IEEE Conference on Computer Vision and and Pattern Recognition, Kauai, 2001, Vol. II. pp. 358–365

Probabilistic Tracking of Motion Boundaries with Spatiotemporal Predictions

Oscar Nestares[*†] **David J. Fleet**[*]

[*] Xerox Palo Alto Research Center, 3333 Coyote Hill Rd, Palo Alto, CA 94304
[†] Instituto de Óptica "Daza de Valdés" (C.S.I.C.), Serrano 121, 28006-Madrid, Spain

Abstract

We describe a probabilistic framework for detecting and tracking motion boundaries. It builds on previous work [4] that used a particle filter to compute a posterior distribution over multiple, local motion models, one of which was specific for motion boundaries. We extend that framework in two ways: 1) with an enhanced likelihood that combines motion and edge support, 2) with a spatiotemporal model that propagates beliefs between adjoining image neighborhoods to encourage boundary continuity and provide better temporal predictions for motion boundaries. Approximate inference is achieved with a combination of tools: Sampled representations allow us to represent multimodal non-Gaussian distributions and to apply nonlinear dynamics. Mixture models are used to simplify the computation of joint prediction distributions.

1 Introduction

The detection of dynamic occlusion and the reliable estimation of 2D motion and relative surface depths at motion discontinuities are long-standing problems in visual motion analysis. Black and Fleet [4] proposed a Bayesian approach with a generative model for local motion in which the 2D flow is either smooth or discontinuous, along with a particle filter to approximate the posterior probability distribution over the models and model parameters. This paper extends that framework in two respects. First, while Black and Fleet [4] considered local image regions in isolation, here we consider a dense array of smaller neighborhoods, combined with a probabilistic model for neighborhood dependencies. Second, we extend the original motion-based likelihood function to include an empirical edge-based likelihood function to improve boundary localization.

Particle filters have become a popular method of approximate inference for dynamical systems [10, 13, 15, 17, 31]. With point-mass approximations to probability distributions, they are effective for non-Gaussian, multimodal distributions that occur with nonlinear dynamics and observation equations. Multimodal distributions are particularly significant with hybrid state-space models having continuous and discrete variables that depend on one another. Here we use discrete and continuous states to represent motion classes and their corresponding motion parameters.

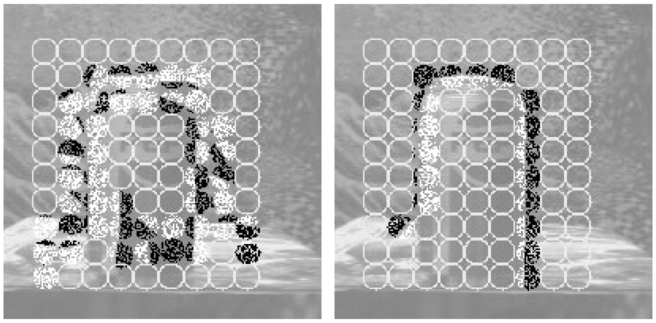

Figure 1. Motion estimates at frame 6 without (left) and with (right) spatiotemporal dependence. Each region depicts the mean of the principal mode of the posterior. Empty circles depict smooth motion. Filled regions are motion boundaries, with white dots on the foreground side.

One issue with particle filters is computational cost. Although effective for low-dimensional tracking, it is not clear how they scale to high-dimensional problems (e.g., see [5, 7, 18]). A related concern is the effectiveness of particle filters for inference with dynamic random fields. Iterative solutions using MCMC can be prohibitively slow, and Bayesian belief propagation may be impractical with Monte Carlo approximations to complex, hybrid distributions.

For approximate inference, Black and Fleet assumed that the motion in each image region could be modeled by an independent Markov chain, estimating the motion in each local region separately. But, by assuming independence this fails to exploit the information that one region could obtain from its neighbors; it is difficult to encourage continuity along motion boundaries, and to make accurate predictions about boundary locations from one time to the next. As an example, Fig. 1(left) depicts motion estimates obtained from isolated particle filters in each circular region. Fig. 1(right), by comparison, depicts the motions estimated with neighborhood dependencies as described below.

This paper explores the use of spatiotemporal predictions with Bayesian filtering to detect and track motion boundaries. We use a simplified dependency graph (see Fig. 2(right)) in which it is assumed that regions at time t, conditioned on nearby regions at time $t-1$, are independent of other regions at current and past times. We also make use of several different inference tools: First, we approximate the joint distribution over multiple regions by a collection of marginals [19]. Second, we use Monte Carlo approximations to these distributions to deal with the nonlinear dynamics and non-Gaussian likelihoods. Finally, we use mixture models to efficiently approximate the prediction distributions from multiple neighborhoods.

2 Previous Work

There are more methods for detecting motion boundaries and estimating discontinuous motion than we can adequately review here. But, broadly speaking, there are two classes of related approaches: 1) those that treat motion boundaries as a source of error for optical flow techniques, and 2) those that explicitly try to detect motion boundaries.

Techniques that view motion discontinuities as a source of noise (outliers) include MRF and regularization formulations where robust statistics, weak continuity, or line processes are used to disable smoothing across motion discontinuities [6, 11, 12, 16, 21, 24, 25, 28]. Robust regression [3, 23] and mixture models [1, 14, 30] allow for multiple motions to occur in a region, and thus provide some degree of robustness at boundaries. These methods improve motion estimation, but they fail to explicitly estimate the structure of local motion boundaries; e.g., they do not estimate which side of the boundary is the foreground.

Other approaches to detecting discontinuities include the detection of bimodality in local distributions of optical flow [27], the application of edge detectors to estimated flow fields [24, 28], and the detection of spectral signatures [2, 22]. Still others have used the presence of unmatched features to detect dynamic occlusions [20, 28]. Few of these methods model the spatial structure of the motion in the immediate neighborhood of a motion boundary, and they have not proved reliable in practice.

3 Bayesian Inference of Motion Boundaries

Following [4] we take a Bayesian approach. We assume a hybrid state space with two motion classes for each local region of the image: 1) a translational model to capture smooth motion; and 2) an explicit nonlinear model of motion boundaries. The state space description, $\mathbf{s} = (\mu,\, \mathbf{c}_\mu)$, includes discrete and continuous random variables. The discrete variable μ encodes the type of motion, and the continuous vector $\mathbf{c}_\mu$ encodes the parameters for the corresponding motion class μ. Smooth motion is parameterized by image translation, $\mathbf{u}$. As shown in Fig. 2(left), motion boundaries are parameterized by the foreground and background velocities, the edge orientation and the normal distance from the edge to the region center, $\mathbf{c}_e \equiv (\mathbf{u}_f, \mathbf{u}_b, \theta, d)$.

With Bayes' rule and the assumption that the motion in each image neighborhood forms a Markov chain, the usual recursive filtering equation takes the form

$$p(\mathbf{s}_t \,|\, \bar{\mathbf{I}}_t) \;=\; k\, p(\mathbf{I}_t \,|\, \mathbf{s}_t, \mathbf{I}_{t-1})\, p(\mathbf{s}_t \,|\, \bar{\mathbf{I}}_{t-1}) \qquad (1)$$

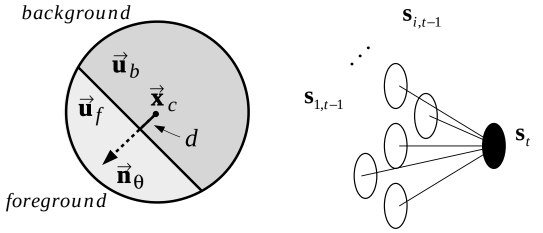

Figure 2. (left) Motion boundary parameterization: $\mathbf{u}_f$ and $\mathbf{u}_b$ denote foreground and background velocities, θ denotes edge orientation with normal $\mathbf{n}_\theta$, and d is the signed perpendicular distance of the edge from the region center $\mathbf{x}_c$. (right) The assumed spatiotemporal dependency.

where $\mathbf{s}_t$ is the state variable for an image region at time t, k is a normalization constant, and $\bar{\mathbf{I}}_t = \{\mathbf{I}_0, \ldots, \mathbf{I}_t\}$ is the observation history. In (1), the conditional observation density, $p(\mathbf{I}_t \,|\, \mathbf{s}_t, \mathbf{I}_{t-1})$, is called the likelihood function. The second factor, $p(\mathbf{s}_t \,|\, \bar{\mathbf{I}}_{t-1})$, is the prediction distribution. With the Markov assumption, and conditional independence of the observations, it can be expressed as

$$p(\mathbf{s}_t \,|\, \bar{\mathbf{I}}_{t-1}) \;=\; \int_{\mathbf{s}_{t-1}} p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1})\, p(\mathbf{s}_{t-1} \,|\, \bar{\mathbf{I}}_{t-1})\, d\mathbf{s}_{t-1} \qquad (2)$$

where the conditional distribution $p(\mathbf{s}_t \,|\, \mathbf{s}_{t-1})$ embodies the temporal dynamics, and $p(\mathbf{s}_{t-1} \,|\, \bar{\mathbf{I}}_{t-1})$ is the posterior distribution at the previous time, $t-1$.

With simple dynamical models, Gaussian process noise, and a likelihood function derived from intensity conservation, the prediction and posterior distributions are easy to formulate. However, because of the nonlinear dynamics of the motion boundary model, and because likelihoods are typically non-Gaussian and multimodal, the posterior (1) does not have an analytical solution. Fleet & Black [4] therefore used a particle filter to approximate the posterior. Although this approach produced promising results, it does have some drawbacks. First, by modeling the motion in each region independently of its neighbors, one ignores the wealth of information that one region can share with its neighbors. Second, their motion-based likelihood function fails to use the available information in images at occlusion boundaries. In particular, they do not exploit the fact that static edges often coincide with motion boundaries.

4 Neighborhood Prediction Distribution

In this paper we continue to approximate the joint posterior distribution over the motion in all image regions by the marginal distributions in each region. But rather than computing these marginals independently, we exploit predictions from several neighborhoods at the previous time. This provides a modest amount of influence from neighbors that helps to encourage edge continuity. It also helps to direct particles (in a particle filter) to the appropriate part of the state space. This is important when an edge is about to leave one neighborhood and enter another.

We assume a spatiotemporal pattern of neighborhood dependence like that in Fig. 2. Each location at time t receives predictions from nearby locations at the previous time $t-1$. Combining these predictions requires that we reformulate the prediction distribution in (2). Following the graphical model in Fig. 2, let the set of nodes that contribute to $\mathbf{s}_t$ be denoted by $\{\mathbf{s}_{i,t-1}\}_{i=1}^{\mathcal{N}}$. We begin by writing the prediction distribution as the marginalization of the joint distribution over $\mathbf{s}_t$ and its neighbors $\{\mathbf{s}_{i,t-1}\}_{i=1}^{\mathcal{N}}$:

$$p(\mathbf{s}_t\,|\,\bar{\mathbf{I}}_{t-1}) \;=\; \int_{\{\mathbf{s}_{i,t-1}\}} p(\mathbf{s}_t, \{\mathbf{s}_{i,t-1}\}_{i=1}^{\mathcal{N}}\,|\,\bar{\mathbf{I}}_{t-1})\,. \tag{3}$$

Factoring the integrand into the state dynamics and the joint posterior from the previous time $t-1$ yields

$$p(\mathbf{s}_t\,|\,\bar{\mathbf{I}}_{t-1}) \;=\; \int_{\{\mathbf{s}_{i,t-1}\}} p(\mathbf{s}_t\,|\,\{\mathbf{s}_{i,t-1}\})\; p(\{\mathbf{s}_{i,t-1}\}\,|\,\bar{\mathbf{I}}_{t-1})\,. \tag{4}$$

The dynamics, $p(\mathbf{s}_t\,|\,\{\mathbf{s}_{i,t-1}\})$, can be factored if we assume that the neighbors at time $t-1$ have uniform priors and are independent when conditioned on $\mathbf{s}_t$. The joint posterior over all neighbors can not be factored in general. However, for computational convenience we approximate the joint posterior as a product of its marginals, as in [19], to yield

$$p(\mathbf{s}_t\,|\,\bar{\mathbf{I}}_{t-1}) \;\approx\; \kappa \prod_{i\in\mathcal{N}} \int_{\mathbf{s}_{i,t-1}} p(\mathbf{s}_t\,|\,\mathbf{s}_{i,t-1})\; p(\mathbf{s}_{i,t-1}\,|\,\bar{\mathbf{I}}_{t-1}) \tag{5}$$

where κ is a constant. The resulting prediction distribution is product of the predictions from each neighboring location, each of which has the form of (2). This can be viewed as one iteration of loopy belief propagation [19, 29].

4.1 Particles and Gaussian Mixtures

The simplified prediction distribution in (5) allows us to combine predictions from each of the neighbors from time $t-1$ in a straightforward manner. If the marginal posteriors were Gaussian, and the dynamics were linear with Gaussian noise, then the prediction in (5) would also be Gaussian. However, the temporal dynamics for the motion boundary model are nonlinear, and the likelihood functions for both models are non-Gaussian and multimodal in general. This precludes an analytical solution to the integrals in (5).

Instead, we use Monte Carlo methods. As with a particle filter, each distribution $p(\mathbf{s}_{i,t-1}\,|\,\bar{\mathbf{I}}_{t-1})$ is represented by a weighted set of N samples $\{\mathbf{s}_{i,t-1}^{(j)}, w_{i,t-1}^{(j)}\}_{j=1}^{N}$. Then, the approximate prediction distribution from a single neighbor (as in (2)) can be viewed as a mixture model [31],

$$\sum_{j=1\ldots N} w_{i,t-1}^{(j)}\, p(\mathbf{s}_t\,|\,\mathbf{s}_{i,t-1}^{(j)})\,. \tag{6}$$

From this perspective, the prediction distribution in (5) amounts to a product of mixture models.

Neigh\Curr	$\mu_t = 0$	$\mu_t = 1$
$\mu_{t-1} = 0$	$1-p_{0\to 1}$	$p_{0\to 1}$
$\mu_{t-1} = 1$	$p_{1\to 0}$	$1-p_{1\to 0}$

Table 1. Transition probabilities $p_\mu(\mu_t|\mu_{t-1}, \mathbf{c}_{t-1})$.

However, because we typically use thousands of particles (e.g., $N = 10^3$), and about $\mathcal{N} = 5$ neighbors, the number of components in such a product, i.e., $N^{\mathcal{N}}$, quickly becomes unmanageable. We overcome this problem by fitting a mixture model to the individual prediction distributions prior to their multiplication in (5). We use mixture models with a small number of Gaussian components (often 3 to 5) plus a uniform outlier process. As a result, the product in (5) reduces to fewer than 10^3 components. The mixture models are fit with a simple version of the EM algorithm.

Note that we first propagate individual samples from the neighboring posteriors at the previous time, and then we fit the mixture model. As with *assumed density filtering* and *unscented filtering*, this is done because it is relatively easy to propagate individual samples through nonlinear dynamics. The final prediction distribution in (5) is obtained multiplying the individual mixture model predictions.

5 Computational Method

Given weighted sample sets that approximate the posterior distribution in each local region at time $t-1$, the steps toward the computation of the posterior distribution in a specific region at time t can be summarized as follows:

1. For each neighbor i at the previous time $t-1$:
   – Draw N samples with replacement from the posterior at time $t-1$ given by $\{\mathbf{s}_{i,t-1}^{(j)}, w_{i,t-1}^{(j)}\}_{j=1}^{N}$.
   – Propagate the samples using the model dynamics (Sec. 5.1), and then sample from the prediction density (6) to get a new sample set at time t.
   – Use EM to fit a mixture of M Gaussians and a uniform outlier component to the new sample set.
2. Take the product of the individual mixture model predictions to form the joint prediction distribution (5).
3. Draw N samples with replacement from this distribution and compute the likelihood of each sample.
4. Normalize the likelihoods to obtain the sample weights.

This yields a weighted sample set $\{\mathbf{s}_t^{(j)}, w_t^{(j)}\}$ that approximates the posterior for a region at time t, i.e., $p(\mathbf{s}_t\,|\,\bar{\mathbf{I}}_t)$.

5.1 Dynamical Model

Given (5), we need only to specify the form of the dynamics between a state $\mathbf{s}_t$ and a single neighbor $\mathbf{s}_{i,t-1}$ from the previous time. To this end it is useful to first expand the state $\mathbf{s}$ into its discrete and continuous components, μ and $\mathbf{c}$. We then rewrite the pair-wise prediction distribution as

$$p(\mathbf{s}_t|\bar{\mathbf{I}}_{t-1}) = \sum_{\mu_{t-1}} \int_{\mathbf{c}_{t-1}} [\,p(\mathbf{c}_t|\mu_t, \mu_{t-1}, \mathbf{c}_{t-1})\, p(\mu_t|\mu_{t-1}, \mathbf{c}_{t-1}) \\ p(\mu_{t-1}, \mathbf{c}_{t-1}\,|\,\bar{\mathbf{I}}_{t-1})\,]\,, \tag{7}$$

Neigh\Curr	$\mu_t = 0$	$\mu_t = 1$
$\mu_{t-1} = 0$	$p_{\mathbf{c}}(\mathbf{u}_t \vert \mathbf{u}_{t-1}) = \mathcal{N}(\mathbf{u}_{t-1}, \sigma_u^2 \mathbf{I})$	$p_{\mathbf{c}}(\mathbf{c}_{e,t} \vert \mathbf{u}_{t-1}) = p(\theta_t, d_t)\, p(\mathbf{u}_{f_t} \vert \mathbf{u}_{t-1})\, p(\mathbf{u}_{b_t} \vert \mathbf{u}_{t-1})$ where $p(\theta_t, d_t) = \mathrm{Uniform}(\theta_t, d_t \vert \text{edge outside neighbor})$ if $\mathbf{x}_{t-1}$ is in current foreground $p(\mathbf{u}_{f_t} \vert \mathbf{u}_{i,t-1}) = \mathcal{N}(\mathbf{u}_{t-1}, \sigma_u^2 \mathbf{I})$; $p(\mathbf{u}_{b_t} \vert \mathbf{u}_{i,t-1}) = \mathcal{N}(\mathbf{0}, 50\mathbf{I})$ (broad prior); else $p(\mathbf{u}_{f_t} \vert \mathbf{u}_{i,t-1}) = \mathcal{N}(\mathbf{0}, 50\mathbf{I})$ (broad prior); $p(\mathbf{u}_{b_t} \vert \mathbf{u}_{i,t-1}) = \mathcal{N}(\mathbf{u}_{t-1}, \sigma_u^2 \mathbf{I})$;

$\mu_{t-1} = 1$	if $\mathbf{x}_t$ is in neighbor's foreground
	$p_{\mathbf{c}}(\mathbf{u}_t \vert \mathbf{c}_{e,t-1}) = \mathcal{N}(\mathbf{u}_{f_{t-1}}, \sigma_u^2 \mathbf{I})$
	else $p_{\mathbf{c}}(\mathbf{u}_t \vert \mathbf{c}_{e,t-1}) = \mathcal{N}(\mathbf{u}_{b_{t-1}}, \sigma_u^2 \mathbf{I})$
		$p_{\mathbf{c}}(\mathbf{c}_{e,t} \vert \mathbf{c}_{e,t-1}) = f(\mathbf{c}_{e,t} \vert \mathbf{c}_{e,t-1}) \mathbb{1}(\vert d_t \vert < R) \,/\, (1 - p_{1\to 0})$
		where
		$\mathbb{1}(\vert d_t \vert < R) = 1$ when $\vert d_t \vert < R$, and 0 otherwise

Table 2. Model dynamics, $p_{\mathbf{c}}(\mathbf{c}_t | \mu_t, \mu_{t-1}, \mathbf{c}_{t-1})$, for the continuous parameters, conditioned on the discrete motion classes. Here, $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ are the centers of the current and neighbor regions at times t and $t-1$. The variances, $\sigma_u^2, \sigma_\theta^2$, and σ_d^2, control the process noise in the dynamics; we let each of them increase as a function of the spatial distance between the region centers $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$. We have omitted the dependence on the neighbor (i) for notational simplicity.

where $p(\mu_t | \mu_{t-1}, \mathbf{c}_{t-1})$ and $p(\mathbf{c}_t | \mu_t, \mu_{t-1}, \mathbf{c}_{t-1})$ denote the discrete and continuous transition distributions. To avoid singularities (where probabilities go to 0) and to allow for modeling errors in the dynamics, we let both distributions be robust; i.e., we use

$$p(\mu_t | \mu_{t-1}, \mathbf{c}_{t-1}) = \alpha p_\mu(\mu_t | \mu_{t-1}, \mathbf{c}_{t-1}) + (1-\alpha) p_{\mu,0}$$
$$p(\mathbf{c}_t | \mu_t, \mu_{t-1}, \mathbf{c}_{t-1}) = \beta p_{\mathbf{c}}(\mathbf{c}_t | \mu_t, \mu_{t-1}, \mathbf{c}_{t-1}) + (1-\beta) p_{\mathbf{c},0}$$

where $p_{\mu,0}$ and $p_{\mathbf{c},0}$ are uniform outlier distributions for discrete and continuous state variables, with mixing probabilities α and β. The inlier dynamics, $p_\mu(\mu_t | \mu_{t-1}, \mathbf{c}_{t-1})$ and $p_{\mathbf{c}}(\mathbf{c}_t | \mu_t, \mu_{t-1}, \mathbf{c}_{t-1})$, are summarized in Tables 1 and 2.

Referring to Table 1, we assume there is a spontaneous rate at which smooth motion regions encounter motion boundaries, denoted by $p_{0\to 1}$. For motion boundaries we assume a simple generative model for the dynamics in which an edge propagates forward with the foreground velocity, but otherwise there is mean-zero Gaussian process noise in the velocities and the boundary orientation. More precisely, with respect to a region $\mathbf{x}_t$ at time t, a motion boundary, parameterized with respect to a neighbor $\mathbf{x}_{t-1}$, is propagated to states distributed according to

$$f(\mathbf{c}_{e,t} | \mathbf{c}_{e,t-1}) \; = \; \mathcal{N}((\mathbf{u}_{f_{t-1}}, \mathbf{u}_{b_{t-1}}), \sigma_u^2 \mathbf{I}) \; \mathcal{N}(\theta_{t-1}, \sigma_\theta^2) \; \mathcal{N}(loc(\mathbf{c}_{e,t-1}), \sigma_d^2) \qquad (8)$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal density with mean μ and variance σ^2, $loc(\mathbf{c}_{e,t-1}) \equiv d_{t-1} + (\mathbf{u}_{f_{t-1}} + \mathbf{x}_{t-1} - \mathbf{x}_t) \cdot \hat{\mathbf{n}}_{t-1}$ is the mean edge location at time t relative to $\mathbf{x}_t$, and $\hat{\mathbf{n}}_{t-1} = (\sin(\theta_{t-1}), \cos(\theta_{t-1}))$. We then define the probability of changing from an edge state at $\mathbf{x}_{t-1}$ to a smooth motion state at $\mathbf{x}_t$ as the probability of the edge moving to points that do not intersect the region at $\mathbf{x}_t$:

$$p_{1\to 0} = \int_{|d_t| > R} \mathcal{N}_{d_t}(loc(\mathbf{c}_{e,t-1}), \sigma_d^2) \qquad (9)$$

where $\mathcal{N}_{d_t}$ is the PDF for d_t, and R is the region radius.

Table 2 defines the continuous prediction distributions conditioned on the discrete motion classes. For example, if the neighbor state at time $t-1$ and the current state are both smooth motions, then the current velocity is normally distributed about the velocity of the previous state. If the previous state was a motion boundary and the current state is smooth, then the current velocity is normally distributed about the foreground or background velocity, depending on whether the current region is on the foreground or background side of the previous region. If the previous state was smooth and the current state is a boundary, then θ_t and d_t are uniformly distributed over values for which the edge does not intersect the previous region, and the velocity distributions depend on the previous velocity state. Finally, if previous and current states are motion boundaries, then the distribution over the current state is Gaussian, but only for parameters such that the edge intersects the current region.

This dynamical model is applied to individual particles. The nonlinear components of the dynamics include the model switching and the computation of the propagated edge distance, which depends on the normal to the edge direction $\hat{\mathbf{n}}_{t-1} = (\sin(\theta_{t-1}), \cos(\theta_{t-1}))$. Nonlinearities make it difficult to propagate distributions analytically, even if the neighbor posterior at the time $t-1$ had been Gaussian.

5.2 Likelihood Function

Given a set of states drawn from the prediction distribution, the weights for a particle approximation to the posterior are proportional to the likelihoods, $p(\mathbf{I}_t \mid \mathbf{s}_t^{(j)}, \mathbf{I}_{t-1})$ for $j = 1...N$. Here, we factor the likelihood into a motion likelihood that depends on intensity differences between frames at times t and $t-1$, and an edge likelihood that depends solely on the band-pass image properties at time t.

The un-normalized motion likelihood is given by

$$p_m(\mathbf{I}_t \mid \mathbf{s}_t^{(j)}, \mathbf{I}_{t-1}) = \left(\exp\left[\frac{-1}{2\sigma_n^2} \sum_{\mathbf{x} \in \mathcal{R}} D(\mathbf{x}, t; \mathbf{s}_t^{(j)}) \right] \right)^{\frac{2}{T}} \qquad (10)$$

where $D(\mathbf{x}, t;\, \mathbf{s}_t^{(j)}) = [I(\mathbf{x}(\mathbf{s}_t^{(j)}), t) - I(\mathbf{x}, t-1)]^2$, $T = |\mathcal{R}|$ is the number of pixels in the circular image region, and $\mathbf{x}(\mathbf{s}_t^{(j)})$ denotes the warped image coordinates that depend

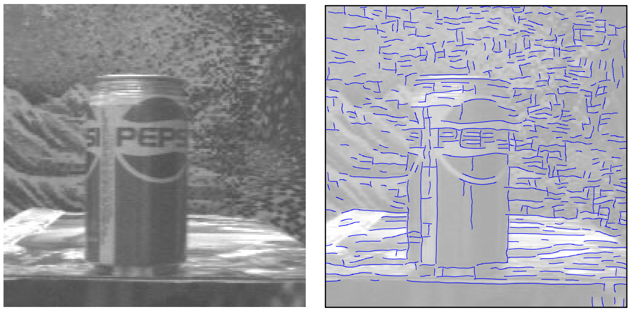

Figure 3. This shows a single image from the pepsi sequence along with the dominant level phase contours at $\pm\pi/2$ at orientations near vertical and horizontal.

on the deformation encoded by $\mathbf{s}_t^{(j)}$. The warped values $I(\mathbf{x}(\mathbf{s}_t^{(j)}), t)$ are computed with bi-linear interpolation. This motion likelihood is derived from a generative model based on brightness constancy and I.I.D. Gaussian image noise. The exponent of $2/T$, however, is computationally, rather than probabilistically, motivated. A large value of T effectively broadens the peaks of the likelihood. With a particle filter, this allows a more effective search of the parameter space, reducing the chances of missing a significant peak.

In addition to motion, we also exploit static edge information. Not all static edges are motion boundaries, but because motion boundaries are generally caused by depth discontinuities, they often coincide with static edges (see Fig. 3). Static edge information also improves the inference of motion boundaries because the correct foreground/background assignment depends on accurate prediction of the edge locations through time.

We chose the edge likelihood to be the observation density over the responses of an oriented band-pass filter tuned to the edge orientation (cf. [26]). This removes all oriented image structure except that near the orientation of the edge. To do this efficiently for many edges we first apply a steerable pyramid transform to the image. From the steerable basis set we can quickly compute responses of filters tuned to any orientation. Here, we use the (G_2, H_2) quadrature-pair filters defined in [9]. These are complex-valued filters so we express their response at each (subsampled) spatial location in terms of amplitude and phase [8]. The edge likelihood is simply the observation density over phase and amplitude of the subsampled filters responses at points along the edge.

However, modeling the observation density is not simple because the appearance of edges at surface boundaries varies significantly with surface reflectance properties and local illumination. Rather than attempt to design an edge model that captures the variability of edge appearance from first principles, we develop an empirical likelihood from the statistics of natural images.

We identified 800 surface boundaries by hand in 25 images. Band-pass filters were steered to each edge orientation. We then extracted phase ϕ and amplitude a responses along each edge with a sampling distance of one wavelength of the filters' tuning frequency. This sparse sampling allows us to treat measurements at different locations as conditionally independent. The resulting ensemble of phase and amplitude measurements exhibits a striking regularity that suggests a factorization of the joint observation density:

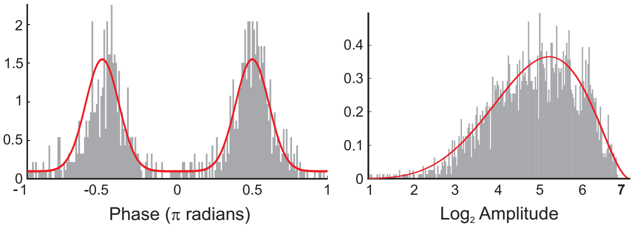

Figure 4. Histograms are shown of (left) phase conditioned on amplitude and the edge, and of (right) log amplitude conditioned on the edge.

$$p_e(\phi, a \,|\, \mathbf{s}) \;=\; p_\phi(\phi \,|\, a,\, \mathbf{s})\, p_a(a \,|\, \mathbf{s}) \;. \qquad (11)$$

As shown in Fig. 4(left), phase responses, ϕ, are typically close to $\pm\pi/2$, depending on sign of the intensity gradient at the edge. These conditional phase distributions are very well described by a mixture of two Gaussian modes at $\pi/2$ and $-\pi/2$, and a uniform outlier density. A maximum likelihood fit of this model to the data with the EM algorithm is shown as the solid curve in Fig. 4A. Wrapping phase about π gives yields the equivalent density for $\psi \equiv (\phi \bmod \pi)$:

$$p_\psi(\psi \,|\, a,\, \mathbf{s}) \;=\; m(a)\, G(\psi;\, \frac{\pi}{2}, \sigma^2) + (1 - m(a)) p_0 \;. \quad (12)$$

where $p_0 = 1/\pi$ is the phase outlier probability, and m is the Gaussian mixing probability.

With this mixture model (12), we find that the mixing proportion m depends significantly on log amplitude; phase becomes more stable with increasing amplitude [8]. Using a Bayesian model selection criteria we find that a good model for the phase density is the mixture in (12), with the standard deviation of the Gaussian held fixed at $\pi/8$, and the mixing probability $m(a)$ given by $m(a) = 0.1(1.25 + \log a)$ where $8 > \log a > 0$ on 8 bit images.

To model response amplitude, we find that a beta distribution fits the conditional distribution of log amplitude well (e.g. see Fig. 4(right)). The beta distribution is a natural choice since it is defined on a finite interval, appropriate for images with a limited dynamic range, and it provides a reasonable approximation to a Gaussian.

Our edge-based likelihood is given by the factorization in (11), along with the parametric models for the phase and amplitude densities. Given a set of K phase and amplitude measurements, conditioned on a motion boundary state, $\mathbf{s}_t^{(j)}$, the joint likelihood is

$$p_e(\{\psi_k, a_k\}_t \,|\, \mathbf{s}_t^{(j)}) = \left(\prod_k p_\psi(\psi_k \,|\, a_k,\, \mathbf{s}_t^{(j)})\, p_a(a_k \,|\, \mathbf{s}_t^{(j)}) \right)^{\frac{1}{K}}$$

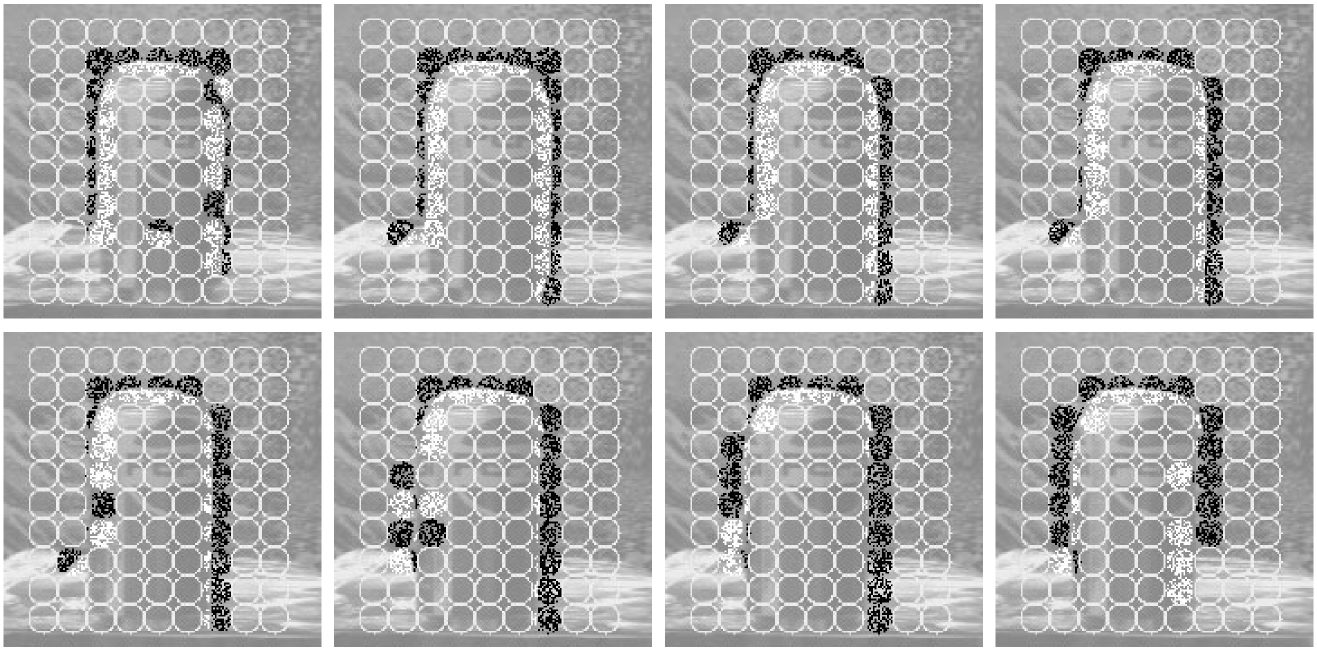

Figure 5. Pepsi results for frames 3–10 (in lexicographic order and cropped slightly for the visibility of the detected boundaries).

When the state is a smooth motion model, the observation density is taken to be uniform.

6 Experimental Results

We demonstrate our approach with the well-known pepsi and flower garden image sequences. We use circular regions with radii of 8 and 12 pixels which overlap by 2 and 3 pixels. We use 5000 samples for particle approximations in each region. We draw 10% of the particles from the initialization prior in [4], and the remaining 90% from the prediction density in (5). The parameters for the dynamics between a location at time t and a neighbor at time $t-1$ depend on the spatial separation between the two locations. For the same spatial location at t and $t-1$ we use $\sigma_u = 0.75$ pixel/frame, $\sigma_\theta = 0.1$ radians, and $\sigma_d = 1$ pixel. For an adjacent region at $t-1$ we use $\sigma_u = 1.5$ pixels/frame, $\sigma_\theta = 0.2$ radians, and $\sigma_d = 1.5$. In both cases $\alpha = 0.975$ and $\beta = 0.95$. Finally, the probability of a motion boundary, conditioned on the motion of a neighboring being smooth, is $p_{0\rightarrow 1} = 0.4$; this value reflects the fact that edges occur in roughly 10% of the image regions, and that such motion boundary predictions are relatively unconstrained, requiring a large number of samples to search the state space effectively.

For display, we use a straightforward Bayesian model selection criterion to decide among 3 motion models, namely smooth motion and the two foreground/background assignments associated with the dominant local orientation in the motion boundary model. Regions where smooth motion is most probable are displayed as empty circles. Filled regions depict motion boundaries, the white/black dots of which lie on the foreground/background. For these motion boundaries we only show the mean position and orientation of the boundary. Note that when the distributions are skewed or multimodal the mean does not necessarily reflect how well the distribution captures the underlying motion.

Figure 5 shows results from frames 3–10 of the pepsi sequence. Compared to the results of [4] in Fig. 1, with the same parameters and both motion and edge-based likelihoods, the current method produces more coherent boundary estimates. Noteworthy in Fig. 5 are the correct assignment of the foreground and the accurate localization of the motion boundaries. Also evident in Fig. 5, is the importance of the neighborhood propagation that allows regions to anticipate the arrival of a boundary from a neighboring region. This is evident in frames 7–9 on the left boundary and later in frames 9–10 on the right side. This propagation allows the correct assignment of the foreground to be infered quickly, unlike the results of [4] which required 2 or more frames to correctly estimate relative depths.

Figure 6 shows results obtained with frames 15–20 of the flower garden sequence. Like the pepsi sequence above, the different regions along the edge of the tree depict good detection and localization of the motion boundary. The foreground is usually assigned correctly, and the evidence of information propagation can be seen in how well the edge is tracked from region to region. The plots in Fig. 6 show marginal distributions of the boundary location parameter, d, from the prediction (dashed) and the posterior (solid) distributions. Inset in each plot is the total probability for the motion boundary model. In frame 15 (leftmost plot) the probability of the edge model is high, but the broad marginal indicates high uncertainty in the edge location, as the edge just entered this region. As time progress the edge

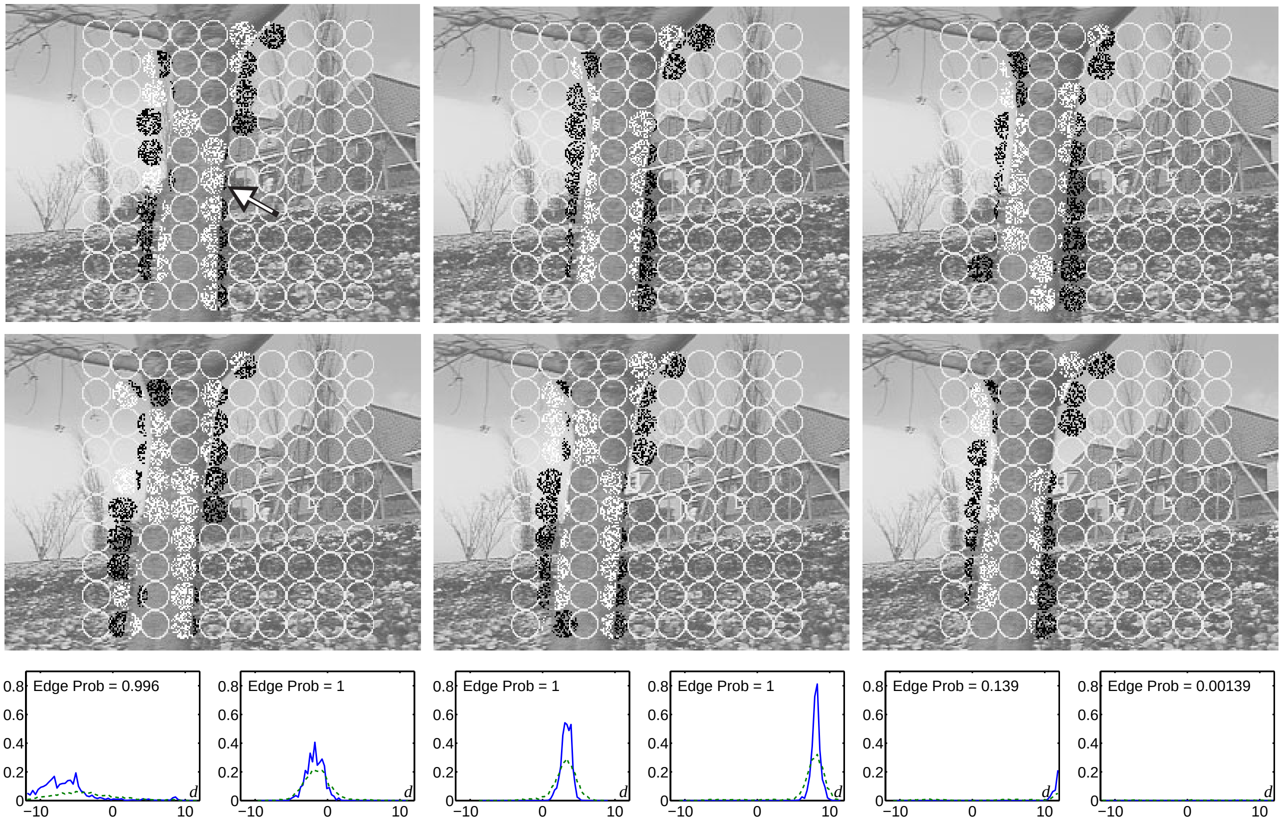

Figure 6. Flower garden results for frames 15–20 (in lexicographic order and cropped for visibility). The bottom plots show the marginal prediction (dashed) and posterior (solid) distributions for d, at each frame, for the region marked by the arrow in frame 15.

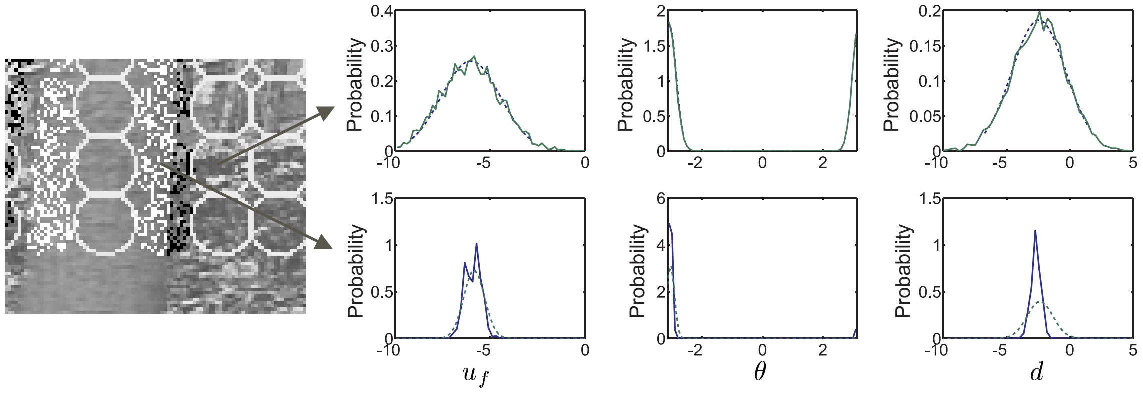

Figure 7. Detailed flower garden results at frame 10 for the central region in the image. (top plots) Marginal prediction densities obtained from the right neighbor are shown for the foreground velocity, the edge orientation, and the edge location. Solid curves show the prediction samples and the dashed curves show the mixture model approximation. (bottom plots) Marginal densities for the joint prediction density from all neighbors are depicted as dashed curves. The posterior marginals are depicted by solid curves.

moves leftward and the location uncertainty decreases as the marginal narrows. When the edge leaves the region in frame 19 the edge probability drops significantly.

Finally, Fig. 7 shows marginal distributions from the prediction densities, the mixture models, and the posterior for a single region at frame 10 of the flower garden sequence. In this case, the edge has just moved into the central region (Fig. 7 (left)), from the region to its right. While the region to its right is best explained by smooth motion at the current time, it was a motion boundary model at the previous frame. It was therefore able to tell the center region to its left to anticipate the occurrence of a boundary. This helps the central region to quickly adapt to the new situation in which an edge is entering the region, and to accurately estimate the boundary position, and to correctly assign the foreground side. The 3 top plots in Fig. 7 (right) show marginal distributions of the prediction density that the region on the right contributes to the central region (solid curve – sampled prediction; dashed curve – mixture model fit). The bottom plots in Fig. 7 show the joint spatiotemporal prediction density (dashed curve) from all neighbors, along with the resulting posterior distribution (solid curve). This shows the advantage of having predictions from neighboring regions.

7 Conclusions

Particle filters are effective for visual tracking, allowing for a Bayesian framework even with non-Gaussian distributions and non-linear dynamics. Here we extend their use, in conjunction with other methods for approximate inference, to the detection and estimation of multiple motion models defined over a random field. In particular, we consider the detection and tracking of motion boundaries for which causal predictions of motion and of boundary locations/orientations are obtained from nearby image regions at the previous time. This helps to encourage boundary continuity, and to direct samples to the appropriate regions of the state as an edges leaves one region and enters another. It also improves the inference of surface depth ordering.

There remain several unresolved research issues concerning the assumptions made here to simplify the mathematical analysis and the implementation. Perhaps most significant is the use of a factored approximation to the posterior distribution over all regions by its individual region marginals. Future work should examine this in connection with recent results on Bayesian belief propagation [19, 29].

Acknowledgments

Thanks to R. Luo for helping to develop the edge likelihood, to A. Jepson for his steerable pyramid code, and to M. Black and A. Jepson for helpful discussions.

References

[1] S. Ayer and H. Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. *ICCV*, pp. 777–784, 1995

[2] S. S. Beauchemin and J. L. Barron. The local frequency structure of 1d occluding image signals. *IEEE PAMI*, 22:200–206, 2000

[3] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63:75–104, 1996

[4] M. J. Black and D. J. Fleet. Probabilistic detection and tracking of motion discontinuities. *IJCV*, 38:229–243, 2000

[5] K. Choo and D. J. Fleet. People tracking with hybrid Monte Carlo. *Proc. IEEE ICCV*, Vol II, pp. 321-328, 2001

[6] N. Cornelius and T. Kanade. Adapting optical flow to measure object motion in reflectance and X-ray image sequences. *Proc. ACM Work. Motion*, pp. 50–58, 1981

[7] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealled particle filtering. *Proc. IEEE CVPR*, Vol II, pp. 126–133, 2000

[8] D. J. Fleet and A. D. Jepson. Stability of phase information. *IEEE PAMI*, 15:1253–1268, 1993

[9] W. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE PAMI*, 13:891–906, 1991

[10] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc.-F*, 140:107–113, 1993

[11] J. G. Harris, C. Koch, E. Staats, and J. Luo. Analog hardware for detecting discontinuities in early vision. *IJCV*, 4:211–223, 1990

[12] F. Heitz and P. Bouthemy. Multimodal motion estimation of discontinuous optical flow using Markov random fields. *IEEE PAMI*, 15:1217–1232, 1993

[13] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998

[14] A. Jepson and M. J. Black. Mixture models for optical flow computation. *Proc. IEEE CVPR*, pp. 760–761, 1993

[15] G. Kitagawa. Non-gaussian state-space modelling of nonstationary time series (with discussion). *J. Amer. Stat. Assoc.*, 82:1032–1063, 1987

[16] J. Konrad and E. Dubois. Multigrid Bayesian estimation of image motion fields using stochastic relaxation. *Proc. IEEE ICCV*, pp. 354–362, 1998

[17] J. S. Liu and R. Chen. Sequential monte carlo methods for dynamics systems. *J. Amer. Stat. Ass.*, 93:1031–1044, 1998

[18] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. ECCV-00, Dublin, V II, pp. 3–19

[19] K. Murphy and Y. Weiss. The factored frontier algorithm for approximate inference in DBNs. *Proc. UAI*, pp. 378–385, 2001

[20] K. Mutch and W. Thompson. Analysis of accretion and deletion at boundaries in dynamic scenes. *IEEE PAMI*, 7:133–138, 1985

[21] H. H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE PAMI*, 8:565–593, 1986

[22] S. A. Niyogi. Detecting kinetic occlusion. *Proc. IEEE ICCV*, pp. 1044–1049, Boston, 1995

[23] H. S. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE PAMI*, 18:814–831, 1996

[24] B. G. Schunck. Image flow segmentation and estimation by constraint line clustering. *IEEE PAMI*, 11:1010–1027, 1989

[25] D. Shulman and J. Hervé. Regularization of discontinuous flow fields. *Proc. IEEE Work. Vis. Motion*, pp. 81–85, 1989

[26] H. Sidenbladh and M. Black. Learning image statistics for Bayesian tracking. *Proc. ICCV*, Vol. II, pp. 709–1716, 2001

[27] A. Spoerri and S. Ullman. The early detection of motion boundaries. *Proc. IEEE ICCV*, pp. 209–218, 1987

[28] W. Thompson, K. Mutch, and V. Berzins. Dynamic occlusion analysis in optical flow fields. *PAMI*, 7:374–383, 1985

[29] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Comp.*, 12:1–41, 2000

[30] Y. Weiss and E. Adelson. Unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. *CVPR*, pp. 321–326, 1996

[31] M. West. Mixture models, Monte Carlo, Bayesian updating and dynamic models. *Comp. Sci. & Stat.*, 24:325–333, 1992