Bayesian Inference of Visual Motion Boundaries

David J. Fleet
Palo Alto Research Center
3333 Coyote Hill Rd., CA 94304

Abstract

We present a Bayesian formulation for image motion estimation in which local motion is explained in terms of multiple, competing, nonlinear models, including one for motion boundaries. We formulate the posterior probability distribution over the models and model parameters, conditioned on the image sequence. Approximate inference is achieved with a combination of tools: Bayesian filtering for online computation; factored sampling to represent multimodal non-Gaussian distributions and to propagate beliefs with nonlinear dynamics; and mixture models to simplify the computation of joint prediction distributions. To efficiently represent such a high-dimensional space we also initialize samples using the responses of a low-level motion detector. The basic formulation and computational model provide a general probabilistic framework for motion estimation with multiple, non-linear, models, the details of which can be found in [4, 18].

Keywords: Optical Flow, Occlusion Boundaries, Hybrid State Estimation, Particle Filters, Hybrid Random Fields, Nonparametric Belief Propagation.

Visual Motion Analysis

Motion is an intrinsic property of the world and an integral part of our visual experience. It provides a remarkably rich source of information that supports a wide variety of visual tasks. Examples include 3D model acquisition, event detection, object recognition, temporal prediction, and oculomotor control.

A key computational problem for visual motion analysis is the estimation of optical flow, the 2D image velocity of projected points in the scene. Since first studied over 20 years ago [7, 11], techniques for estimating optical flow have improved significantly. The use of benchmark data sets and publically available code have helped to estabalish the quantitative accuracy of recent methods [1]. Accordingly, it is now relatively well accepted that, for smooth textured surfaces, current methods provide accurate and relatively fast estimators for 2D image velocity. Although many interesting variations exist, perhaps the simplest, most commonly used techniques are known as area-based regression methods. Broadly speaking, these techniques are derived from two main assumptions, widely known as *brightness constancy* and *smoothness*.

While regression techniques produce reliable optical flow estimates for smooth textured surfaces, there are motions for which they are not effective. There are many situations where brightness constancy does not hold and the motion is not smooth. Examples include motion discontinuities, the motion of bushes or trees in the wind, and the deformations and self-occlusions of clothing as people walk. As we aim to cope with more and more complex dynamic phenomena, it is becoming clear that we need more explicit models for such dynamic processes. Moreover, we require new computational methods for coping with such complex models.

Motion and Surface Boundaries

One outstanding problem is the estimation of motion at surface boundaries. Because of depth discontinuities at surface boundaries, the image motion is often discontinuous as well, thereby violating the motion smoothness assumption. In addition, pixels near the boundary that are visible at one time may not be visible at the next time as the foreground moves and thereby occludes a different portion of the background. This violates the brightness constancy assumption. As a result most optical flow techniques produce poor estimates at occlusion boundaries.

Nevertheless, motion boundaries are a rich source of scene information. First, they provide information about the position and orientation of surface boundaries. Second, analysis of the occlusion or dis-occlusion of pixels at motion boundaries can provide information about the relative depth ordering of the adjacent surfaces. In turn, information about surface boundaries and depth ordering may be useful for tasks as diverse as navigation, structure from motion, video compression, perceptual organization, and object recognition.

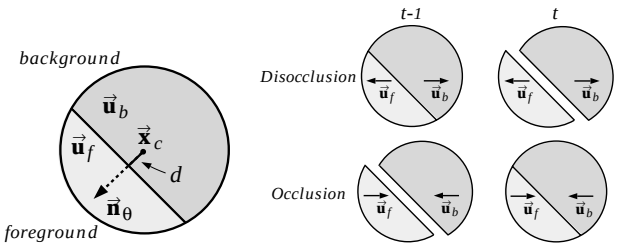

Figure 1: A motion boundary is parameterized by foreground and background velocities, $\mathbf{u}_f$ and $\mathbf{u}_b$, an orientation θ with normal $\mathbf{n}_\theta$, and a signed distance d from the neighborhood center $\mathbf{x}_c$. From this model we can predict which pixels are visible in frames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

The detection of motion boundaries has been a long-standing problem in optical flow estimation [1, 19]. Most previous methods cope with motion boundaries by treating them as a form of *noise*; that is, as the violation of a smoothness assumption. This occurs with regularization schemes where robust statistics, weak continuity, or line processes are used to locally disable smoothing across motion discontinuities [10, 14, 17]. Robust regression [3] and mixture models [20, 13] have been used to account for the multiple motions. But they often break down at boundaries where outliers can comprise up to half of the motion constraints.

Such methods fail to explicitly model the image structure in the spatiotemporal neighborhood of the boundary; they do not model the boundary orientation, which pixels are occluded/disoccluded, or the depth ordering of the surfaces. Additionally, most of the above methods have no explicit temporal model so that information can be integrated over time to improve inference of surface depth ordering.

In this presentation we describe a probabilistic, model-based, approach to image motion analysis in which the 2D motion in each local image neighborhood is represented using one of several possible models. This allows us to use different motion models that are suited to the diverse types of optical flow that occur with natural scenes. Here, we consider two models, namely, for smooth motion and motion boundaries. Regions of smooth motion are modeled using conventional translational models, while the complex phenomena that occur at motion boundaries are accounted for by an explicit, non-linear, boundary model.

The motion boundary model depicted in Fig. 1 encodes the boundary orientation, the image velocities on each side of the boundary, the depth ordering of the two sides, and the distance from the boundary to the region center. With this model we can predict the visibility of occluded and disoccluded pixels so that these pixels may be excluded when estimating the probability of a particular motion. Moreover, the explicit offset parameter allows us track the movement of the boundary through the region. This foreground/background ambiguities to be resolved.

Nonparametric Bayesian Inference

To cope with image noise, matching ambiguities, and model uncertainty, we adopt a Bayesian probabilistic framework that integrates information over space and time and represents multiple, competing, model hypotheses. The goal is to compute the posterior probability distribution over motion models and their parameters, conditioned on image measurements. The posterior is expressed in terms of a likelihood function and a prediction distribution. The likelihood is the probability of observing the current image given the correct model. The prediction represents our prior belief about the motion at the current time and location based on previous observations; it embodies the temporal dynamics of how models and model parameters evolve over time, and our belief about the spatial coherence of boundaries.

Computational problems with this formulation arise in several ways, the first of which is the use of hybrid states, involving discrete and continuous variables. The discrete variable encodes the type of motion, and the continuous variables encode the corresponding motion parameters (2 for smooth motion, and 6 for the motion boundary model). Distributions over hybrid state spaces where continuous variables depend on discrete variables are usually multi-modal. Moreover, it is similarly significant that our likelihood functions and temporal dynamics are nonlinear, which further complicates the expected structure of the posterior distribution.

The final source of computational difficulty arises because we want to estimate motion at locations throughout the image that are statistically dependent on one another. These dependencies arise because motion boundaries will tend to move from one region to another, and because motion boundaries tend to be smooth and coherent through space. As a result, we are actually interested in multi-modal hybrid state inference on a time-varying random field.

For these reasons we turn to methods for approximate probabilistic inference. One form of approximate inference that has become popular for dynamical vision problems (e.g., motion and tracking), is the particle filter [6, 9, 12]. Particle filters approximate the posterior with a weighted set of samples; samples

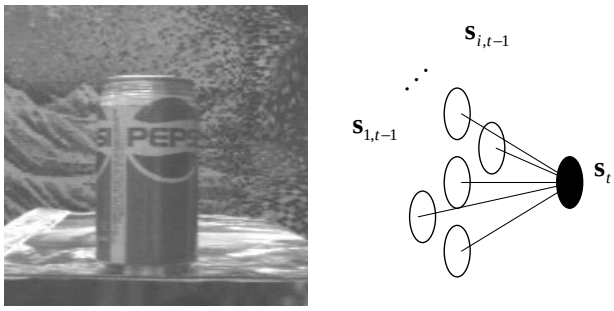

Figure 2: (left) Frame 1 of the Pepsi can image sequence. (right) stuff

are drawn randomly from a proposal distribution, often the prediction distribution, and then weighted by a function of the likelihood. With such *point-mass approximations* to probability distributions, particle filters are useful in coping with non-Gaussian, multimodal distributions.

If we were to treat each image neighbourhoods independently of other neighbourhoods, then we could use a particle filter to approximate a to a low-dimensional posterior for the motion in each image region [4]. However, because the motions in adjoining regions are not independent, we have a random field of interdependent states. While particle filters are suitable for low-dimensional hybrid states, they will not cope with random fields. As is well known, a primary problem with particle filters is the exponential increase in the required number of particles (i.e., computational cost) as a function of the dimensionality of the state space (e.g., see [5, 15]). This becomes prohibitive with a random field.

In this presentation we explore different forms of approximate inference, drawing on research described in [4, 18]. In the first case we approximate the posterior and the temporal dynamics by factoring each so that each local region of the image can be treated separately. This has problems since it prohibits us from encouraging boundary continuity and from allowing one region to predict when edges are going to move from one region to another.

We assume a very simple form of graphical model depicted in Fig. 2(right), in which neighbourhoods at time t have direct statistically dependence on nearby neighbours at the previous time $t-1$. For inference we use a form of nonparametric Bayesian belief propagation which implicitly assumes that the posterior at a given location can be approximated as the product of its marginal distributions for that region (cf. [16]). Towards this end, we use Monte Carlo (sampled) approximations to these distributions to deal with non-linear dynamics and non-Gaussian likelihoods, and we use mixture models to efficiently approximate the prediction distributions that arise from multiple neighborhoods. More detail can be found in [18].

While the method described here can be thought of simply as a motion boundary detector, the framework has wider application. The Bayesian formulation and computational model provide a general probabilistic framework for motion estimation with multiple, non-linear, models. This generalizes previous work on recovering optical flow using linear models [2, 8]. Moreover, the Bayesian formulation provides a principled way of choosing between multiple hypothesized models for explaining the image variation within a region. This work can also be viewed as an exploration of the suitability of different forms of approximate inference in the context of otherwise intractable inference problems in vision.

Experimental Results

Figure 3 shows some experimental results of an implementation of the approach applied to the well-known Pepsi can image sequence (Fig. 2). At frame 1, before belief propagation has begun to spread information through time and to neighbouring locations, the individual regions show relatively little coherence. In frames 2 and 3 the neighborhood interactions introduce some coherence. Noteworthy in Fig. 3 are the correct assignment of the foreground and the accurate localization of the motion boundaries. Also evident in Fig. 3 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 then in frames 9–10 on the right side. This propagation allows the correct depth ordering to be infered quickly.

Conclusions

Research on image motion estimation usually relies on relatively weak models of the spatiotemporal image structure. Our goal is to move towards a richer description of image motion using a vocabulary of motion primitives. This work represents one step in that direction with the introduction of an explicit non-linear model of motion boundaries and a Bayesian framework for representing a posterior 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.

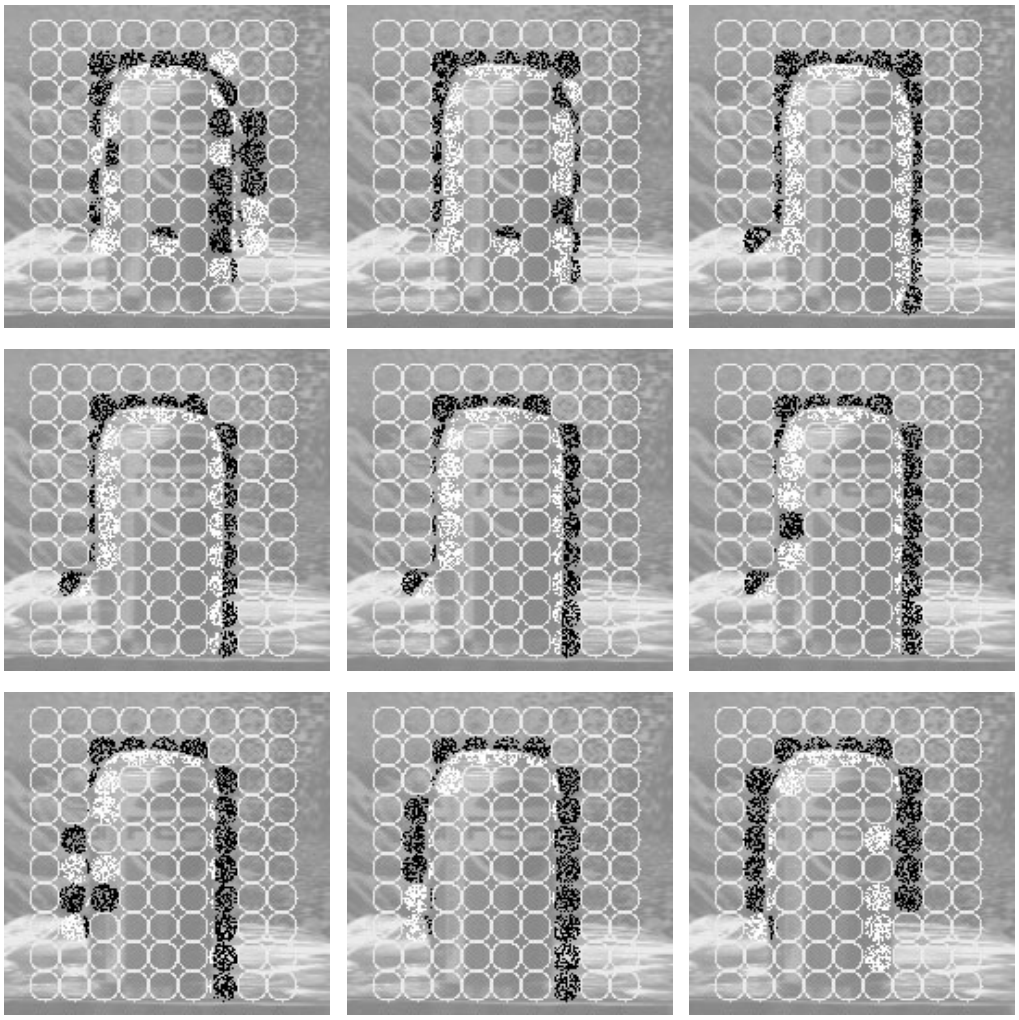

Figure 3: Pepsi can results for frames 2–10 (in lexicographic order and cropped slightly to improve the resolution of the display).

However, due to the complex multimodal nature of the distributions, one cannot perform exact Bayesian inference for such problems. Instead we explore different forms of approximate inference. Here we extend the use of particle filters with other methods for approximate inference to the detection and estimation of multiple motions defined on a dynamic random field. With such techniques we can begin to explore problems of motion estimation that were previously inaccessible.

Acknowledgements: Thanks to Michael Black, Allan Jepson, Ray Luo, and Oscar Nestares for their participation at various stages of this research, and for discussions about motion discontinuities, generative models and sampling methods.

References

[1] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *IJCV*, 12:43–77, 1994.

[2] J. R. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. *Proc. ECCV*, pp. 237–252. Springer-Verlag, 1992.

[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] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, Berlin, 2001.

[7] C. L. Fennema and W. B. Thompson. Velocity determination in scenes containing several moving objects. *CVGIP*, 9:301–315, 1979.

[8] D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson. Design and use of linear models for image motion analysis. *IJCV*, 36:169–191, 2000.

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

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

[11] B. K. P. Horn and B. G. Schunk. Determining optical flow. *AI*, 17:185–203, 1981.

[12] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *IJCV*, 29:2–28, 1998.

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

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

[15] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. *Proc. ECCV*, vol. II, pp. 134–149, 2000.

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

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

[18] O. Nestares and D. J. Fleet. Probabilistic tracking of motion boundaries with spatiotemporal predict ions. *Proc. IEEE CVPR*, vol. II, pp. 358–365, 2001.

[19] M. Otte and H. H. Nagel. Optical flow estimation: Advances and comparisons. *Proc. ECCV*, pp. 51–60, 1994.

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