Probabilistic Detection and Tracking of Motion Discontinuities

Michael J. Black **David J. Fleet**

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

{black,fleet}@parc.xerox.com
http://www.parc.xerox.com/{black,fleet}/

Abstract

We propose a Bayesian framework for representing and recognizing local image motion in terms of two primitive models: translation and motion discontinuity. Motion discontinuities are represented using a non-linear generative model that explicitly encodes the orientation of the boundary, the velocities on either side, the motion of the occluding edge over time, and the appearance/disappearance of pixels at the boundary. We represent the posterior distribution over the model parameters given the image data using discrete samples. This distribution is propagated over time using the Condensation algorithm. To efficiently represent such a high-dimensional space we initialize samples using the responses of a low-level motion discontinuity detector.

1 Introduction

Motion discontinuities provide information about the position and orientation of surface boundaries in a scene. Additionally, analysis of the occlusion/disocclusion of pixels at a motion boundary provides information about the relative depth ordering of the neighboring surfaces. While these properties have made the detection of motion discontinuities an important problem in computer vision, experimental results have been somewhat disappointing. As discussed below, previous approaches have treated motion discontinuities as "noise" (violations of spatial smoothness) or have used approximate models of the motion discontinuities.

In this paper we formulate a generative model of motion discontinuities as illustrated in Figure 1. The model includes the orientation of the boundary, the velocities of the surfaces on either side, the foreground/background assignment, and an offset of the boundary from the center of the region. With this explicit model, we can predict the visibility of occluded and disoccluded pixels so that these pixels can be excluded when estimating the probability of a particular model. Moreover, an explicit displacement parameter

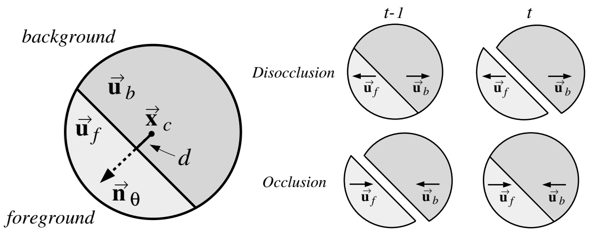


Figure 1: Model of an occlusion boundary, parameterized by foreground and background velocities, $\vec{\mathbf{u}}_f$ and $\vec{\mathbf{u}}_b$, an orientation θ with normal $\vec{\mathbf{n}}_\theta$, and a signed distance d from the neighborhood center $\vec{\mathbf{x}}_c$. With this model we predict which pixels are visible between frames at times t and $t-1$.

allows us to predict the location of the edge, and hence track its movement through a region of interest. Tracking the motion of the edge allows foreground/background ambiguities to be resolved.

Explicit generative models such as this have not previously been used for detecting motion discontinuities due to the non-linearity of the model and the difficulty of estimating the model parameters. To solve this problem we exploit a probabilistic sampling-based method for estimating image motion [6]. Adopting a Bayesian framework, we define the likelihood of observing the image data given the parameters of the generative model. This likelihood distribution can be efficiently evaluated for a particular set of parameters. The prior probability distribution over the parameters is defined as a mixture of a temporal prior and an initialization prior. The temporal prior is defined in terms of the posterior distribution at the previous time instant and the temporal dynamics of the discontinuity model. The initialization prior incorporates predictions from a low-level motion feature detector [8]. The posterior distribution over the parameter space, conditioned on image measurements, is typically non-Gaussian. The distribution is represented using factored sampling and is predicted and updated over time using the Condensation algorithm to propagate conditional probability densities [12].

Given the relatively high dimensional parameter space,

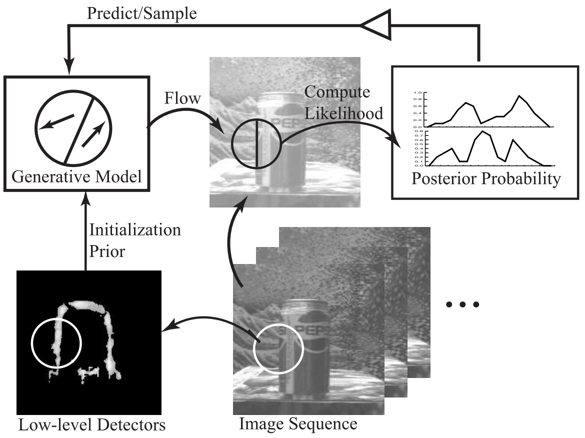


Figure 2: Multi-stage probabilistic model. Low-level detectors help initialize a sampled distribution. Likelihoods are computed directly from pairs of images. The Condensation algorithm is used to incrementally predict and update the posterior distribution over model parameters.

naive sampling methods will be extremely inefficient. But if the samples can be directed to the appropriate portion of the parameter space, small numbers of samples can well characterize such distributions [14]. It is for this purpose that we use an initialization prior (as shown in Figure 2). At the low level, there is a set of dense motion discontinuity detectors that signal the presence of potential occlusion boundaries and give estimates of their orientations and velocities. In the example here, these detectors are based on approximate linear models of motion discontinuities, the coefficients of which can be estimated with robust optical flow techniques [8]. Neighborhoods of these filter outputs provide a *prior* distribution over model parameters that is sampled from when initializing non-linear models at the higher level. The likelihood of the non-linear model is then computed directly from the image data.

We illustrate the method on natural images and show how the Bayesian formulation and conditional density propagation allow motion discontinuities to be detected and tracked over multiple frames. Explicit image feature models, combined with a Bayesian formulation, and the Condensation algorithm offer a new set of tools for image motion estimation and interpretation.

2 Previous Work

Previous approaches for detecting occlusion boundaries have often treated the boundaries as a form of "noise", that is, as the violation of a smoothness assumption. This approach is taken in regularization schemes where robust statistics, weak continuity, or line processes are used to disable smoothing across motion discontinuities [4, 11]. Similarly, parameterized models of image motion (e.g. translational, affine, or planar) assume that flow is represented by a low-order polynomial. Robust regression [4, 19] and mixture models [1, 15, 23] have been used to account for the multiple motions that occur at motion boundaries but these methods fail to explicitly model the boundary and its spatiotemporal structure.

Numerous methods have attempted to detect discontinuities in optical flow fields by analyzing local distributions of flow [21] or by performing edge detection on the flow field [18, 20, 22]. It has often been noted that these methods are sensitive to the accuracy of the optical flow and that accurate optical flow is hard to estimate without prior knowledge of the occlusion boundaries. Other methods have focused on detecting occlusion from the structure of a correlation surface [3], or of the spatiotemporal brightness pattern [7, 9, 17]. Still others have used the presence of unmatched features to detect dynamic occlusions [16].

None of these methods explicitly model the image motion present at a motion feature, and have not proved sufficiently reliable in practice. For example, they do not explicitly model which image pixels are occluded or disoccluded between frames. This means that these pixels, which in one frame have no "match" in the next frame, are treated as noise. With our explicit non-linear model, these pixels can be predicted and ignored.

Additionally, most of the above methods have no explicit temporal model. With our generative model, we can predict the motion of the occlusion boundary over time and hence integrate information over a number of frames. When the motion of the discontinuity is consistent with that of the foreground we can explicitly determine the foreground/background relationships between the surfaces.

3 Generative Model

For the purposes of this work, we decompose an image into a grid of circular neighborhoods in which we estimate motion information. We assume that the motion in any region can be modeled by translation (for simplicity) or by dynamic occlusion. Generative models of these motions are used to compute the likelihood of observing two successive images given a motion model and its parameter values.

The translation model has two parameters, i.e., the horizontal and vertical components of the velocity, denoted $\vec{\mathbf{u}}_0 = (u_0, v_0)$. For points $\vec{\mathbf{x}}$ at time t in a region R, assuming brightness constancy, the translation model is

$$I(\vec{\mathbf{x}}', t) = I(\vec{\mathbf{x}}, t-1) + \nu(\vec{\mathbf{x}}, t), \quad (1)$$

where $\vec{\mathbf{x}}' = \vec{\mathbf{x}} + \vec{\mathbf{u}}_0$. In words, the intensity at location $\vec{\mathbf{x}}'$ at time t is equal to that at location $\vec{\mathbf{x}}$ at time $t-1$ plus noise ν. We assume here that the noise is white and Gaussian with a mean of zero and a standard deviation of σ_n.

The occlusion model contains 6 parameters: the edge orientation, the two velocities, and the distance from the center of the neighborhood to the edge. In our parameterization, as shown in Figure 1, the orientation, $\theta \in [-\pi, \pi)$, specifies the direction of a unit vector, $\vec{\mathbf{n}} = (\cos(\theta), \sin(\theta))$, that is normal to the occluding edge. We represent the location of the edge by its signed perpendicular distance d from the center of the region (positive meaning in the direction of the normal). Thus, the edge is normal to $\vec{\mathbf{n}}$ and passes through the point $\vec{\mathbf{x}}_c + d\,\vec{\mathbf{n}}$, where $\vec{\mathbf{x}}_c$ is the center of the region.

Relative to the center of the region, we define the foreground to be the side to which the normal $\vec{\mathbf{n}}$ points. Therefore, a point $\vec{\mathbf{x}}$ is on the foreground if $(\vec{\mathbf{x}} - \vec{\mathbf{x}}_c) \cdot \vec{\mathbf{n}} > d$. Similarly, points on the background satisfy $(\vec{\mathbf{x}}-\vec{\mathbf{x}}_c)\cdot\vec{\mathbf{n}} < d$. Finally, we denote the velocities of the foreground (occluding) and background (occluded) sides by $\vec{\mathbf{u}}_f$ and $\vec{\mathbf{u}}_b$.

Assuming that the occluding edge moves with the foreground velocity, the occurrences of occlusion and disocclusion depend solely on the difference between the background and foreground velocities. In particular, occlusion occurs when the background moves faster than the foreground in the direction of the edge normal. If $u_{fn} = \vec{\mathbf{u}}_f \cdot \vec{\mathbf{n}}$ and $u_{bn} = \vec{\mathbf{u}}_b \cdot \vec{\mathbf{n}}$ denote the two normal velocities, occlusion occurs when $u_{bn} - u_{fn} > 0$. Disocclusion occurs when $u_{bn} - u_{fn} < 0$. The width of the occluded/disoccluded region, measured normal to the occluding edge, is $|u_{bn} - u_{fn}|$.

With this model, parameterized by $(\theta,\, \vec{\mathbf{u}}_f,\, \vec{\mathbf{u}}_b,\, d)$, a pixel $\vec{\mathbf{x}}$ at time $t-1$ moves to location $\vec{\mathbf{x}}'$ at time t, as follows:

$$\vec{\mathbf{x}}' = \begin{cases} \vec{\mathbf{x}} + \vec{\mathbf{u}}_f & \text{if } (\vec{\mathbf{x}} - \vec{\mathbf{x}}_c) \cdot \vec{\mathbf{n}} > d \\ \vec{\mathbf{x}} + \vec{\mathbf{u}}_b & \text{if } (\vec{\mathbf{x}} - \vec{\mathbf{x}}_c) \cdot \vec{\mathbf{n}} < d + w \end{cases} \tag{2}$$

where $w = \max(u_{bn} - u_{fn},\, 0)$ is the width of the occluded region. Finally, with $\vec{\mathbf{x}}'$ defined by (2), the brightness constancy assumption for a motion edge is given by (1).

Referring to Figure 1, in the case of disocclusion, a circular neighborhood at time $t-1$ will map to a pair of regions at time t, separated by the width of the disocclusion region $|u_{bn} - u_{fn}|$. Conversely, in the case of occlusion, a pair of neighborhoods at time $t-1$, separated by $|u_{bn} - u_{fn}|$, map to a circular neighborhood at time t. Being able to look forward or backwards in time in this way allows us to treat occlusion and disocclusion symmetrically.

4 Probabilistic Framework

For a given image region and images up to time t, we wish to estimate the posterior probability distribution over models and model parameters at time t. This distribution is not directly observable and, as described below, we expect it to be multi-modal. It is discrete over the model types and continuous over model parameters.

Let *states* be denoted by $\mathbf{s} = (\mu, \vec{\mathbf{p}})$, where μ is the model type (translation or occlusion), and $\vec{\mathbf{p}}$ is a parameter vector appropriate for the model type. For the translation model $\vec{\mathbf{p}} = (\vec{\mathbf{u}}_0)$, and for the occlusion model $\vec{\mathbf{p}} = (\theta, \vec{\mathbf{u}}_f, \vec{\mathbf{u}}_b, d)$. Our goal is to find the posterior probability distribution over states at time t given the measurement history up to time t, i.e., $p(\mathbf{s}_t \mid \vec{\mathbf{Z}}_t)$. Here, $\vec{\mathbf{Z}}_t = (\mathbf{z}_t, ..., \mathbf{z}_0)$ denotes the measurement history. Similarly, let $\vec{\mathbf{S}}_t = (\mathbf{s}_t, ..., \mathbf{s}_0)$ denote the state history (a stochastic process).

Following [12], we assume that the temporal dynamics of the motion models form a Markov chain, in which case $p(\mathbf{s}_t \mid \vec{\mathbf{S}}_{t-1}) = p(\mathbf{s}_t \mid \mathbf{s}_{t-1})$. We also assume conditional independence of the observations and the dynamics, so that, given $\mathbf{s}_t$, the current observation $\mathbf{z}_t$ and previous observations $\vec{\mathbf{Z}}_{t-1}$ are independent. With these assumptions one can show that the posterior distribution $p(\mathbf{s}_t \mid \vec{\mathbf{Z}}_t)$ can be factored and reduced using Bayes' rule to

$$p(\mathbf{s}_t|\vec{\mathbf{Z}}_t) \;=\; k\, p(\mathbf{z}_t|\mathbf{s}_t)\, p(\mathbf{s}_t|\vec{\mathbf{Z}}_{t-1}) \tag{3}$$

where k is a constant factor to ensure that the distribution integrates to one. Here, $p(\mathbf{z}_t|\mathbf{s}_t)$ represents the likelihood of observing the current measurement given the current state, while $p(\mathbf{s}_t|\vec{\mathbf{Z}}_{t-1})$ is referred to as a temporal prior (the prediction of the current state given all previous observations).

According to the generative models discussed above (1), the likelihood of observing the current image pair given the current state is normally distributed. The current state defines a mapping from visible pixels in one frame to those in the next. The intensity difference between a corresponding pair of pixel locations is taken to be normally distributed with a mean of zero and a standard deviation of σ_n.

Using Bayes' rule and the conditional independence assumed above, it is straightforward to show that the temporal prior can be written in terms of temporal dynamics that propagate states from time $t-1$ to time t and the posterior distribution over states at time $t-1$. In particular,

$$p(\mathbf{s}_t|\vec{\mathbf{Z}}_{t-1}) \;=\; \int p(\mathbf{s}_t|\mathbf{s}_{t-1})\, p(\mathbf{s}_{t-1}|\vec{\mathbf{Z}}_{t-1})\, d\,\mathbf{s}_{t-1}. \tag{4}$$

The probability distribution $p(\mathbf{s}_t|\mathbf{s}_{t-1})$ embodies the temporal dynamics, describing how states evolve through time.

For now assume that the model type (i.e. translation or occlusion) remains constant between frames (this can be extended to allow transitions between model types [5, 13]). For the translational model, we assume the velocity at time t equals that at time $t-1$ plus Gaussian noise:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = G_{\sigma_u}(\Delta\vec{\mathbf{u}}_0) \tag{5}$$

where G_{σ_u} denotes a mean-zero Gaussian with standard deviation σ_u, and $\Delta\vec{\mathbf{u}}_0 = \vec{\mathbf{u}}_{0,t} - \vec{\mathbf{u}}_{0,t-1}$ denotes the temporal velocity difference.

Similarly, for the temporal dynamics of the occlusion model we assume that the expected orientation and velocities remain constant, while the location of the edge propagates with the velocity of the foreground. Moreover, independent noise is added to each. Therefore, we can express the conditional $p(\mathbf{s}_t|\mathbf{s}_{t-1})$ as

$$G_{\sigma_u}(\Delta\vec{\mathbf{u}}_f)\, G_{\sigma_u}(\Delta\vec{\mathbf{u}}_b)\, G_{\sigma_d}(\Delta d - \vec{\mathbf{n}} \cdot \vec{\mathbf{u}}_{f,t-1})\, G^{w}_{\sigma_\theta}(\Delta\theta) \tag{6}$$

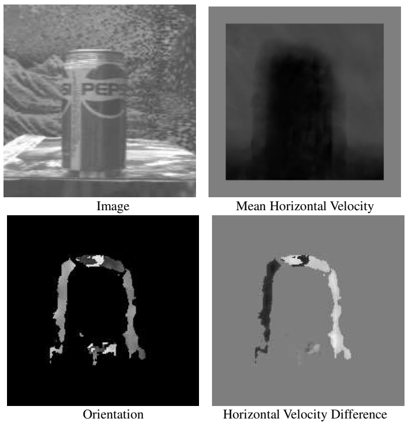


Figure 3: One frame of the Pepsi Sequence, with responses from the low-level motion edge detector, which feed the initialization prior. The motion is primarily horizontal.

where G^w denotes a wrapped-normal (for circular distributions), and as above, $\Delta\theta = \theta_t - \theta_{t-1}$ and $\Delta d = d_t - d_{t-1}$.

5 Low-Level Motion-Edge Detectors

Unlike a conventional Condensation tracker for which the prior is derived by propagating the posterior from the previous time, we have added an initialization prior that provides a form of bottom-up information to initialize new states. This is useful at time 0 when no posterior is available from a previous time instant. It is also useful to help avoid getting trapped at local maxima thereby missing the occurrence of novel events that might not have been predicted from the posterior at the previous time. This use of bottom-up information, along with the temporal prediction of Condensation, allows us to effectively sample the most interesting portions of the state-space.

To initialize new states and provide a distribution over their parameters from which to sample, we use a method described by Fleet *et al.* [8] for detecting motion discontinuities. This approach uses a robust, gradient-based optical flow method with a linear parameterized motion model. Motion edges are expressed as a weighted sum of basis flow fields, the coefficients of which are estimated using a area-based regression technique. Fleet *et al.* then solve for the parameters of the motion edge that are most consistent (in a least squares sense) with the linear coefficients.

Figure 3 shows an example of applying this method to an image sequence in which a Pepsi can translates horizontally relative to the background. The method provides a mean velocity estimate at each pixel (i.e., the average of the velocities on each side of the motion edge). This is simply the translational velocity when no motion edge is present. A confidence measure, $c(\vec{\mathbf{x}}) \in [0, 1]$ can be used to determine where edges are most likely, and is computed from the squared error in fitting a motion edge using the linear coefficients. The bottom two images in Figure 3 show estimates for the orientation of the edge and the horizontal difference velocity across the edge at all points where $c(\vec{\mathbf{x}}) > 0.5$.

While the method provides good approximate estimates of motion boundaries, it produces false positives and the parameter estimates are noisy, with estimates of disocclusion being more reliable than those of occlusion. Also, it does not determine which is the foreground side and hence does not predict the velocity of the occluding edge. Despite these weaknesses, it is a relatively quick, but sometimes error prone, source of information about the presence of motion discontinuities.

Initialization Prior. When initializing a new state we use the distribution of confidence values $c(\vec{\mathbf{x}})$ to first decide on the motion type (translation or discontinuity). If a discontinuity was detected, we would expect some fraction of confidence values, $c(\vec{\mathbf{x}})$, within our region of interest, to be high. We therefore rank order the confidence values within the region and let the probability of a discontinuity state be the 95^{th} percentile confidence value, denoted C_{95}. Accordingly, the probability of translation is then $1 - C_{95}$.

Given a discontinuity model, we assume that motion boundary locations are distributed according to the confidence values in the region (pixel locations with large $c(\vec{\mathbf{x}})$ are more likely). Given a spatial position, the detector at that position provides estimates of the edge orientation and the image velocity on each side, but does not specify which side is the foreground. Thus, the probability distribution over the state space, conditioned on the detector estimates and location, will have two distinct modes, one for each of the two possible foreground assignments. We take this distribution to be a mixture of two Gaussians which are separable with standard deviations $1.5\sigma_u$ for the velocity axes, $5\sigma_\theta$ for the orientation axis, and $2\sigma_d$ for the position axis. The standard deviations are larger than those use in the temporal dynamics described in Section 4 because we expect greater noise from the low-level estimates.

To generate a translational model, we choose a spatial position according to the distribution of $1 - c(\vec{\mathbf{x}})$. The distribution over translational velocities, given the detector estimate and spatial position, is then taken to be a Gaussian distribution centered at the mean velocity estimate of the detector at that location. The Gaussian distribution has standard deviations of $1.5\sigma_u$ along each velocity axis.

6 Computational Model

In this section we describe the computational embodiment of the probabilistic framework above. The non-linear nature of the discontinuity model means that $p(\mathbf{s}_t | \vec{\mathbf{Z}}_t)$ will not be Gaussian and we represent this distribution using a discrete set, $\{\mathbf{s}_t^{(i)}, i = 1, \ldots, S\}$, of random samples [6, 12, 24].

The posterior is computed by choosing discrete samples from the prior and then evaluating their likelihood. Normalizing the likelihoods of the samples so that they sum to one produces weights $\pi_t^{(n)}$:

$$\pi_t^{(n)} = \frac{p(\mathbf{z}_t | \mathbf{s}_t^{(n)})}{\sum_{i=1}^S p(\mathbf{z}_t | \mathbf{s}_t^{(i)})}.$$

The set of S pairs, $(\mathbf{s}_t^{(n)}, \pi_t^{(n)})$, provides a fair sampled representation of the posterior distribution as $S \to \infty$ [12].

6.1 Likelihood

To evaluate the likelihood $p(\mathbf{z}_t | \mathbf{s}_t^{(i)})$ of a particular state, we draw a uniform random sample $\mathcal{R}$ of visible image locations (as constrained by the generative model and the current state). Typically we sample 50% of the pixels in the region (cf. [2]). Given this subset of pixels, we compute the likelihood as

$$p(\mathbf{z}_t | \mathbf{s}_t^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[\frac{-1}{2\sigma_n^2 |\mathcal{R}|} \sum_{\vec{\mathbf{x}} \in \mathcal{R}} E(\vec{\mathbf{x}}, t; \mathbf{s}_t^{(i)})^2\right] \tag{7}$$

where $E(\vec{\mathbf{x}}, t; \mathbf{s}_t^{(i)}) = I(\vec{\mathbf{x}}\,', t) - I(\vec{\mathbf{x}}, t-1)$, $|\mathcal{R}|$ is the number of pixels in $\mathcal{R}$, and the warped image location $\vec{\mathbf{x}}\,'$ is a function of the state $\mathbf{s}_t^{(i)}$, e.g., as in (2). The warped image value $I(\vec{\mathbf{x}}\,', t)$ is computed using bi-linear interpolation.

6.2 Prior

The prior used here is a mixture of the temporal prior and the initialization prior. In the experiments that follow we use mixture proportions of 0.8 and 0.2 respectively; that is, 80% of the samples are drawn from the temporal prior.

Temporal Prior. We draw samples from the temporal prior by first sampling from the posterior, $p(\mathbf{s}_{t-1} | \vec{\mathbf{Z}}_{t-1})$, to choose a particular state $\mathbf{s}_{t-1}^{(n)}$. The discrete representation of the posterior means that this is done by constructing a cumulative probability distribution using the $\pi_{t-1}^{(n)}$ and then sampling from it. Given $\mathbf{s}_{t-1}^{(n)}$, we then sample from the dynamics, $p(\mathbf{s}_t | \mathbf{s}_{t-1}^{(n)})$, which, as explained above, is a normal distribution about the predicted state.

This is implemented using the following dynamics which propagate the state forward and add a sample of Gaussian noise:

$$\vec{\mathbf{u}}_t = \vec{\mathbf{u}}_{t-1} + \mathcal{N}(0, \sigma_u) \tag{8}$$

$$d_t = d_{t-1} + \vec{\mathbf{u}}_{f,t-1} + \mathcal{N}(0, \sigma_d) \tag{9}$$

$$\theta_t = [\theta_{t-1} + \mathcal{N}(0, \sigma_\theta)] \bmod 2\pi \tag{10}$$

where $\vec{\mathbf{u}}_t$ denotes any one of $\vec{\mathbf{u}}_{0,t}$, $\vec{\mathbf{u}}_{f,t}$, or $\vec{\mathbf{u}}_{b,t}$. With a wrapped-normal distribution over angles, the orientation θ_{t-1} is propagated by adding Gaussian noise and then removing an integer multiple of 2π so that $\theta_t \in [-\pi, \pi)$. This sampling process has the effect of diffusing the parameters of the states over time to perform a local search of the parameter space.

Initialization Prior. From the initialization prior for a particular neighborhood, discontinuity states are drawn with probability C_{95}. In these instances, we sample from the discrete set of confidence values, $c(\vec{\mathbf{x}})$, to choose a spatial position for the motion edge within the region. Then, as explained above, since the assignment of foreground/background is ambiguous, the distribution of the remaining parameters is taken to be an equal mixture of two Gaussians from which a sample is drawn.

Translational models are drawn with probability $1 - C_{95}$, from which we sample a spatial position according to the distribution of $1 - c(\vec{\mathbf{x}})$. We then sample from a Gaussian distribution over image velocities that is centered at the mean velocity estimate of the detector.

6.3 Algorithm Summary

Initially, at time 0, a set of S samples is drawn from the initialization prior, their likelihoods are computed, and normalized to give the weights $\pi_0^{(n)}$. At each subsequent time, the algorithm then repeats the process of sampling from the combined prior, computing the likelihoods, and normalizing.

Note that given the sampled approximation to the distribution $p(\mathbf{s}_t | \vec{\mathbf{Z}}_t)$, we can compute the expected value for some state parameter, $f(\mathbf{s}_t)$, as

$$E[f(\mathbf{s}_t) | \vec{\mathbf{Z}}_t] = \sum_{n=1}^{S} f(\mathbf{s}_t^{(n)}) \pi_t^{(n)}.$$

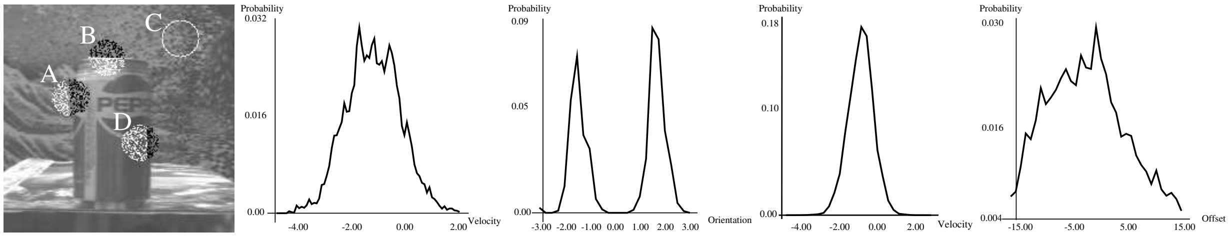
Care needs to be taken in computing this for the orientation as there are often 2 modes 180 degrees apart.

For displaying results, we compute the mean state for each type of model (translation or discontinuity) by computing the expected value of the parameters of the state divided by the sum of all normalized likelihoods for that state. These mean states can be overlaid on the image.

Detection can be performed by comparing the sum of the likelihoods for each model type. Given the way the Condensation algorithm allocates samples, this is not the most reliable measure of how well each model fits the data. If the likelihood of a model drops rapidly, the distribution may temporarily have many low likelihood states allocated to that portion of the state space. The combined likelihood of these states may easily be greater than the likelihood of a new model that does a much better job of fitting the data. Instead, we therefore compute and compare the likelihoods of the mean models to determine which model is more likely.

7 Experimental Results

We illustrate the method with experiments on natural images. For these experiments, the standard deviation, σ_n, of the image noise was 7.0, we use circular image regions with a 16 pixel radius, and we use 3500 state samples to represent the distribution in each image region.



A: u_f **B**: θ **C**: u_0 **D**: d *(offset)*

Figure 4: Pepsi Sequence. Discontinuity (filled) and translational (empty) models shown superimposed on image. Various marginal probability distributions for each region.

7.1 Pepsi Sequence

Since we represent distributions over models and model parameters, it is often difficult to visualize the results. Figure 4 shows marginal probability distributions for different parameters at various image locations. For each region we indicate which model was most likely. Translational models are shown as empty circles (Figure 4 **C**). For discontinuities, we compute the mean state then sample pixel locations from the generative model; pixels that lie on the foreground are shown as white and background pixels as black. Figure 4 **A**, **B**, **D**, shows three such regions.

To the right of the image are views of the marginal probability distributions for various parameters which illustrate the complexity of the posterior. Shown for region **A** is the probability of the horizontal velocity of the foreground; the ambiguity regarding foreground/background means that there are two possible interpretations at -1.7 and -0.8 pixels per frame. The distribution has multiple peaks, two of which correspond to these interpretations and both of which are significant. The foreground/background ambiguity is pronounced in region **B** where the motion is parallel to the edge; this results in strong bi-modality of the distribution with respect to the edge normal, θ, since the direction of the normal points towards the foreground. In region **C** there is no ambiguity with respect to the horizontal velocity of the translation model as shown by the tight peak in the distribution. Finally we plot the offset of the edge, d, for region **D**. In this case, the distribution also non-Gaussian and skewed to one side of the boundary.

Figure 5 illustrates the temporal behavior of the method. Note that the correct assignment of model type is made at each frame, the mean orientation appears accurate, and the boundaries of the Pepsi can are tracked. In the first frame, the assignment of foreground for region **A** is incorrect which is not surprising given that it is ambiguous from two frames. By propagating information over time, however, the initial ambiguities are resolved since the motion of the edge over time is consistent with the foreground being on the right. Note that the ambiguity remains for region **B**, as can be seen by inspecting the distribution in Figure 4, despite the fact that the displayed mean value matches the correct interpretation. In general, propagation of neighboring distributions would be needed to resolve such ambiguities.

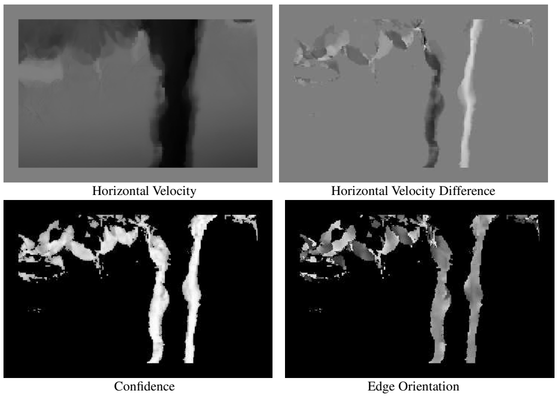


Figure 6: Low level detector responses for one pair of frames in the Flower Garden sequence.

7.2 Flower Garden Sequence

Results on the Flower Garden sequence are shown in Figure 7. The low-level detector responses for the initialization prior are shown in Figure 6; they provide reasonable initial hypotheses but do not precisely localize the edge.

Figure 7 shows the results of our method within several image regions. Regions **C**, **D**, **E**, and **F** correctly model the tree boundary (both occlusion and disocclusion) and, after multiple frames, correctly assign the tree region to the foreground. Note that the orientation of the edge correctly matches that of the tree and that, after the edge passes through the region, the best model switches from discontinuity to translation.

The bottom of Figure 7 shows the probability distribution corresponding to the horizontal velocity of the foreground in region **C**. At frames 2 and 3, there are two clear peaks corresponding the motions of the foreground and background indicating that this relationship is ambiguous. Frames 4 – 6 illustrate how the distribution has formed around the correct motion corresponding to the foreground. By frame 7, the peak has diminished as the translation model becomes more likely.

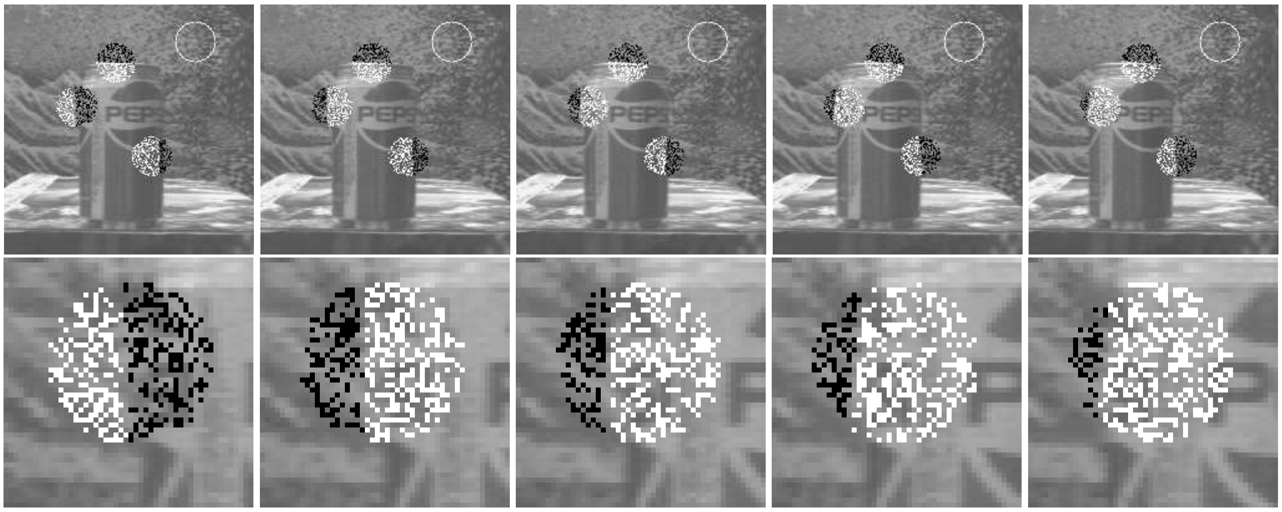


Figure 5: Pepsi Sequence. Top: mean states at frames 1, 3, 4, 7, and 9. Bottom: detail showing region **A**

Region **B** corresponds to translation and is correctly modeled as such. While translation can be equally well accounted for by the discontinuity model, the low-level detectors do not respond in this region and hence the distribution is initialized with more samples corresponding to the translational model. Region **A** is more interesting; if the sky were completely uniform, this region would also be modeled as translation. Note, however, that there are significant low-level detector responses in this area (Figure 6) due to the fact that the sky is not uniform. The probability of translation and discontinuity are roughly equal here and the displayed model flips back and forth between them. For the discontinuity model, the orientation corresponds to the orientation of the tree branches in the region.

8 Conclusions

Work on image motion estimation has typically exploited limited models of spatial smoothness. Our goal is to move towards a richer description of image motion using a vocabulary of motion primitives. Here we describe a step in that direction with the introduction of an explicit non-linear model of motion discontinuities and a Bayesian framework for representing a posterior probability distribution over models and model parameters. Unlike previous work that attempts to find a maximum-likelihood estimate of image motion, we represent the probability distribution over the parameter space using discrete samples. This facilitates the correct Bayesian propagation of information over time when ambiguities make the distribution non-Gaussian.

The applicability of discrete sampling methods to high dimensional spaces, as explored here, remains an open issue. We find that an appropriate initialization prior is needed to direct samples to the portions of the state space where the solution is likely. We have proposed and demonstrated such a prior here but the more general problem of formulating such priors and incorporating them into a Bayesian framework remains open.

This work represents what we expect to be a rich area of inquiry. For example, we can now begin to think about the spatial interaction of these local models. For this we might formulate a probabilistic spatial "grammar" of motion features and how they relate to their neighbors in space and time. This requires incorporating the spatial propagation of probabilities in our Bayesian framework. This also raises the question of what is the right vocabulary for describing image motion and what role learning may play in formulating local models and in determining spatial interactions between them (see [10]). In summary, the techniques described here (generative models, Bayesian propagation, and sampling methods) will permit us to explore problems within motion estimation that were previously inaccessible.

Acknowledgements We thank Allan Jepson for many discussions about motion discontinuities, generative models, sampling methods, and probability theory.

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] A. Bab-Hadiashar and D. Suter. Optic flow calculation using robust statistics. *CVPR*, pp. 988–993, 1997.

[3] M. Black and P. Anandan. Constraints for the early detection of discontinuity from motion. *AAAI*, pp. 1060–1066, 1990.

[4] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, Jan. 1996.

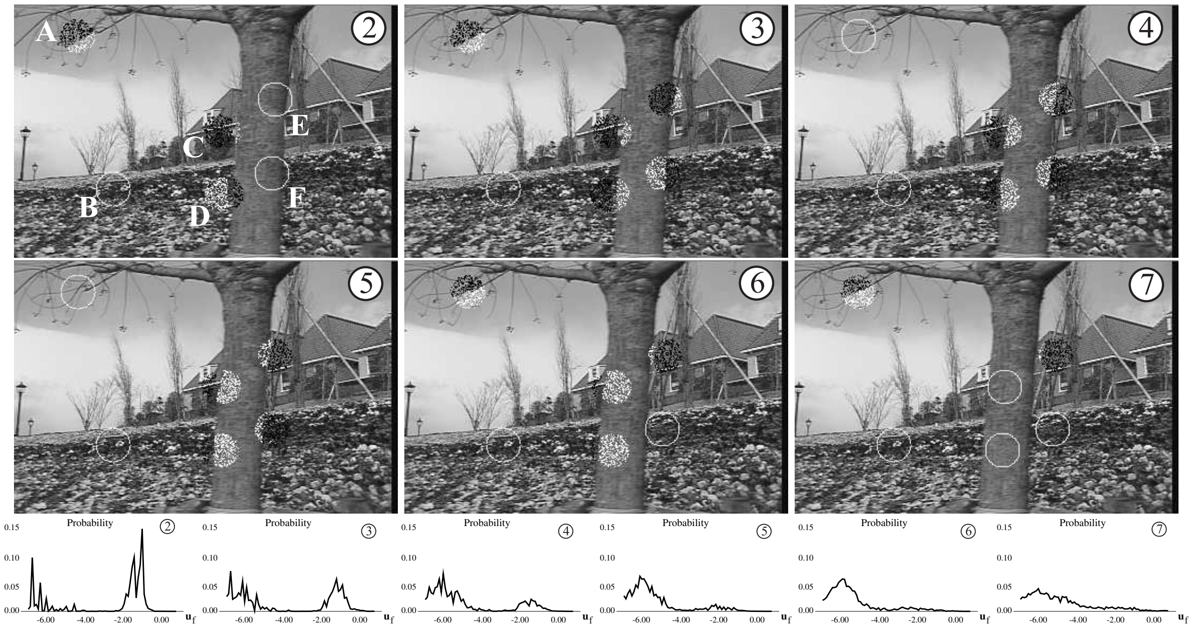


Figure 7: Flower Garden sequence (frames 2–7). Most likely mean models overlaid on images. Bottom: evolution of the marginal probability of the foreground velocity in region **C**.

[5] M. Black and A. Jepson, A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. *ECCV*, vol. 1406, *LNCS*, pp. 909–924, 1998.

[6] M. Black. Explaining optical flow events with parameterized spatio-temporal models. *CVPR*, pp. 326–332, 1999.

[7] G. Chou. A model of figure-ground segregation from kinetic occlusion. *ICCV*, pp. 1050–1057, 1995.

[8] D. Fleet, M. Black, and A. Jepson. Motion feature detection using steerable flow fields. *CVPR* pp. 274–281, 1998.

[9] D. Fleet and K. Langley. Computational analysis of non-Fourier motion. *Vision Res.*, 22:3057–3079, 1994.

[10] W. Freeman and E. Pasztor, Learning to estimate scenes from images. *NIPS*, 1999.

[11] J. Harris, C. Koch, E. Staats, and J. Luo. Analog hardware for detecting discontinuities in early vision. *IJCV*, 4(3):211–223, June 1990.

[12] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *ECCV*, vol. 1064, *LNCS*, pp. 343–356, 1996.

[13] M. Isard and A. Blake. A mixed-state Condensation tracker with automatic model-switching. *ICCV*, pp. 107–112, 1998.

[14] M. Isard and A. Blake. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. *ECCV*, vol. 1406, *LNCS*, pp. 893–908, 1998.

[15] A. Jepson and M. Black. Mixture models for optical flow computation. In *Partitioning Data Sets: With Applications to Psychology, Vision and Target Tracking*, pp. 271–286, DIMACS Workshop, April 1993.

[16] K. Mutch and W. Thompson. Analysis of accretion and deletion at boundaries in dynamic scenes. *PAMI*, 7(2), pp. 133–138, 1985.

[17] S. Niyogi. Detecting kinetic occlusion. *ICCV*, pp. 1044–1049, 1995.

[18] J. Potter. Scene segmentation using motion information. *IEEE Trans. SMC*, 5:390–394, 1980.

[19] H. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE PAMI*, 18(8):814–831, 1996.

[20] B. Schunck. Image flow segmentation and estimation by constraint line clustering. *IEEE PAMI*, 11(10):1010–1027, Oct. 1989.

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

[22] W. Thompson, K. Mutch, and V. Berzins. Dynamic occlusion analysis in optical flow fields. *IEEE PAMI*, 7(4):374–383, July 1985.

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

[24] A. Yuille, P-Y. Burgi, and N. Grzywacz. Visual motion estimation and prediction: A probabilistic network model for temporal coherence. *ICCV*, pp. 973–978, 1998.