrane spline grid was 32 pixels.

4.3 Edge–Preserving Denoising

Experimentally, one finds that the optimal estimated field p^* has an interesting property: if the models for 2 spatially adjacent regions R_j, R_k are similar, in the sense that $D_{jk}(r) = |\Phi(r, \theta_j^*) - \Phi(r, \theta_k^*)|$ is small, for r close to the boundary between R_j and R_k, then the corresponding $p_j^*(r)$ and $p_k^*(r)$ will also be relatively close, while if $D_{jk}(r)$ is large, $p_j^*(r)$ and $p_k^*(r)$ will be very different, indicating a sharp transition between R_j and R_k. A consequence of this is that the mean estimated intensity $\hat{I}$, for an observed image I, which is given by:

$$\hat{I}(r) = \sum_{k=1}^{M} \Phi(r, \theta_k^*) p_k^*(r) \tag{24}$$

will in fact be a smoothed version of I, in which sharp intensity changes (i.e., edges) are preserved. This is illustrated in Fig. 13, where we show: a noise corrupted "Lena" image (panel a); the reconstruced intensity from the optimal segmentation (i.e., $\Phi(r, \theta^*_{f^*(r)})$, panel b); the mean intensity computed using (24) (panel c) and the value of p_k^* for a particular model (panel d); note, for example in Lena's hat, how the value of p_k^* falls sharply in the right edge of the hat, and smoothly in the transition to a similar model inside the hat. In this case we used 8 constant models, $\lambda = 0.3$ and $\gamma = 0.05$. The filtering time was 12 sec.

This application is included here to illustrate an interesting property of the mean estimated intensity; there are many published edge–preserving filtering methods, based on a wide variety of techniques, such as: wavelets; non–linear partial differential equations; robust statistics, etc. We do not include a comparison here, since to do this meaningfully, would fall beyond the scope of this paper, whose main purpose is to present the HMMF models and illustrate their versatility.

5 Discussion

We have presented a new energy–minimization method for image segmentation, in the case when the parameters of the models that describe the spatial variation of a given attribute within each segment are not known, and when it is necessary to include prior constraints for the spatial coherence of the supports for each model. This method is rigorously based on Bayesian estimation theory, and its key idea is to introduce a hidden Markov random measure field, so that the (also hidden) label field is generated by a 2–step stochastic procedure. The resulting posterior energy, given by (15), may be directly minimized with respect to p and θ, subject to the constraints $p(r) \in S_M$, instead of using costly 2–step iterative procedures, such as EM, and without having to use approximations, such as MF. For the minimization of this function, any non–linear constrained optimization method may be used. We have tried, for instance, a Quasi–Newton scheme, using a barrier function to handle the constraints [54]; the results we have gotten so far, however, are practically indistinguishable from the ones obtained with the simple gradient projection scheme reported here, and the computational cost is significantly higher.

We have presented examples that illustrate the performance of this method in a variety of situations: in intensity–based segmentation, using simple constant models, and also parametric models of high order (membrane splines), and in motion segmentation, where the model parameters enter in a highly non–linear fashion. In all these cases, one gets a consistently robust behavior, both with respect to noise, and with respect to initialization, with a reasonable computational cost. The enhanced performance of this method, may be due, in part, to the non–linear data term $-\sum_r \log(v(r) \cdot p(r))$, which is used instead of the classical $-\sum_r \log v(r)$ (which is quadratic for Gaussian noise): this term, in combination with the quadratic regularization term $\sum_{<r,s>} |p(r) - p(s)|^2$ permits the energy function (15) to strike a good balance between 2 opposing tendencies: on one hand, the data term pushes each distribution $p(r)$ towards low entropy configurations, in which one component $p_k(r)$ dominates, because the minimum of $-\log(v(r) \cdot p(r))$, subject to $p(r) \in S_M$, is attained by $p_j(r) = \delta(j-k)$, $j = 1, \ldots, M$ where $k = \arg\max_j v_j(r)$. On the other hand, the regularization term acts as a diffusion, and hence, tends to produce high entropy (uniform) configurations; this balance permits the solution to evolve from an initial uniform state to a final low entropy configuration at an appropriate rate, so that the model parameters θ can escape from local minima at the beginning, when the segmentation induced by p is "soft", and be optimally adjusted at the end, when each $p(r)$ is sharply peaked.

As in the classical case, the potentials W_C may be adjusted to include constraints that are relevant to particular applications. We have found that the simple quadratic potentials given by (11) are sufficient to enforce general spatial coherence prior assumptions, but variations of these potentials, and the inclusion of additional terms

should improve the results in specific cases.

We presented 3 examples of applications, to show the versatility of the models presented here. We showed how HMMF prior models, together with membrane splines, may be used to construct procedures for MRI and motion segmentation, as well as for edge–preserving filtering. It should be clear, however, that these examples were introduced only for illustrative purposes, since the complete solution of any of these problems, and a meaningful validation of the corresponding procedures, involves many subtle points that fall beyond the scope of this presentation.

The main limitation of the procedures presented here is the determination of appropriate values for the control parameters, which at present must be done "by hand". This is, of course, undesirable, even when the performance is relatively robust with repect to their precise setting. We think, however, that since this methodology is rigorously based on Bayesian estimation theory, it should be possible to devise strategies for their automatic determination, as has been done in certain cases for classical MRF's. The design of computationally efficient procedures for doing this is, in our view, the main open problem of the field.

Acknowledgements. J.L. Marroquin and S. Botello were supported in part by Conacyt, Mexico, under grant 34575-A.

References

[1] J. Besag., Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36 (2):192-236, 1974

[2] J. Besag, On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48 (3):259-302, 1974

[3] D. Geman and S. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 6(6):721-741, 1984.

[4] T. Simchony, R. Chellappa and Z. Lichtenstein, Relaxation algorithms for MAP estimation gray-level images with multiplicative noise. *IEEE Trans. on Information Theory*, 36(3): 608-613, May 1990.

[5] S.Z. Li,Invariant surface segmentation through energy minimization with discontinuities, *Int. J. of Comp. Vis.*, 5,2:161-194, 1990.

[6] J.L. Marroquin, S. Mitter and T. Poggio, Probabilistic solution of ill-posed problems in computational vision *Journal of the American Statistical Association*82(397):76-89, 1987.

[7] C. Bouman and. M Shapiro, A multiscale random field model for Bayesian segmentation. *IEEE Transactions on Image Processing* 3(2) 162-177, 1994.

[8] M.L. Comer and E.J. Delp, Parameter estimation and segmentation of noisy or textured images using the EM algorithm and MPM estimation. *Proc. of IEEE Int. Conf. on Image Proc., Vol. II*, 650-654, Austin Texas, November 1994.

[9] J. Zhang, J.W. Modestino and D.A. Langan Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segnmentation. *IEEE Trans. on Image Processing*, 3(4):404-420, July 1994.

[10] H. Derin and W. Cole, Segmentation of textured images, using Gibbs random fields. *Comput. Vision Graphics and Image Process.*, 32:72-98, 1986.

[11] F. Cohen and D. Cooper, Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-9 (2): 195-219, March 1987.

[12] S. Lakshmanan and H. Derin, Simultaneous parameter estimation and segmentation of Gibbs Random fields using simulated annealing. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 11:799-813, August 1989.

[13] J.C. Rajapakse and F. Krugge, Segmentation of MR images with intensity inhomogeneities. *Image and Vis. Comp.*, 16:165-180, 1998.

[14] K. Held, E.R. Kopsa, B.J. Krause, W.M. Wells, R. Kikinis and H.W. Muller-Gartner, Markov random field segmentation of brain MR images. *IEEE Trans. on Med. Im.*, 16,6:878-886, 1997.

[15] W.M. Wells III, W.E.L. Grimson, R. Kikinis and F.A. Jolesz, Adaptive segmentation of MRI data, *IEEE Trans. on Med. Im.* 15:429-442, Aug. 1996.

[16] Y. Zhang, M. Brady and S. Smith, Segmentation of brain MR images through a hidden markov random field model and the Expectation–Maximization algorithm, *IEEE Trans. on Med. Im.*, 20,1:45-57, Jan. 2001.

[17] J. Zerubia and R. Chellappa, Mean Field annealing using compound Gauss-Markov random fields for edge detection and image estimation. *IEEE Trans. on Neural Networks*, 4(4):703-709, July 1993.

[18] D. Geman, S. Geman, C. Graffigne and P. Dong, Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 12(7):609-628, 1990.

[19] D. Geiger and F. Girosi, Parallel and deterministic algorithms from MRF?s: surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 13(5), 401-412 1991.

[20] D. Geman and G. Reynolds, Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 14(3):767-783, 1992.

[21] P. Green, Bayesian reconstructions form emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging*. 9(1),84-93, 1990.

[22] S. Geman and D McClure, Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst.*, LII-4:5-21, 1987.

[23] S.Z. Li, *Markov random field modeling in image analysis*, Springer, Tokyo, 2001.

[24] R. Chellappa and A. Jain (Eds.) *Markov Random Fields: Theory and Applications. Academic Press.* 1993.

[25] A.P. Dempster N.M. Larid and D.B. Rubin, Maximum likelihood form incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1-38, 1977.

[26] N. Metropolis, A. Rosenbluth, M. Rosenbluth A. Teller and E. Teller, Equations of state calculations by fast computational machine. *Journal of chemical Physics* 21:1087-1092.

[27] J. Zhang, The mean field theory in EM procedures for Markov random fields. *IEEE transactions on Image processing* 40(10): 2570-2583, 1992.

[28] J. Zhang, The mean field theory in EM procedures for blind Markov random field image restoration. *IEEE transactions on Image processing* 2(1): 27-40, 1993.

[29] A.L. Yullie, Generalized deformable models, statistical physics and maching problems. *Neural Computation* 6:341-356, 1994.

[30] J.L. Marroquin, F. Velasco, M. Rivera and M. Nakamura, Gauss-Markov Measure Field Models for Low-Level Vision. *IEEE Trans. on PAMI.* 23,4, p. 337-348 (2001).

[31] J.L. Marroquin, Deterministic Interactive Particle Models for Image Processing and Comp. Graph., *Graph. Mod. and Im. Proc.*, 55, 5: 408-417, (1993).

[32] D.G. Luenberger, *Linear and Nonlinear Programming*, Addison Wesley, Reading, Mass. , 1989.

[33] J.L. Marroquin, E. Arce and S. Botello, Hidden Markov Measure Field Models for Image Segmentation. Centro de Investigacion en Matematicas, Technical Report 1-02-05(CC), Guanajuato, Gto., Mexico, 2002. Available in: http://www.cimat.mx/biblioteca/RepTec/

[34] P.J. Burt, The pyramid as a structure for efficient computation. In *Multiresolution image processing and analysis*, edited by A. Rosenfeld, Springer Series in Information Sciences, Vol. 12, Springe, New York, 1984.

[35] A. Jalobeanu, L. Blanc-Feraud and J. Zerubia, Hyperparameter estimation for satellite image restoration using a MCMC maximum–likelihood method. *Pat. Recog.*, 35,2:341-352, Feb. 2002.

[36] N. Vasconcelos and A. Lippman, Empirical Bayes motion segmentation. *IEEE Trans. on PAMI*, 23,2:217,221, Feb. 2001.

[37] H.E. Cline, C.L. Doumulin, H.R. Hart, W.E. Lorensen and S. Ludke, 3-D reconstruction of the brain from magnetic resonance images using a connectivity algorithm. it Magnetic Res. Imag., 5:345-352, 1987.

[38] M. Joliot and B.M. Majoyer, Three dimensional segmentation and interpolation of magnetic resonance brain images. *IEEE Trans. on Med. Im.*, 12,2:269-277, 1993.

[39] M. Styner, C. Brechbüler, G. Szekely and G. Gerig, Parametric estimate of intensity inhomogeneities applied to MRI. *IEEE Transactions on Medical Imaging* 19(3),153-165, 2000.

[40] 2. J.L. Marrroquin, B.C. Vemuri, S. Botello, F. Calderon and A. Fernandez-Bouzas, An accurate and efficient Bayesian method for automatic segmentation of brain MRI. *IEEE Trans. on Med. Imag.* 21,8:934-945 2002.

[41] I.J. Schoenberg, *Cardinal spline interpolation*, Philadelphia, PA. Society for Industrial and Applied Mathematics, 1973.

[42] C.A.Cocosco, V. Kollokian, R.K. Kwan, A.C. Evans, BrainWeb: Online Interface to a 3D MRI Simulated Brain Database. NeuroImage, 5, 4, part2/4, S425, 1997.

[43] D.L. Collins, A.P. Zidjenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes and A.C. Evans, Design and construction of a realistic digital brain phantom. *IEEE Trans. Med. Im.* 17, 3:463-468, 1998.

[44] K.V. Leemput, F. Maes, D. Vandermuelen and P. Suetens, Automated model-based tissue classification of MR images of the brain. *IEEE Trans. on Med. Imag.*, 18 (10):897-908, 1999.

[45] Hsu S. , Anandan P. and Peleg S., 1994, Accurate computation of optical flow by using layered motion representation. In 12th International Conference on Pattern Recognition, A:743-746.

[46] Jepson A. and Black M. J., 1993. Mixture models for optical flow computatiom. In Proc. IEEE. Conf. Computer Vision and Pattern Recognition: 760-761.

[47] Wang J.Y.A. and Adelson E. H., 1997, Representing moving images with layers. IEEE Transactions on Image Processing Special Issue: Image Sequence Compression: 520-526.

[48] J. Shi and J. Malik, Motion Segmentation and Tracking Using Normalized Cuts. *International Conference on Computer vision ICCV98, 1998.*

[49] J. Zhang and G.G. Hanauer, The application of mean field theory to image motion estimation.*IEEE Trans. on Image Processing*, 4(1)19-33, January 1995.

[50] F. Heitz and P. Bouthemy, Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Transactions on Pattern Analysis and machine Intelligence* 15(12):1217-1232, 1993.

[51] J. Konrad and E. Dubois, Bayesian estimation of motion vector fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, 14(9):910-927, September 1992.

[52] M.R. Luettgen, W.C. Karl and A.S. Willsky, Efficient multiscale regularization with applications of Markov random fields. *IEEE Trans. on Image Processing*, 3(1), January 1994.

[53] Y. Weiss and E. H. Adelson. A unified mixture framework for motion segmentation: Incorporing spacial coherence and estimating the number of models. In *Proc. IEEE CVPR96*, pp 321-326. 1996.

[54] J. Nocedal and S.J. Wright, *Numerical optimization*, Springer, New York, 1999.

Table 1 Success rates and average processing times for different estimation procedures for the experiment explained in the text.

	Success rate $\sigma = 1$	Success rate $\sigma = 1.5$	Avg. Time (sec)
EM/MCMC	60 %	60 %	800
EM/Mean Field	60 %	0 %	200
HMMF	100 %	100 %	5

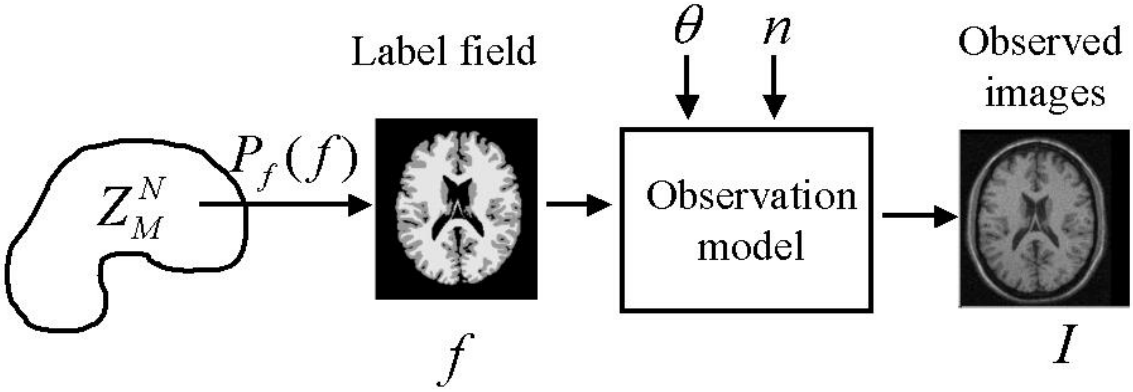

Figure 1: Classical MRF model for image segmentation.

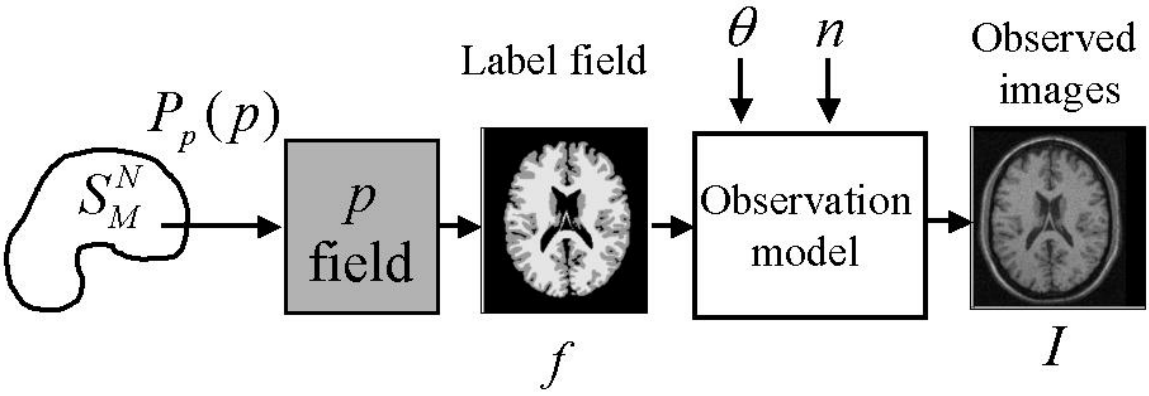

Figure 2: Hidden Measure Field Model for image segmentation.

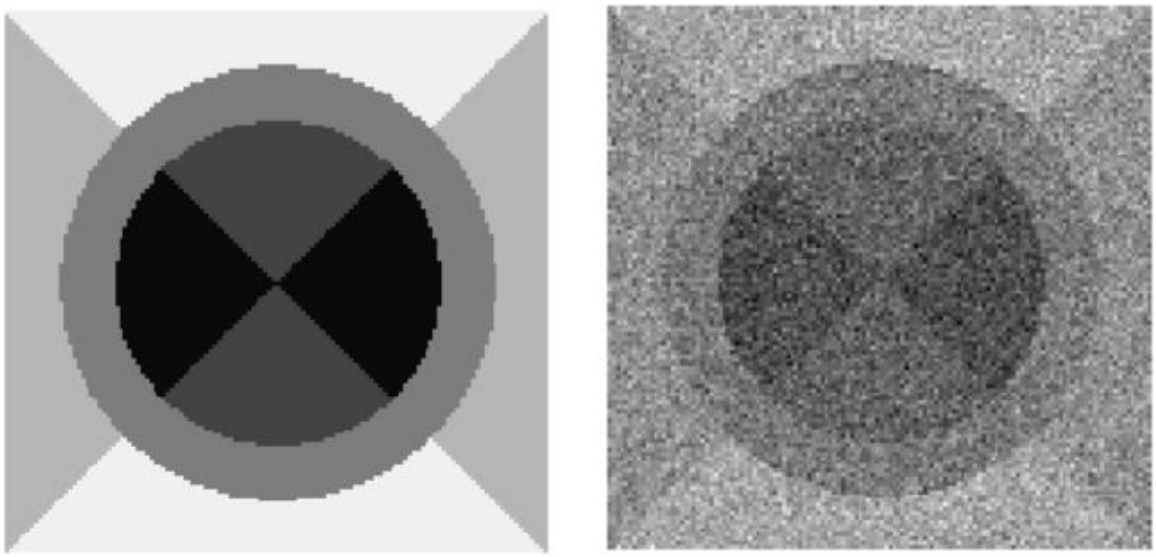

Figure 3: Left: class distribution; right: observed image for noise std. dev. = 1.5, for the set of experiments described in the text.

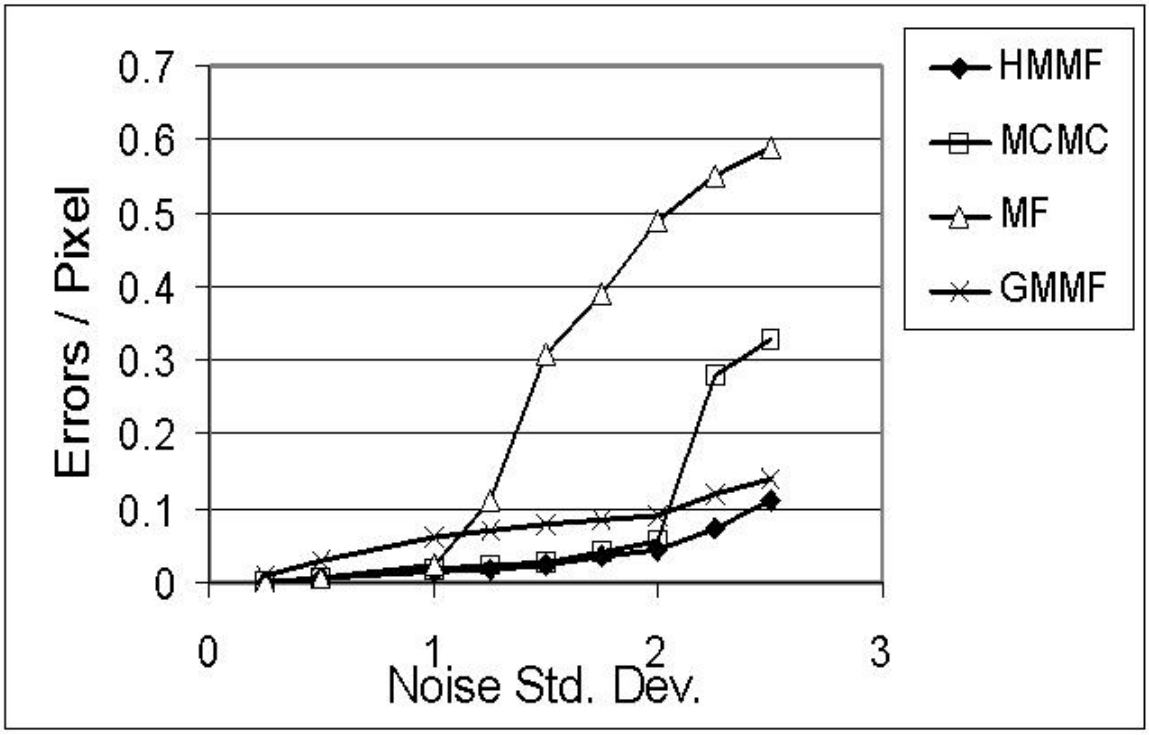

Figure 4: Comparative performance for 4 Bayesian estimators, for different noise levels (see text).

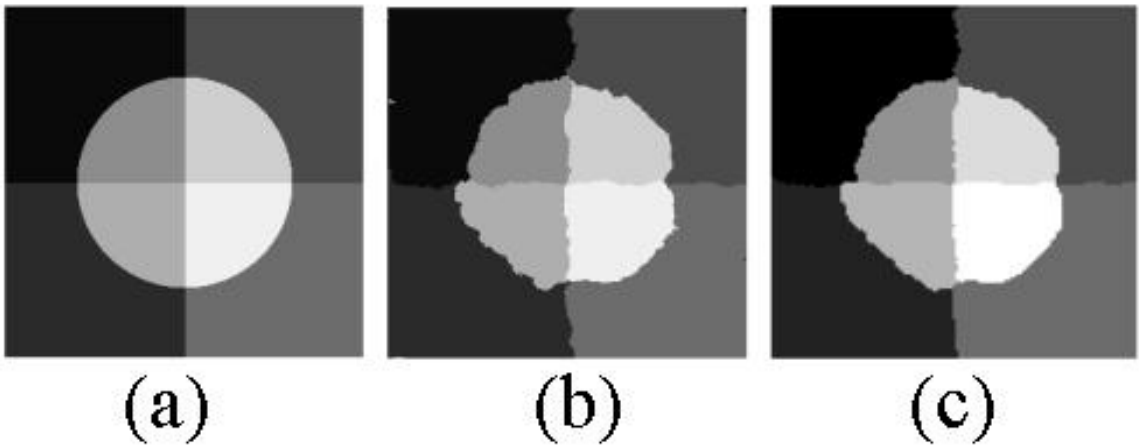

Figure 5: (a) Class distribution for the classification problem discussed in the text. (b) GMMF reconstruction. (c) HMMF reconstruction.

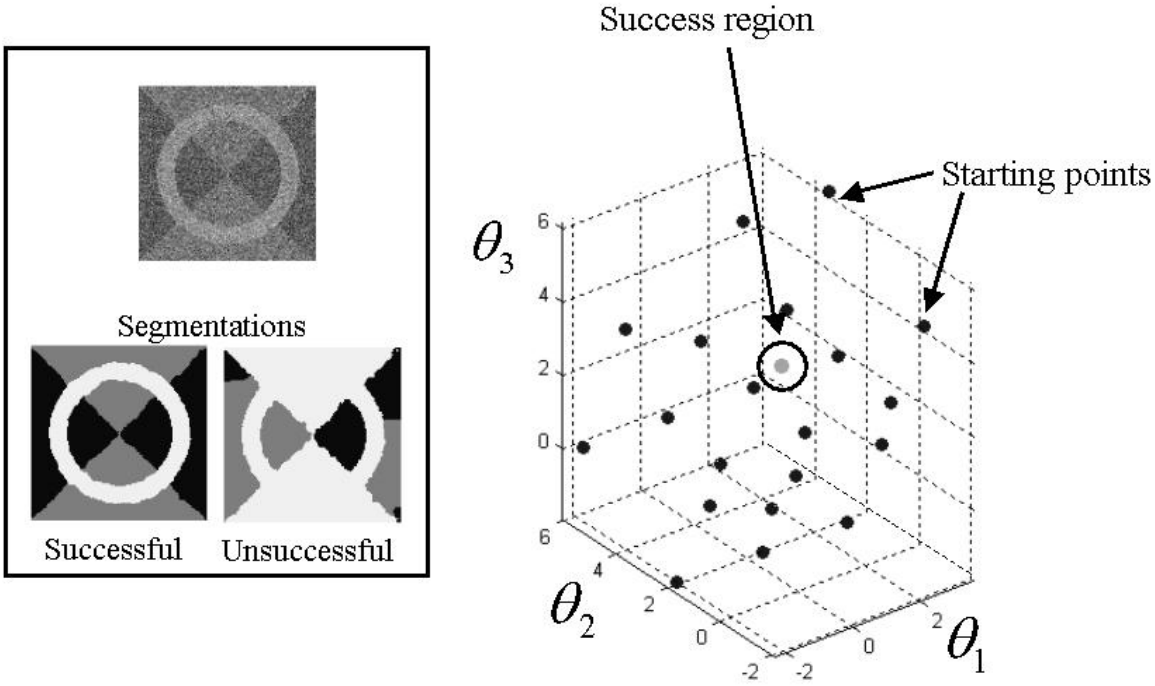

Figure 6: Left: observed image and typical successful and unsuccessful segmentations. Right: 20 random initial values for the parameters.

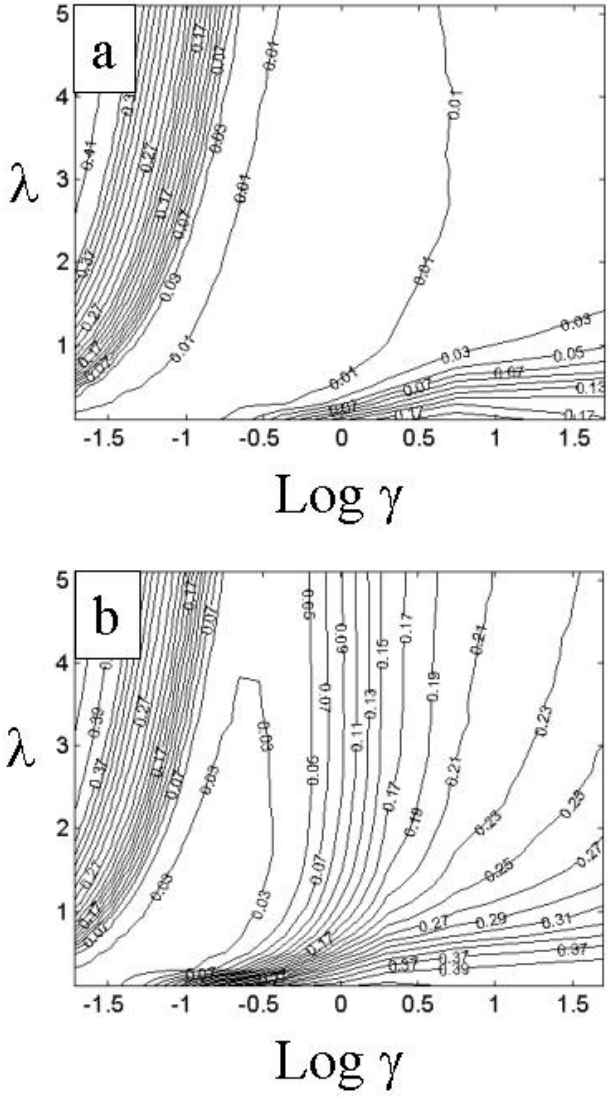

Figure 7: Level curves for the error surface (average number of segmentation errors) plotted against the values for the control parameters for 2 noise levels: (a) SNR = 1, (b) SNR = 2.

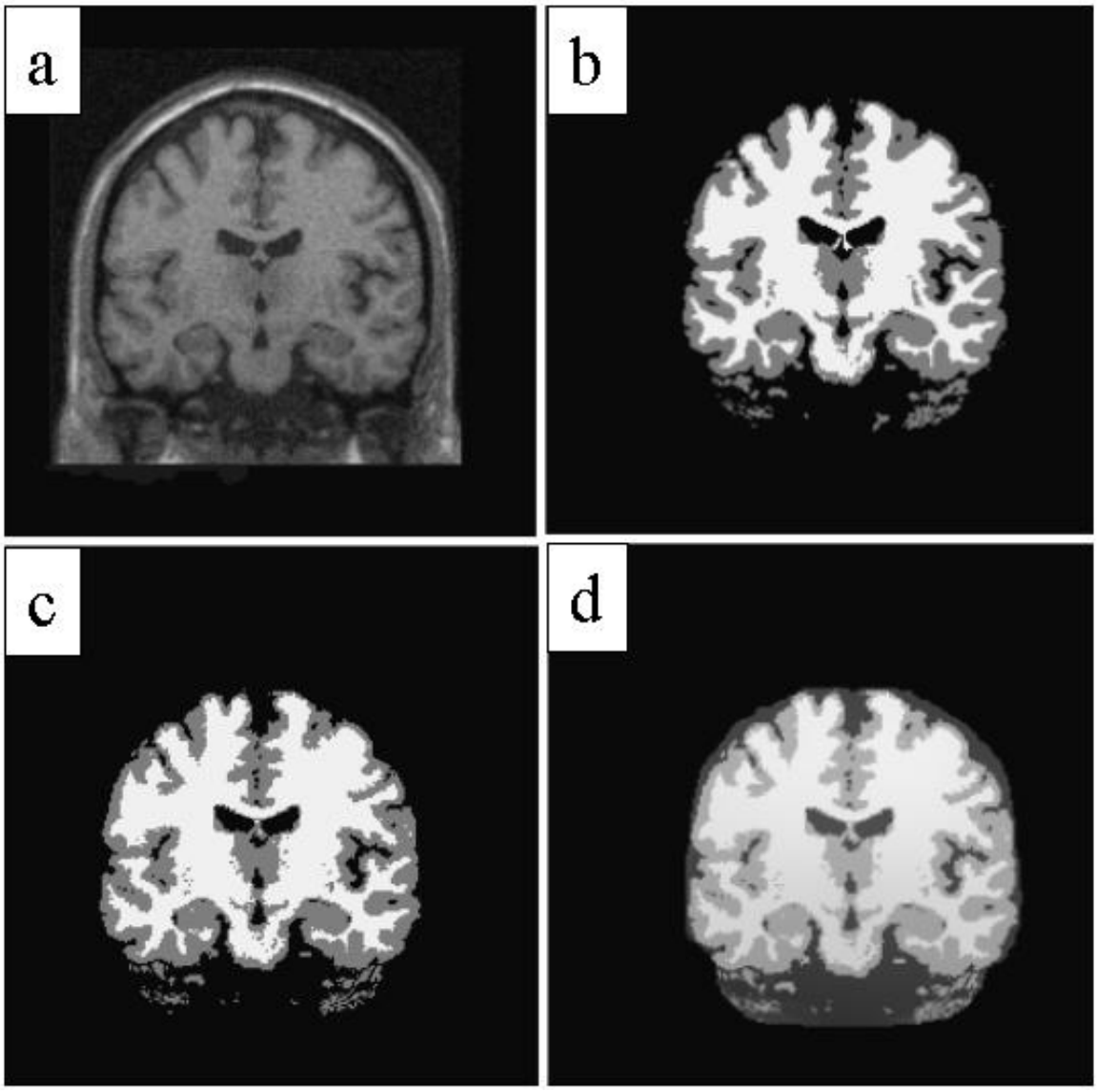

Figure 8: (a) Sample slice of a simulated brain MR volume. (b) Anatomical model (ground truth). (c) HMMF segmentation. (d) Reconstructed intensity $\Phi(r, \theta^*_{f^*(r)})$.

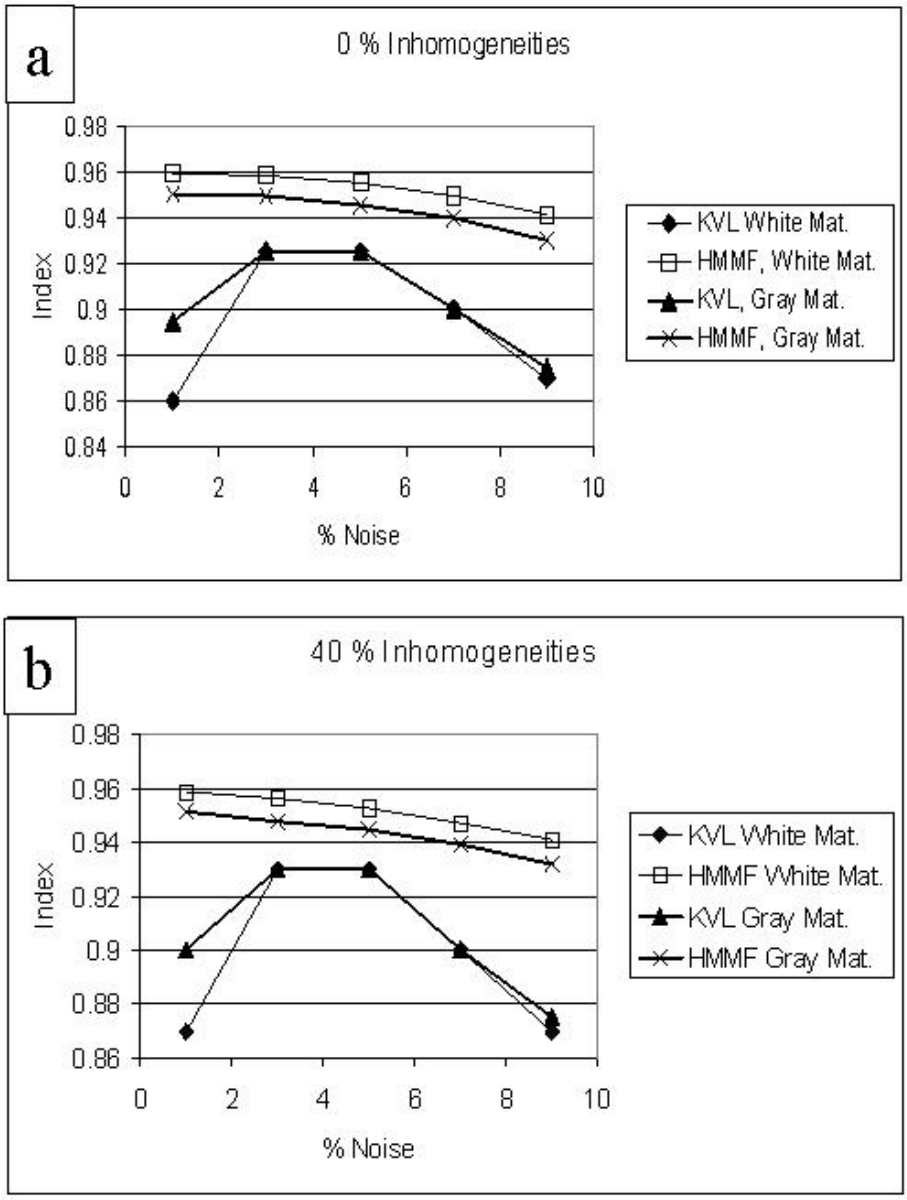

Figure 9: Performance index for different noise levels for the HMMF segmentation and the one reported in [44] (labeled KVL) for a simulated MR volume, for (a) 0 % and (b) 40% spatial inhomogeneities.

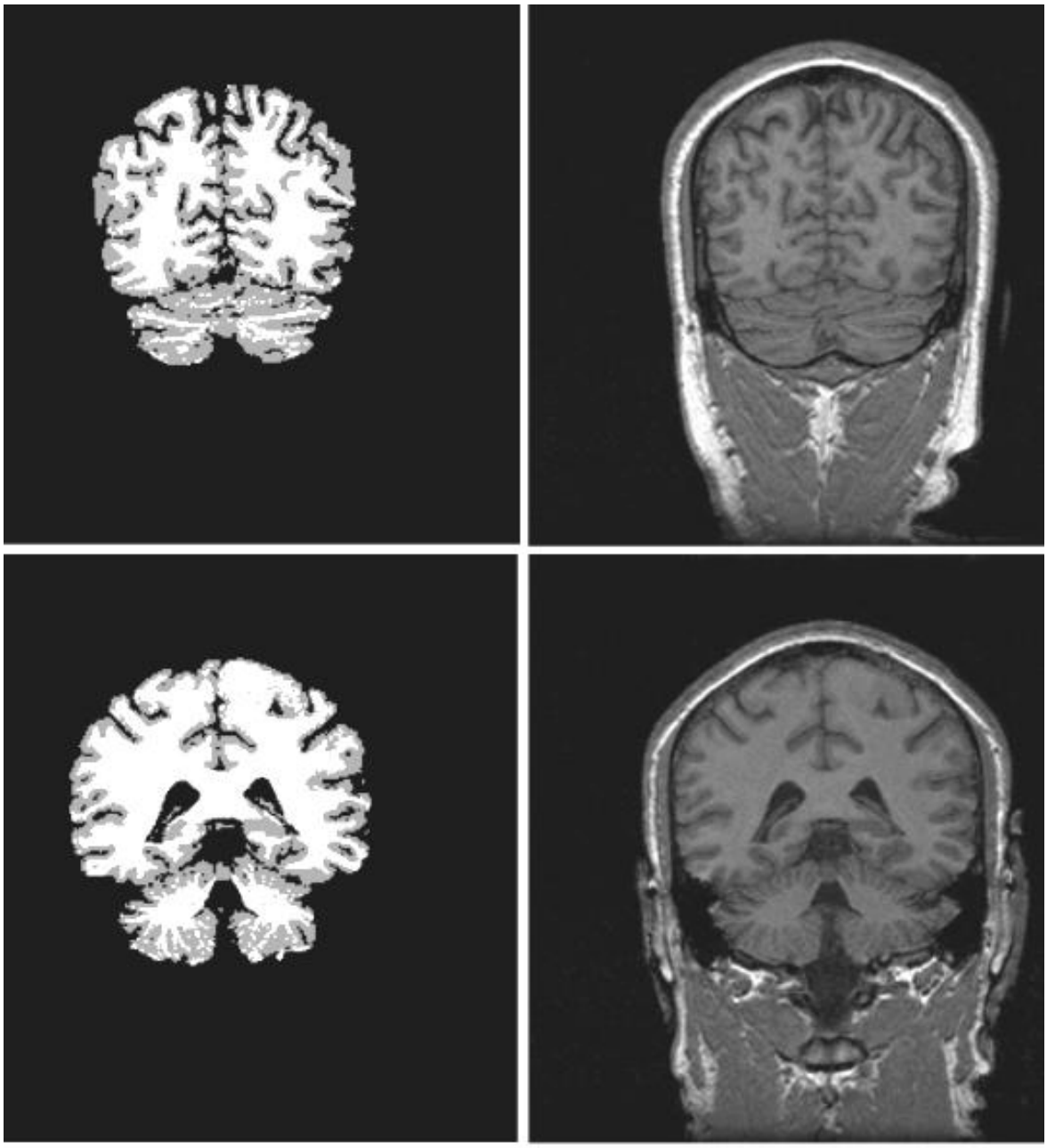

Figure 10: HMMF segmentation (left column) for 2 sample slices of a real brain MR volume (right column).

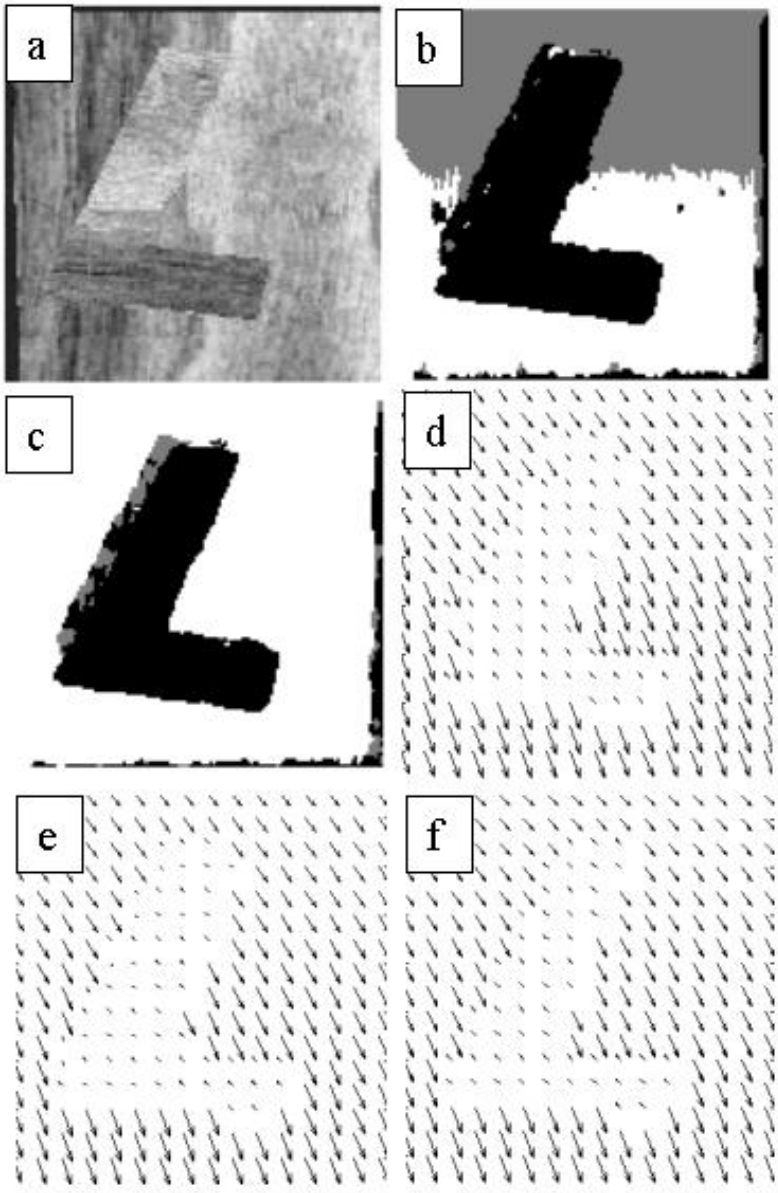

Figure 11: (a) One frame of a synthetic sequence. (b) and (c) Piecewise constant and piecewise smooth segmentations. (d) and (e) Piecewise constant and piecewise smooth reconstructed flows. (f) True piecewise affine flow.

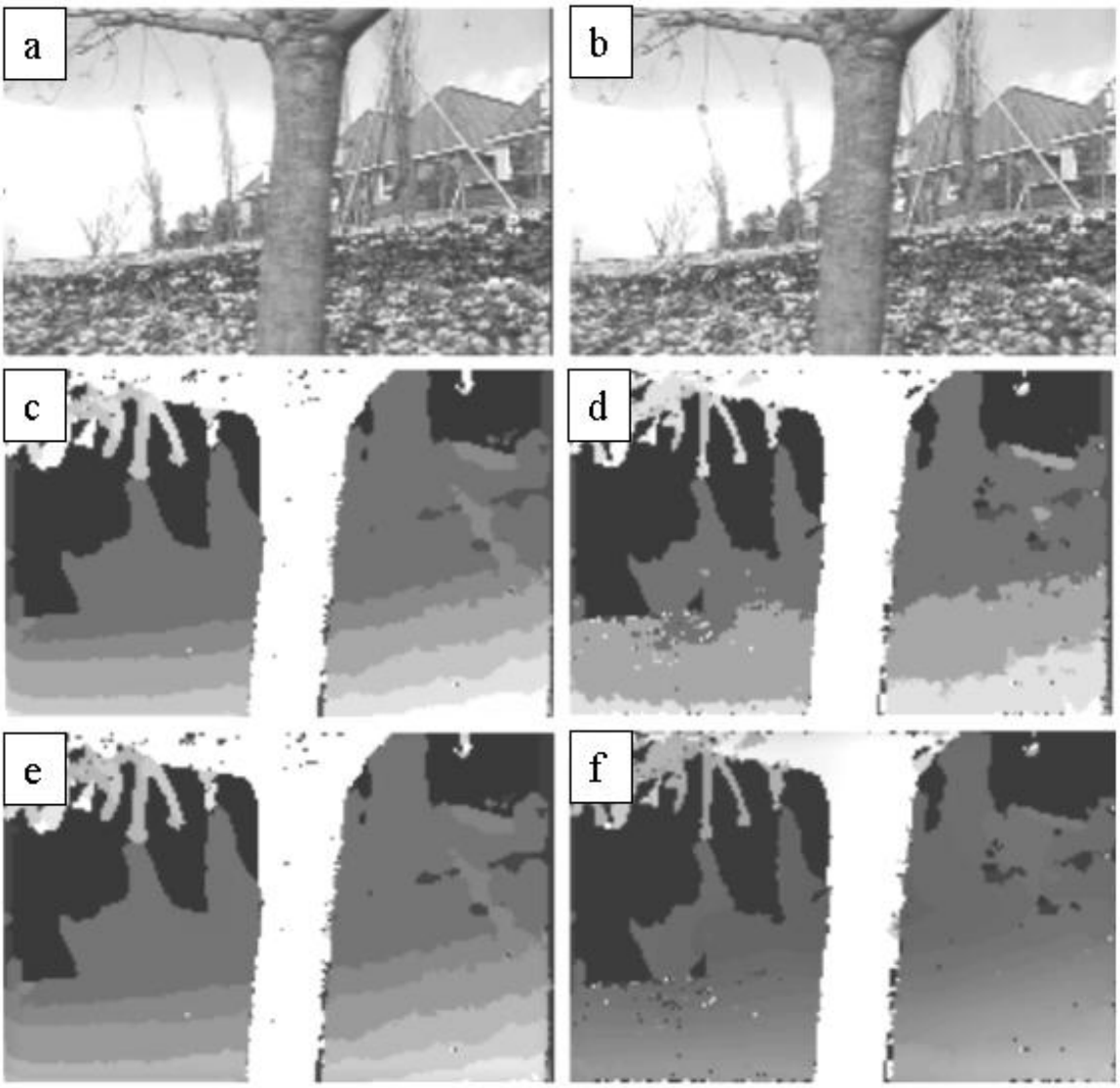

Figure 12: (a) and (b) 2 frames of a real motion sequence. (c) and (d) Piecewise constant and piecewise smooth segmentations. (e) and (f) Magnitude of the horizontal component of the velocity for the piecewise constant and piecewise smooth cases.

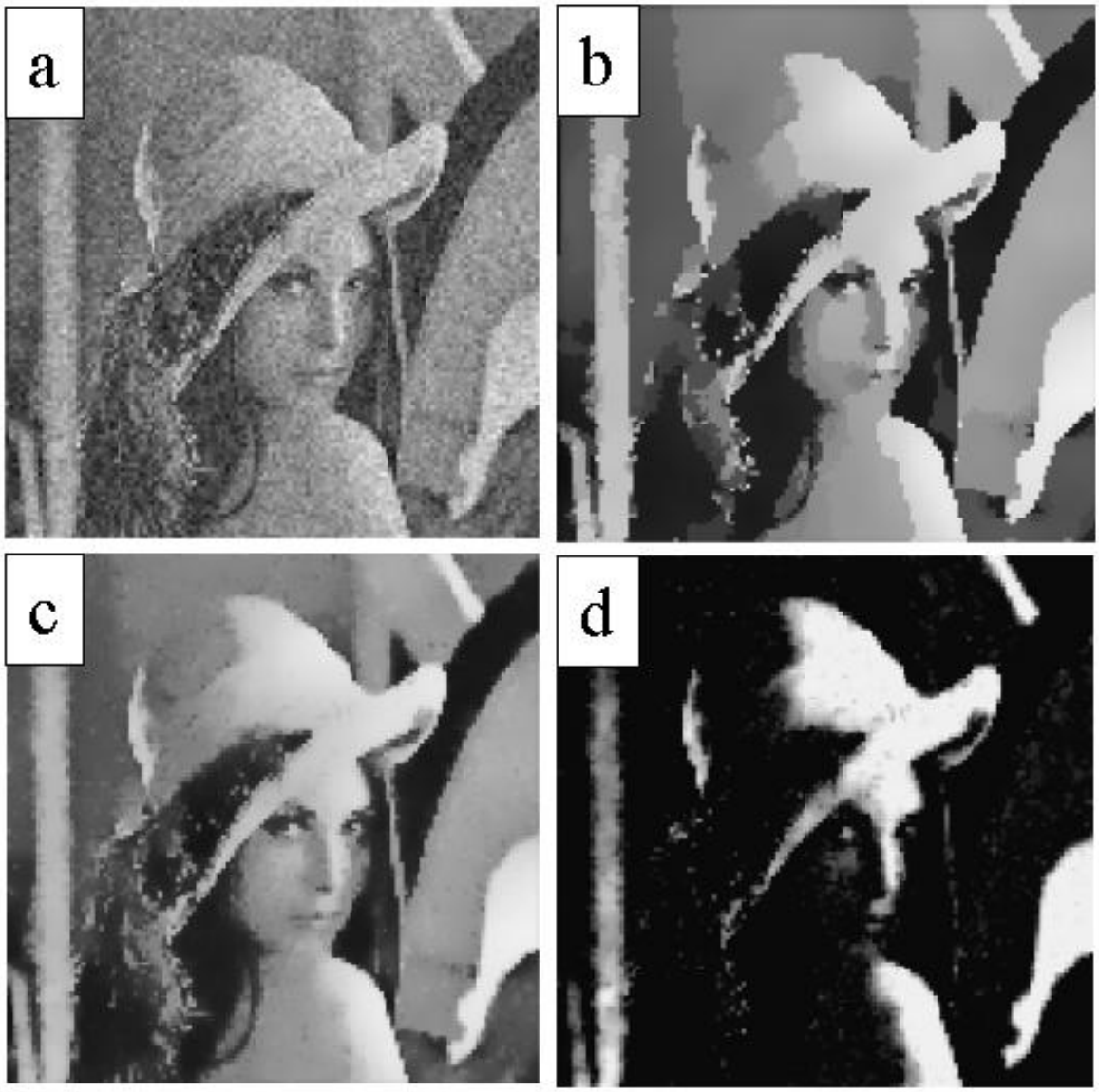

Figure 13: (a) Lena image corrupted with additive Gaussian noise. (b) Reconstructed intensity $\Phi(r, \theta^*_{f^*(r)})$. (c) Mean reconstructed intensity. (d) p^* field for one of the models