Efficient MCMC Sampling with Implicit Shape Representations

Jason Chang
Massachusetts Institute of Technology
32 Vassar St. Cambridge, MA
jchang7@csail.mit.edu

John W. Fisher III
Massachusetts Institute of Technology
32 Vassar St. Cambridge, MA
fisher@csail.mit.edu

Abstract

We present a method for sampling from the posterior distribution of implicitly defined segmentations conditioned on the observed image. Segmentation is often formulated as an energy minimization or statistical inference problem in which either the optimal or most probable configuration is the goal. Exponentiating the negative energy functional provides a Bayesian interpretation in which the solutions are equivalent. Sampling methods enable evaluation of distribution properties that characterize the solution space via the computation of marginal event probabilities. We develop a Metropolis-Hastings sampling algorithm over level-sets which improves upon previous methods by allowing for topological changes while simultaneously decreasing computational times by orders of magnitude. An M-ary extension to the method is provided.

1. Introduction

Level set representations and Markov chain Monte Carlo (MCMC) sampling methods are useful in a wide variety of applications. Level set representations eschew explicit curve and surface parameterizations while allowing topological changes with superior numerical stability [17]. MCMC methods enable one to reason about complex distributions for which exact analysis is intractable [6, 7] and additionally provide a more extensive characterization of energy minimization formulations when viewed from a Bayesian perspective. Integrating the two formalisms faces two distinct challenges. First, the high dimensionality of implicit representations induces a large configuration space resulting in slow convergence for naive implementations. Second, certain technical conditions induce a correspondence problem that, in prior efforts, has overly constrained the applicable class of curves (e.g. simply connected shapes). Here, we address these and additional issues resulting in a *computationally tractable* MCMC sampling algorithm over the space of implicitly defined shapes. This, in turn, simplifies the estimation of marginal statistics defined over the distribution of implicitly defined curves $\mathcal{C}$ for a given image I. While many level set methods are formulated as an energy minimization over some functional $E(\mathcal{C}; I)$, it is often the case that, either due to the ill-posedness of unsupervised segmentation or the stochastic nature of a well posed formulation, multiple plausible explanations exist. In either case, characterization of the posterior distribution is desirable; e.g., marginal statistics over the distribution may offer a more informative characterization than the optimal configuration. Consequently, a common alternative is to recast the optimization formulation as one of Bayesian inference by viewing the energy functional as the negative log of a probability density

$$p(\mathcal{C}|I) \propto \exp\left(-E(\mathcal{C}; I)\right). \quad (1)$$

Depending on the form of Equation 1, when explicit characterization and/or direct sampling is intractable, one may utilize (under certain technical conditions) the Metropolis-Hastings algorithm [11] to both sample from the distribution and evaluate marginal statistics.

By way of example, consider the synthetic and natural images in the first column of Figure 1. For both images, we include the associated probability of boundary (PB) image, a thresholded PB image, and the minimal energy segmentation obtained via the MCMC sampler described in the sequel. While neither the thresholded PB image nor the optimal segmentation is error-free, the optimal configuration produces a larger number of false positives. This a well known phenomenon [6] in the MCMC literature.

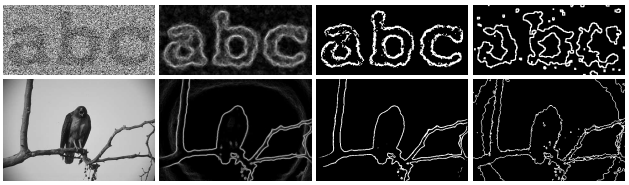

Figure 1: Examples of MCMC sampling on a synthetic and natural image. Qualitatively, thresholding the probability of boundary (PB) is superior to the optimal segmentation.

Here, we emphasize that our primary contribution is to develop a *computationally tractable* Metropolis-Hastings sampling algorithm by which energy based, level-set formulations may be analyzed within a Bayesian framework. As with many MCMC samplers, the proposal distribution has a critical impact on the length of the *mixing time*, i.e. convergence to the stationary distribution from which we would like to sample. In addition to relaxing constraints on the allowed shape class (as compared to previous methods [4, 5]), we suggest a design method for the proposal distribution that dramatically reduces the mixing time. In summary, the contributions of this work are threefold. First, we develop an MCMC sampling method for implicit shape representations that includes topological changes. Second, we extend the approach to the case of M-ary segmentations. Third, we achieve these improvements while simultaneously accelerating the sampling procedure by *orders* of magnitude over previous methods. While we utilize explicit formulations in order to demonstrate the method, the method itself is quite general and can be used for almost any static image feature and region based energy functional.

2. Related Work

Sampling from the space of implicit segmentations has been suggested previously. Fan et al. [5] develop a hybrid method: alternating between implicit (level set) and explicit (marker based) representations of a simply connected shape. The proposal distribution generates a sample perturbation over a set of marker points which, when computing what is known as the Hastings ratio, induces a correspondence problem over the explicit representation. Upon completion, the new sample is converted into an implicit form by resolving the Eikonal equation. While establishing the feasibility of applying MCMC methods to implicit representations, [5] is constrained to binary segmentations of a single, simply connected shape. Furthermore, iterations between implicit and explicit representations incur a substantial computational burden. Fan suggests the use of jump diffusion processes [10] as a means of incorporating topological changes. However, no formulation satisfying detailed balance (see Section 3.2) is provided.

Chen et al. [4] improve upon the method of Fan et al. by obviating the need to transition between implicit and explicit representations. They construct a smooth normal perturbation at a *single* point on the curve (denoted the "foot point") that preserves the signed distance property between proposal samples, thereby simplifying the correspondence problem and evaluation of the Hastings ratio. However, the resulting perturbations are extremely smooth, and as such, explore the configuration space very slowly. As in [5] obstacles remain for incorporating topological changes, restricting this method to binary segmentations with a single simply connected shape.

3. Metropolis-Hastings Sampling over Curves

The energy functional $E(\mathcal{C}; I)$ is a surrogate for evaluating what constitutes a good segmentation. In both parametric and nonparametric settings, this term can often be decomposed into a data-fidelity term (i.e. a likelihood) and a penalty/regularization term (i.e. a prior). While we discuss unsupervised nonparametric methods (whose data-fidelity terms have a natural information-theoretic interpretation), the underlying framework applies to both parametric and nonparametric models. For example, [14] uses the mutual information (MI) between the pixel intensities and labels, L, as the data-fidelity term combined with a curve length penalty as the prior. In the supplemental material we provide conditions under which this energy functional is equivalent to a posterior distribution. Other commonly used energy functionals include KL divergence [12], J divergence [13], and Bhattacharyya distance [16]. In the following section, we develop our sampling framework with a general energy functional, E, such that any of these information-theoretic measures can be applied. Following prior approaches, we compute distribution estimates via a nonparametric kernel density estimate (KDE) [18]. Fast methods for computing a KDE can be found in [9].

While MCMC sampling in finite dimensional spaces has been well studied, the same cannot be said with respect to sampling from the infinite dimensional space of shapes. One can construct a Metropolis-Hastings sampler [11] as follows. Let $\hat{\varphi}^{(t+1)}$ be a proposed sample of the implicit representation (i.e. the level-set function) generated from a distribution $q(\hat{\varphi}^{(t+1)}|\varphi^{(t)})$ conditioned on the current sample, $\varphi^{(t)}$. The superscript values (t) and $(t+1)$ index the sampling iteration and the hat indicates a proposed sample. This new sample is then accepted with probability

$$\begin{aligned}&\Pr\left[\varphi^{(t+1)} = \hat{\varphi}^{(t+1)} \,\middle|\, \varphi^{(t)}\right] \\ &= \min\left[\overbrace{\underbrace{\frac{\pi\left(\hat{\varphi}^{(t+1)}\right)}{\pi\left(\varphi^{(t)}\right)}}_{\text{Posterior Sample Ratio}} \cdot \underbrace{\frac{q\left(\varphi^{(t)}\middle|\hat{\varphi}^{(t+1)}\right)}{q\left(\hat{\varphi}^{(t+1)}\middle|\varphi^{(t)}\right)}}_{\text{Forward-Backward Ratio}}}^{\text{Hastings Ratio}}, 1\right]. \qquad (2)\end{aligned}$$

Otherwise, $\varphi^{(t+1)} = \varphi^{(t)}$. Convergence to the stationary distribution occurs after a suitable number of iterations (i.e. the mixing time) which produces a *single* sample from the posterior. Evaluating the Hastings ratio, the product of the two ratios in the acceptance probability, has been the primary barrier for implementing MCMC methods over implicit representations. In particular, one needs to solve a correspondence problem to compute the probability of generating the forward and reverse transition (in the forward-backward ratio). Doing so satisfies the condition of detailed balance which, in addition to ergodicity, is sufficient for convergence to the desired posterior distribution.

As with any level-set representation, one needs to choose the magnitude of the level-set, φ, away from the curve. Previous sampling methods have constrained the level-set function to be a signed distance function (SDF). Chen [4] solves the correspondence problem by generating perturbations that are SDF-preserving, thus having a one-to-one mapping from forward and reverse transitions. An alternative is to produce a non-SDF-preserving perturbation and reinitialize the level set function to an SDF at each iteration. However, this creates a many-to-many correspondence problem which significantly increases the computational complexity of the forward-backward ratio.

Our idea is straightforward: do not constrain the level-set function to be an SDF. SDFs provide advantages in terms of numerical stability and the computation of the curvature (see [17] for details) for optimization based methods. As the method here is not PDE-based and optimization is not the specific goal, there is essentially no penalty for using an alternative. While our level-set function no longer satisfies the SDF property, we still benefit from the way implicit representations handle topological changes and reparameterization. Furthermore, this greatly simplifies the design and evaluation of a proposal distribution by allowing for straightforward evaluation of the Hastings ratio.

3.1. Strategic Bias in the Proposal Distribution

We note that the closer $q(\circ|\triangle)$ is to $\pi(\circ)$, the closer the Hastings ratio is to unity and the higher the acceptance rate. Consequently, designing proposal distributions which capture essential, application-specific characteristics of the posterior distribution can improve convergence speeds by reducing the number of rejected samples. By relaxing the SDF constraint on the level-set function, many potential proposal distributions will result in a tractable evaluation of the Hastings ratio. Without care, however, the majority of these proposal distributions will have very poor mixing times. Thus, our aim is to design a proposal distribution that is easily evaluated, has a high acceptance rate, and explores the configuration space via large perturbations.

In Equation 2, the Hastings ratio consists of the posterior sample ratio (PSR) and the forward-backward ratio (FBR). The PSR represents the ratio of the posterior probability of the new sample over that of the old. Generating samples that have higher posteriors will produce high values of this ratio. The FBR represents the probability of generating the previous sample conditioned on the new one (the backward transition) over the probability of generating the new sample conditioned on the previous one (the forward transition).

Fan et al. [5] suggest using a proposal distribution biased by the curvature to favor samples that fit the prior model. Here, we develop a proposal which favors both the likelihood and prior model. This generally produces higher PSR values, but biases the FBR toward smaller values (see the supplemental materials for an illustrative example). Thus, our goal is to develop a proposal distribution with a higher overall Hastings ratio (the product of the PSR and the FBR), where deleterious effects on the FBR are compensated with increases in the PSR. Exploiting the simple observation that neighboring pixels tend to have the same label, we can develop a proposal that has this property.

We construct an additive perturbation, $\mathbf{f}$, to $\varphi^{(t)}$,

$$\hat{\varphi}^{(t+1)} = \varphi^{(t)} + \mathbf{f}^{(t)}, \tag{3}$$

by first sampling from a point process, attributing the points with values sampled from a biased Gaussian distribution and then smoothing with a lowpass filter. We refer to this process as Biased and Filtered Point Sampling (BFPS). The lowpass filter captures the property that pixels in close proximity have higher probability of being in the same region while the *choice* of bias favors points with high likelihood under the energy functional. The result is dramatically increased PSRs using large biased moves while only slightly decreasing the FBR. Mathematically this is expressed as

$$\mathbf{f}^{(t)} = \mathbf{h}^{(t)} * \left(\mathbf{c}^{(t)} \circ \mathbf{n}^{(t)}\right), \tag{4}$$

$$n_i^{(t)} \sim \mathcal{N}\left(\mu_i^{(t)}, \sigma^2\right), \quad c_i^{(t)} \sim \text{Bernoulli}\left(p_{c_i}^{(t)}\right), \tag{5}$$

where '$*$' denotes convolution and '$\circ$' denotes the elementwise product. We bias the Gaussian RVs with the gradient velocity, $\mathbf{v}^{(t)}$, (the negative gradient of the energy functional) to prefer moving to more probable configurations:

$$\mu_i^{(t)} = \alpha_n \left[-\frac{\partial E\left(\varphi^{(t)}\right)}{\partial \varphi^{(t)}}\right]_i = \alpha_n v_i^{(t)}, \tag{6}$$

where α_n is a weighting parameter. The probability associated with each point, c_i, is also carefully selected to favor selecting points which are better explained in another region. Specifically, it is chosen to be higher for points that have a gradient velocity that is large in magnitude *and* has the opposite sign of the current level-set value:

$$p_{c_i}^{(t)}(1) \propto \alpha_c \exp\left[-v_i^{(t)} \cdot \text{sign}\left(\varphi_i^{(t)}\right)\right] + (1 - \alpha_c), \tag{7}$$

where α_c is a parameter that trades off the bias with a uniform distribution. Additionally, we define the variable γ as $\frac{1}{|\Omega|}\sum_{i\in\Omega} p_{c_i}^{(t)}(1) = \gamma$, which approximates the average probability that a random point will be selected, where Ω is the set of all pixels. Because $p_{c_i}^{(t)}(1)$ is only defined up to a scale factor, we can renormalize its value to achieve any γ. In practice, α_n, α_c, and γ are dynamically adapted to maintain a minimum acceptance rate, and $\mathbf{h}^{(t)}$ is chosen to be a circularly symmetric (truncated) Gaussian kernel with a scale parameter randomly chosen from a finite set of values. Randomly chosen scale parameters introduce a minor complication (which we address), but empirically result in faster mixing times.

3.2. Sufficient Conditions for MCMC Sampling

In MCMC methods, convergence to the correct stationary distribution is a key issue. It is sufficient, and often easier, to satisfy the following conditions: (1) that the chain is **ergodic** and (2) that each individual step in the MCMC procedure satisfies **detailed balance**.

Ergodicity requires the Markov chain to be aperiodic and irreducible. Proving a complicated Markov chain is aperiodic is very difficult [8]. Similar to [4] and [5], we argue that our Markov chain is unlikely to be periodic because the space of segmentations is so large. In the rare case that the chain is periodic but still irreducible, the average sample path will still converge to the distribution from which we are trying to sample. Irreducibility of a Markov chain implies that any state in the chain has finite probability of reaching any other state in the chain. Fan[5] and Chen[4] only show that the chain is irreducible in the space of single simply connected components. Additionally, they require multiple iterations to show that any curve can be altered to any other curve. The method here, however, allows for any topological change and has finite probability of transitioning from any curve to any other curve in a *single* perturbation. This is trivially shown as $\mathbf{c}$ is a Bernoulli process that has finite probability of being one everywhere and $\mathbf{n}$ is a Gaussian process that has finite probability of taking on any value.

Detailed balance is satisfied as long as the Hastings ratio (Equation 2) is calculated correctly. Often, the energy functional we are sampling from depends on the probability density of the observed data conditioned on a segmentation. In these cases, the densities change at every iteration, which we assume are properly updated using a kernel density estimate. The PSR can be computed as the following:

$$\frac{\pi\left(\hat{\varphi}^{(t+1)}\right)}{\pi\left(\varphi^{(t)}\right)} = \frac{\exp\left[-E\left(\hat{\varphi}^{(t+1)}\right)\right]}{\exp\left[-E\left(\varphi^{(t)}\right)\right]}. \tag{8}$$

Recall that the new proposed level-set, $\hat{\varphi}^{(t+1)}$ is only dependent on the previous level-set, $\varphi^{(t)}$, and the random perturbation, $\mathbf{f}^{(t)}$. Thus, the FBR can be written as

$$\frac{q\left(\varphi^{(t)}\middle|\hat{\varphi}^{(t+1)}\right)}{q\left(\hat{\varphi}^{(t+1)}\middle|\varphi^{(t)}\right)} = \frac{p_{\mathbf{F}}\left(-\mathbf{f}^{(t)}|\hat{\varphi}^{(t+1)}\right)}{p_{\mathbf{F}}\left(\mathbf{f}^{(t)}|\varphi^{(t)}\right)}. \tag{9}$$

For a single lowpass filter, there exists a simple, one-to-one mapping between the forward and backward transitions. However, recall that *multiple* filters of different variances are used to speed up the algorithm. A realized perturbation, $\mathbf{f}$, can therefore be generated from *multiple* combinations of $\{\hat{\mathbf{h}}, \hat{\mathbf{c}}, \hat{\mathbf{n}}\}$. Exact calculation of the FBR requires the probability of generating $\mathbf{f}$ using each of these combinations. We refer to $\{\mathbf{h}, \mathbf{c}, \mathbf{n}\}$ without hats as the actual combination that was used to generate $\mathbf{f}$. Here, we show that the probability of generating the perturbation is dominated by $\{\hat{\mathbf{h}}, \hat{\mathbf{c}}, \hat{\mathbf{n}}\} = \{\mathbf{h}, \mathbf{c}, \mathbf{n}\}$.

We first note the following relationship: $\hat{\mathbf{c}} \circ \hat{\mathbf{n}} = \hat{\mathbf{h}}^{-1} * \mathbf{f}$, where $\hat{\mathbf{h}}^{-1}$ is the highpass filter that is the inverse of $\hat{\mathbf{h}}$. The probability of generating a perturbation can be written as:

$$\begin{aligned} p_{\mathbf{F}}(\mathbf{f}|\varphi) &= \sum_{\hat{\mathbf{h}}} p_{\mathbf{H}}(\hat{\mathbf{h}}) p_{\mathbf{C}}(\hat{\mathbf{c}}|\varphi, \hat{\mathbf{h}}) p_{\mathbf{N}}\left(\hat{\mathbf{n}}|\varphi, \hat{\mathbf{c}}\right) \\ &= \sum_{\hat{\mathbf{h}}} p_{\mathbf{H}}(\hat{\mathbf{h}}) p_{\mathbf{C}}(\hat{\mathbf{c}}|\varphi, \hat{\mathbf{h}}) \prod_{\substack{i\in\Omega \\ \hat{c}_i=1}} \mathcal{N}\left(\hat{n}_i; \mu_i(\varphi), \sigma^2\right). \end{aligned} \tag{10}$$

Additionally, recall that $\hat{\mathbf{c}}$ is a set of sparse points with an average probability of being nonzero of $\gamma \ll 1$. When $\hat{\mathbf{h}} \neq \mathbf{h}$, $\hat{\mathbf{c}}$ will be nonzero at almost every pixel, allowing us to conclude the following inequality:

$$p_{\mathbf{C}}(\hat{\mathbf{c}}|\varphi, \hat{\mathbf{h}} \neq \mathbf{h}) \approx \gamma^{|\Omega|} \ll \gamma^{\gamma|\Omega|} \cdot \overline{\gamma}^{\overline{\gamma}|\Omega|} \approx p_{\mathbf{C}}(\mathbf{c}|\varphi, \mathbf{h}), \tag{11}$$

where $\overline{\gamma} \triangleq 1 - \gamma \approx 1$. Noting that $p_{\mathbf{N}}\left(\hat{\mathbf{n}}|\varphi, \hat{\mathbf{c}} \neq \mathbf{c}\right) \ll p_{\mathbf{N}}\left(\mathbf{n}|\varphi, \mathbf{c}\right)$, we conclude that the probability of generating the perturbation with a filter $\hat{\mathbf{h}} \neq \mathbf{h}$ is much less than the probability of generating it with $\hat{\mathbf{h}} = \mathbf{h}$. We can therefore approximate the probability of a particular perturbation as

$$p_{\mathbf{F}}(\mathbf{f}|\varphi) \approx \frac{1}{N_h} p_{\mathbf{C}}(\mathbf{c}|\varphi) \prod_{\substack{i\in\Omega \\ c_i=1}} \mathcal{N}\left(n_i; \mu_i(\varphi), \sigma^2\right), \tag{12}$$

where N_h is the number of possible filters, a filter is chosen with uniform probability, and $p_{\mathbf{C}}(\mathbf{c}|\varphi)$ is evaluated using Equation 7. Combining these equations with Equations 8 and 9 allows a straightforward and efficient calculation of the Hastings ratio, ensuring detailed balance.

3.3. Extension to M-ary Shape Sampling

In the context of level-set representations, separate extensions to M-ary segmentation have been suggested by Chan and Vese [2] and Brox and Weickert [1]. These extensions do not lend themselves to sampling approaches; consequently, we suggest a novel alternative. Let M level-set functions represent $M + 1$ regions. The last M regions, $R_1, ..., R_M$ each contain the positive values of its respective level-set function. The null region, R_0, contains those pixels that are not contained by any other region. More precisely, we have the following definition of regions:

$$R_0 = \bigcap_{\ell\in\mathcal{L}} \{i \mid \varphi_\ell(i) < 0\} \tag{13}$$

$$R_\ell = \{i \mid \varphi_\ell(i) \geq 0\}, \quad \forall \ell \in \mathcal{L} = \{1, 2, \ldots, M\} \tag{14}$$

When developing an M-ary representation, one must ensure that both vacuum (a pixel belongs to no region) and overlap conditions (a pixel belongs to multiple regions) will not occur. Due to the null region, vacuum conditions never occur; however, an overlap condition may occur among the

regions $R_1, R_2, \ldots, R_M$. We develop a perturbation similar to the binary case that precludes both of these conditions.

At each iteration, randomly select one of the level-sets, φ_ℓ. Each pixel, i, in this level-set can be categorized into one, and only one, of the following three types: (1) the pixel belongs to R_ℓ, (2) the pixel belongs to R_0, or (3) the pixel belongs to $\{R_l | l \in \mathcal{L}, l \notin \{\ell, 0\}\}$. By only allowing transitions between pixels of type 1 and 2, an overlap condition cannot occur. The proposed perturbation is then of the same form as before:

$$\hat{\varphi}_\ell^{(t+1)} = \varphi_\ell^{(t)} + \mathbf{f}_\ell^{(t)}. \tag{15}$$

The new $\mathbf{f}_\ell^{(t)}$ is drawn from the following

$$\mathbf{f}_\ell^{(t)} = \left(\mathbf{h}^{(t)} * \left(\mathbf{c}_\ell^{(t)} \circ \mathbf{n}_\ell^{(t)}\right)\right) \circ \mathbb{1}_{\{R_\ell \cup R_0\}}, \tag{16}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function, and is included to ensure only pixels of type 1 and 2 are affected. This restriction, along with the modified proposal distribution described subsequently, can be implemented with essentially no penalty. In very specific instances, it can lead to poor convergence times, but these situations can be precluded with a proposal that randomly selects the null region.

To ensure a good proposal distribution, we alter the Bernoulli process, $\mathbf{c}$, and the mean of the Gaussian random variables, μ_i. We use the notation that (ℓ) is the label of the level-set we are currently perturbing and (l) is the label of another level-set (i.e. $l \neq \{0, \ell\}$). In the binary case, only one gradient velocity existed because there was only one level-set. With multiple sets, we define the quantity $\mathbf{v}\,(\ell, l)$ as the gradient velocity between regions R_ℓ and R_l. In the M-ary framework presented above, we would only consider $\mathbf{v}\,(\ell, 0)$ because only moves to and from R_0 are allowed. However, the null region, R_0, acts as a temporary region for pixels switching between other regions. More specifically, if a pixel in the level-set that is currently being perturbed, φ_ℓ, would be better described in R_l, then there should be a force to move the pixel to the null region so it can ultimately move to region R_l. This observation is reflected in our proposal distribution by replacing the true gradient velocity, $\mathbf{v}(\ell, 0)$, in Equations 6 and 7 with the following minimal gradient velocity, $\mathbf{m}(\ell)$, at pixel i:

$$m_i(\ell) = \min_{\substack{l \in \{0,1,2,\ldots,M\} \\ l \neq \ell}} v_i\,(\ell, l)\,. \tag{17}$$

This minimal gradient velocity essentially trades off the current region label with the other most likely label. When $M = 1$, this formulation simplifies to the binary case.

4. Applying BFPS

BFPS is a general method with application to a variety of energy functionals over implicit representations. Such an application is predicated on the evaluation of the gradient of the energy functional. Here, we present a few energy functionals and features that are easily incorporated into the approach along with a description of some marginal statistics of interest. In particular, whereas in the past, region-based methods were rarely evaluated over image data sets where edge detection is the goal, utilizing BFPS allows for straightforward evalution of region-based methods in edge-detection tasks.

Table 1 shows three information-theoretic energy functionals used in previous optimization-based segmentation algorithms (mutual information [14], J-Divergence [13], and Bhattacharya Distance [16]) and their corresponding gradients. Here, X represents the image feature, and $p_X^{\pm}$ represents the densities of the feature in the $R^{\pm}$ region. Using these energy functionals within BFPS is a matter of replacing the functionals in Equation 8 and the gradient velocities in Equations 6, 7, and 12. In each case, we assume that a curve length penalty is used for regularization.

Alternative image features are also adaptable to BFPS. For example, results using the features of [14] (scalar intensity), [13] (scalar texture measure), and [12] (vector texture measure) will be shown in Section 5. Furthermore, distributions of image feature can be described using parametric (e.g. [12]) or non-parametric models (e.g. [14], [13]).

4.1. Marginal Statistics

As is typical in MCMC approaches, marginal statistics can be evaluated over samples using a simple counting measure. Similar to [5], one can compute the histogram image of a segmentation, where each pixel in the histogram contains a count of the number of times it was included in a particular region. Similarly, the 50% quantile curve corresponds to thresholding the histogram image at 0.5.

Here, we consider another marginal event probability: the probability that a pixel lies on the boundary. We refer to this as the probability of boundary image (PB). The PB at pixel i is calculated by simply counting the number of samples for which pixel i lies on a boundary and normalizing by the number of samples. This statistic is of particular interest as it allows one to evaluate results over the Berkeley Segmentation Dataset (BSDS) [15] which compares precision-recall (PR) curves on precisely this event probability. In this dataset, the maximum harmonic mean of points on the PR curve, or F-measure, is used as the metric for rating boundary detectors. Unlike boundary detectors, however, optimization-based segmentation algorithms produce a single point on the PR curve. Recent segmentation algorithms rarely report benchmark results on the BSDS due to poor F-measures owing to the inability to trade off between precision and recall. BFPS enables these segmentation algorithms to produce a PB image for more robust comparison on the BSDS.

Description	Energy ($E(\varphi)$)	Gradient Velocity at Pixel i (v_i), $\quad R_-^+ = \frac{p_X^+(x_i)}{p_X^-(x_i)}$
Mutual Information	$-\lvert\Omega\rvert I(X;L) + \alpha \oint_{\mathcal{C}} dl$	$\log R_-^+ - \alpha\kappa_i$
J Divergence	$-\lvert\Omega\rvert J\left(p_X^+; p_X^-\right) + \alpha \oint_{\mathcal{C}} dl$	$\frac{1}{\pi^+}\left[\log R_-^+ - D\left(p_X^+ \lVert p_X^-\right) - R_+^- + 1\right] - \frac{1}{\pi^-}\left[\log R_+^- - D\left(p_X^- \lVert p_X^+\right) - R_-^+ + 1\right] - \alpha\kappa_i$
Bhattacharya Distance	$-\lvert\Omega\rvert \int_{\mathcal{X}} \sqrt{p_X^+(x)\, p_X^-(x)}\, dx + \alpha \oint_{\mathcal{C}} dl$	$\frac{1}{2\pi^-}\sqrt{R_-^+} - \frac{1}{2\pi^+}\sqrt{R_+^-} - \alpha\kappa_i$

Table 1: Energy Functionals and Corresponding Gradient Velocities

Figure 2: Synthetic example illustrating the importance of allowing topological changes. The histogram image obtained using each sampling algorithm is shown.

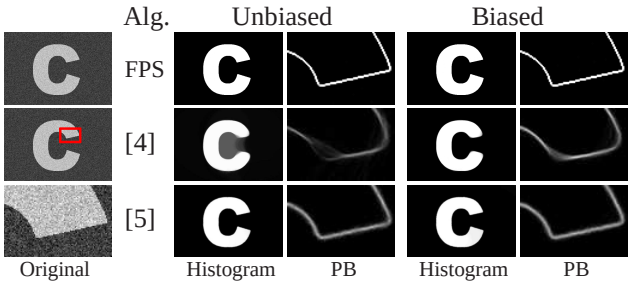

Figure 3: Synthetic example of each algorithm after 100,000 iterations. Each row shows the histogram image and a detail of the PB image using both an unbiased and biased version of a sampling algorithm.

5. Empirical Results

In this section, we demonstrate the use of the BFPS procedure. Unless otherwise stated, we use nonparametric pixel intensities as the image feature and mutual information with a curve length penalty as the energy functional. While [14] has shown that this combination produces good results in an optimization framework for a wide variety of images, we choose it merely to illustrate the sampling aspects of BFPS. Other functionals might yield differing segmentations, though relative comparisons between sampling approaches, specifically [4] and [5], would remain the same.

5.1. Topological Changes and Computation Times

While the M-ary extension is useful, the primary advantages of BFPS over [4] and [5] are the ability to handle topological changes and the improvement in computation time. We demonstrate these advantages with two examples. As [4] and [5] are restricted to simply connected shapes, we initialize the segmentation with a single circle of radius 50 pixels, centered at a random location.

Figure 2 shows a noisy image containing the letter 'O'. Each region is composed of normally distributed, i.i.d. pixels. Since the approaches of [4] and [5] do not allow for topological changes, the iterations either settle on the exterior or interior boundary of the 'O' (but never both) depending on the initialization. This is a simple example of the importance of handling topological changes.

The computation time needed to draw a sample from the posterior depends on two factors: (1) the time to draw and evaluate a sample from the proposal distribution and (2) the number of iterations needed from the proposal distribution before the Markov chain reaches its stationary distribution. We examine the computation times for six algorithms based on BFPS and the algorithms of [4] and [5]. While [4] and [5] do not incorporate a gradient bias, we implement both with and without a bias to illustrate its impact. We refer to the algorithms as BFPS, UFPS, B[5], U[5], B[4], and U[4] where the preceding 'B' and 'U' indicate a biased or unbiased algorithm. In [5], the bias corresponds to moving each marker point with the gradient, and in [4], to both selecting and moving the so-called "foot point" with the gradient.

Consider the synthetic image of Figure 3 containing a simply connected 'C'. We run each algorithm for 100,000 iterations (which for [4] takes over 8 hours to evaluate a single sample path). The histograms in Figure 3 imply that all algorithms, aside from U[4], have converged. Examination of a detail (see Figure 3) of the 'C' and the PB associated with each algorithm shows this not to be the case; it is clear that both biased and unbiased versions of [4] and [5] have not converged. The results of [5] have a blurred PB, and the results of [4] are both blurred and miss corners.

The plot in Figure 4 shows the average energy across all sample paths for each algorithm as a function of the number of iterations. We note that while the average energy appears to be non-decreasing, the energy in each sample path both increases and decreases. While all of the algorithms will eventually converge to the stationary distribution, Figure 4 illustrates the stark difference in mixing times. BFPS converges in approximately 150 iterations while the unbiased version, UFPS, converges in approximately 40,000 iterations. After 100,000 iterations, all other algorithms have yet to converge.

When calculating total computation time, one must also

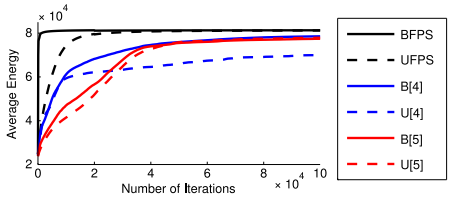

Figure 4: The average energy across all sample paths vs. the number of iterations for multiple sampling algorithms.

Algorithm	Iterations Until Convergence	Seconds per Iteration	Total Gain
BFPS	150	0.03	×1
UFPS	40,000	0.025	×222
B[4]	254,000	0.30	×16,933
U[4]	896,000	0.26	×51,769
B[5]	321,000	5.00	×356,667
U[5]	336,000	5.00	×373,333

Table 2: Computation Times of Algorithms

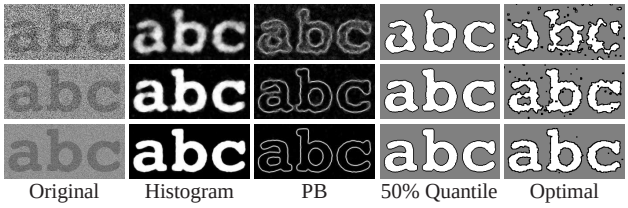

Figure 5: Results for three synthetic images with varying SNR values (0.5, 1.0, and 2.0, top to bottom, respectively).

consider the time it takes to generate and evaluate a single sample from the proposal. These times are summarized in Table 2. We linearly interpolate the average energy using the last 5,000 iterations to estimate how many iterations are needed for the algorithms based on [4] and [5], noting that this is an optimistic *lower* bound on the number of iterations as the average energy grows sub-linearly. While the bias term increases speed in all algorithms, BFPS is still over 15,000 times faster than any other biased method. BFPS is over 50,000 faster than the original formulations (i.e. excluding the bias) of U[4] and over 300,000 faster than U[5].

5.2. Low SNR Segmentations

The previous results illustrate the computational advantages of BFPS over other sampling algorithms. We now show results of using BFPS in a few applications. Consider the synthetic images shown in Figure 5. Each image contains two regions that are drawn from Gaussian distributions with different means. We alter the variance to consider three different SNR values: 0.5, 1.0, and 2.0. The last column shows the sample path with the highest energy which approximates the optimal configuration. In the lowest SNR

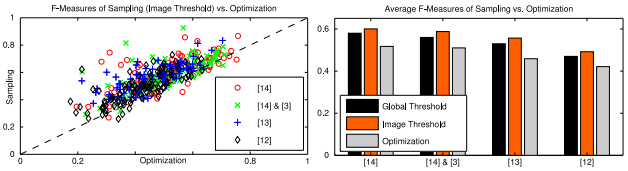

Figure 6: Sampling vs. optimization on the BSDS. The scatter plot shows the F-measures of each image using an image-based threshold on the PB image vs. the optimal segmentation. The bar plot shows the average F-measures using the global threshold on the PB image, the image-based threshold on the PB image, and the optimal segmentation.

case, the 50% quantile clearly produces much better results than the optimal sample path. As the SNR increases, the optimal sample path approaches the average sample path. Consequently, in low SNR scenarios, marginal event probabilities tend to be more robust than optimal configurations.

5.3. Boundary Detection

As stated previously, marginal events such as boundaries are of interest. Due to their inherent topological constraints, [4] and [5] are less applicable to natural images where it is often desirable to group regions which are separated spatially and/or segment an image into more than two region labels. As such, the remaining results focus on the use of the M-ary version of BFPS. We consider four different image features: the raw intensity of a pixel [14], the intrinsic intensity of a pixel [14] & [3], the shape operator [13], and the steerable pyramid output [12]. The intrinsic intensity is estimated a priori, meaning that a gain and bias field [3] are estimated and removed prior to segmentation. As BFPS extends almost *any* segmentation algorithm to a boundary detector, the emphasis here is not on a particular energy functional or image feature, but rather the improved performance via marginal statistics (made feasible by BFPS) compared to optimization. To avoid local minima, we run gradient descent with 100 random initializations and select the minimal energy configuration for each image. Results across the entire BSDS are shown in Figure 6. In addition to reporting performance on BSDS with the average F-measure (as is typical) we also report results using the optimal image-based threshold. While a measure of image complexity or contextual content might provide a means of approximating such a threshold, our purpose is to illustrate the achievable gains using the PB image. Regardless, results are reported using both global and image-based thresholds, and in either case, sampling improves upon the optimization approach across the majority of images in the dataset.

Figure 7 shows results on four specific images from the BSDS. Qualitatively, the PB image provides a superior de-

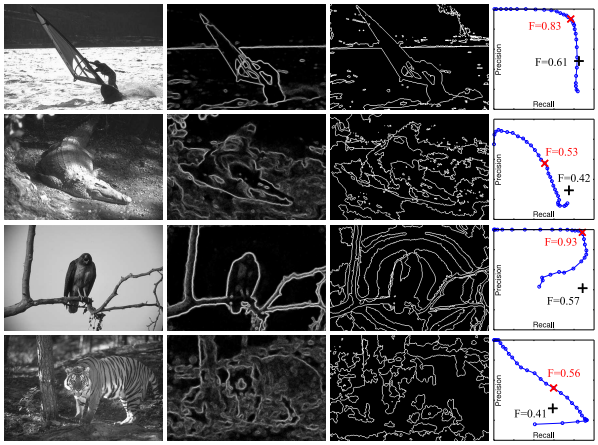

Figure 7: Example images from [15]. In the PR curves, the '×' marks the F-measure obtained using BFPS, and the '+' marks that of the optimal sample. The first two rows use the image feature of [14]. The third row also uses the image feature of [14] but with the gain and bias field of [3]. The fourth row uses the textural image feature of [13].

marcation of edges in the image. Quantitatively, the F-measure is also improved by thresholding the PB image rather than using the optimal sample path. Additional results are included in the supplemental material.

6. Conclusion

We have presented an MCMC framework that allows one to sample from the space of segmentations. The formulation was developed with a general energy functional and image feature such that almost any optimization based segmentation algorithm can be used. In contrast to previous methods, BFPS easily and efficiently handles topological changes, large perturbations, and multiple regions, while exhibiting a 50,000 times speed up. In addition to more robust segmentation quantiles, we have demonstrated that MCMC sampling also allows one to extend a region based, level set segmentation algorithm to a boundary detector. This development enables the evaluation of region-based methods on the Berkeley Segmentation Dataset as a common benchmark comparison. Publicly available code for the BFPS algorithm, easily adaptable to any image feature and energy functional, can be found at http://people.csail.mit.edu/jchang7/.

7. Acknowledgments

This research was partially supported by the Air Force Office of Scientific Research under Award No. FA9550-06-1-0324 and the Air Force Research Laboratory under Award No. FA8650-07-D-1220. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Air Force.

References

[1] T. Brox and J. Weickert. Level Set Segmentation with Multiple Regions. *IEEE Trans. on Image Processing*, 15(10):3213–3218, Oct. 2006.

[2] T. Chan and L. Vese. An efficient variational multiphase motion for the Mumford-Shah segmentation model. *Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers, 2000*, 1:490–494 vol.1, 2000.

[3] J. Chang and J. W. Fisher III. Analysis of Orientation and Scale in Smoothly Varying Textures. *Proc. of ICCV*, 2009.

[4] S. Chen and R. J. Radke. Markov Chain Monte Carlo Shape Sampling using Level Sets. *NORDIA, in conjunction with ICCV*, 2009.

[5] A. C. Fan, J. W. Fisher III, W. M. Wells III, J. J. Levitt, and A. S. Willsky. MCMC curve sampling for image segmentation. *MICCAI*, 2007.

[6] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2003.

[7] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. on PAMI*, (6):721 –741, Nov. 1984.

[8] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. CRC Press, 1996.

[9] L. Greengard and J. Strain. The Fast Gauss Transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

[10] U. Grenander and M. I. Miller. Computational anatomy: an emerging discipline. *Q. Appl. Math.*, LVI(4):617–694, 1998.

[11] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[12] M. Heiler and C. Schnorr. Natural Image Statistics for Natural Image Segmentation. *Proc. of ICCV*, 2003.

[13] N. Houhou, J.-P. Thiran, and X. Bresson. Fast Texture Segmentation Model Based on the Shape Operator and Active Contour. In *Proc. of CVPR*, 2008.

[14] J. Kim, J. W. Fisher III, A. Yezzi, M. Cetin, and A. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Trans. on Image Processing*, 14:1486–1502, Oct. 2005.

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. of ICCV*, 2001.

[16] O. Michailovich, Y. Rathi, and A. Tannenbaum. Image Segmentation Using Active Contours Driven by the Bhattacharyya Gradient Flow. *IEEE Trans. on Image Processing*, 16(11):2787–2801, 11 2007.

[17] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 31 Oct. 2002.

[18] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.