To appear: Proceedings of the International Conference on Computer Vision, Osaka, Japan, Dec. 1990.

A Model for the Detection of Motion over Time

Michael J. Black and P. Anandan

Department of Computer Science
Yale University
New Haven, CT 06520–2158

Abstract

We propose a model for the incremental estimation of visual motion fields from image sequences. Our model exploits three standard constraints on image motion within an optimization framework: i) Data Conservation: *the intensity structure of a surface patch changes gradually over time;* ii) Spatial Coherence: *neighboring points have similar motions;* iii) Temporal Coherence: *the image velocity of a surface patch changes gradually. Our formulation takes into account the possibility of multiple motions at a particular location. We present an incremental scheme for the minimization of our objective function, based on simulated annealing. All computations are parallel, local, and incremental, and occlusion and disocclusion boundaries are estimated.*

Introduction

The estimation of visual motion fields from image sequences is generally treated as a problem of combining multiple constraints, each of which is insufficient to determine the field uniquely, but can together determine a unique motion field. One of these constraints involves measurements computed from the image sequence, while the others reflect various assumptions about the spatial and temporal coherence of surfaces and their motion.

Specifically, we can identify three constraints: The *data conservation* constraint states that the image measurements (e.g., the intensity structure) corresponding to an environmental surface patch change slowly over time. The *spatial coherence* constraint is derived from the observation that surfaces have spatial extent and hence neighboring points on a surface will have similar motion. The *temporal coherence* constraint is based on the observation that the velocity of motion tends to change gradually.

The accuracy of flow-field computation depends critically on the precise form of the constraints and on their proper integration. The traditional formulations of the constraints embody a global smoothness assumption. As is well known, this assumption is violated when multiple motions are present.

In this paper, we suggest ways of reformulating the data constraint and the spatial smoothness constraint to allow for the possibility of multiple motion fields in a single image sequence. Our formulation is based on *robust statistics* [3] and the *weak continuity constraints*[4] used in Markov Random Field (MRF) models. We also suggest a way to formulate the temporal coherence constraint.

This paper also presents a highly parallel and *incremental* model to exploit these constraints in which computation occurs locally, knowledge about the motion increases over time, and occlusion and disocclusion boundaries are estimated. The reformulation of the constraints results in an objective function which is non-convex, and whose minimization is achieved through the use of a stochastic temporal updating scheme. Traditional stochastic relaxation techniques that are based on simulated annealing typically require many iterations over the same image data. We employ a new technique which is more appropriate for sequences of slowly varying images and in which we replace the notion of iteration over the same data with iteration over time.

The following section describes our new formulation of the constraints as energy terms defined over a small neighborhood. This section also describes techniques for coping with discontinuities in the motion field. We then present a computational framework for exploiting the constraints and our temporal updating scheme. Experimental results with synthetic and real images sequences are presented.

Formulating the Constraints

The paradigm within which we operate is the standard one of converting each constraint into a term in an overall objective function that is minimized to obtain the motion field. We reformulate the traditional constraints to account for multiple motions by using techniques for *outlier rejection* [3].

The idea is that, within a window centered on a motion discontinuity, any measurements pertaining to the motion of the surfaces will fall into two distributions. The types of measurements can be either intensity differences between a patch in one image and a translated patch from a previous image, or differences in neighboring flow vectors. In either case, these measurements can be classified as belonging to one of two distributions: measurements consistent with the motion we are examining, and measurements which are inconsistent. While attempting to interpret one motion, measurements corresponding to the other motion can be viewed as statistical outliers which, if not rejected, will contaminate the final interpretation of the motion.

The standard assumption, in the absence of motion discontinuities, is that errors can be described by Gaussian noise. By "errors" we refer to the deviations of the measured data from the model predictions. This Gaussian assumption leads to a quadratic objective function $D(x) = x^2$, which disproportionately weights outliers. Robustness in the presence of multiple distributions can be achieved by adopting a

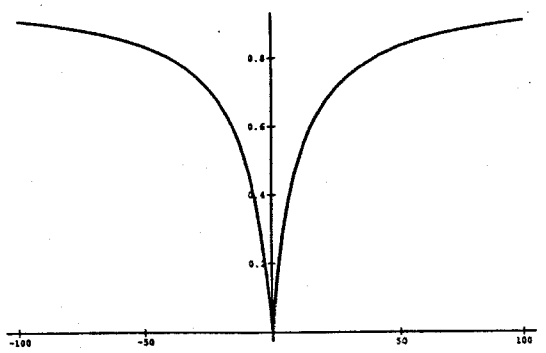

Figure 1: Shape of the ϕ' function

more realistic model of the error. One technique for relaxing the assumption of simple Gaussian error is the *weak continuity constraint* [4, 5, 6], wherein the quadratic objective function is replaced by:

$$H^*(x,l) = (1-l)D(x) + \beta(l),$$

where l is a binary valued "line process" which has value $l = 0$ if x is consistent with our model and a value of $l = 1$ if not. So the new function H^* is just the original objective function D if our model is valid but becomes some fixed value β when the model is violated. The term β amounts to a penalty for violating the model.

Blake and Zisserman [4] show that the line processes can be eliminated from the objective function by first minimizing over them, resulting in an objective function which is solely a function ϕ of the actual variables under consideration.

Geman and Reynolds [6] remove the binary restriction from the line processes and in doing so derive another ϕ function (see figure 1) with similar properties:

$$\phi(x) = \frac{-1}{1+|x|/\Delta}. \qquad (1)$$

This function has been used for the spatial smoothness constraint in [6] for image restoration preserving discontinuities. Regardless of the ϕ function chosen, the objective function simply becomes $H(x) = \phi(x)$.

Qualitatively, the effect of the type of ϕ function given above is to weight highly measurements which correspond to our model and remain uncommitted about areas which do not conform to expectations. This is a simple, yet important, heuristic for increasing robustness in the presence of outliers resulting from multiple motions.

The Data Conservation Constraint

The data error term embodies the assumption that the intensity of a surface element remains constant over time, although its image location may change.

Let $S = \{s_1, s_2, \ldots, s_{n^2}\}$ be a fixed set of *sites* corresponding the $n \times n$ pixels in image I. Let $(i(s), j(s))$ denote the pixel coordinates of site s. Given image intensity functions I_n and I_{n+1} between two successive frames (n and $n+1$), the local contribution (at site s) to the data conservation constraint is defined as an energy term $E_D^s(u,v)$ over the space of possible displacements (u,v) at pixel "site" s:

$$\sum_{t \in G_s} \phi_D(I_n(i(t),j(t)) - I_{n+1}(i(t)+u, j(t)+v)), \quad (2)$$

where G_s denotes a neighborhood of s (usually a square window of a pre-specified size).

Note that if we define $\phi_D(x) = (x/\Delta)^2$ (where Δ is simply a scaling factor), then ϕ_D defines the standard *sum of squared differences* (SSD) measure [1]. This measure assumes that all the points in the neighborhood G_s are translated by the uniform velocity (u,v) and the resulting image is corrupted by additive Gaussian noise. (In practice, this measure serves as a good approximation even if the velocities vary gradually around (u,v).)

The Gaussian noise model on which the SSD measure is based is inappropriate for dealing with multiple motions. To allow the possibility of multiple motions within the neighborhood we relax this assumption and adopt the Geman and Reynolds ϕ function described above.

In comparison with the SSD measure, the proposed ϕ function has the following property: if two motions are present, then a pixel that is consistent with one of the two motions tends to contribute less (than the SSD measure) to the data error of the other motion, resulting in less "cross-talk."

The Spatial Smoothness Constraint

We formulate the spatial smoothness constraint as consisting of a sum of error terms $E_S^s(u,v)$ defined locally at site s as:

$$E_S^s(u,v) = \sum_{t \in G_s} \phi_S(u(s)-u(t)) + \phi_S(v(s)-v(t)), \quad (3)$$

where $(u(s), v(s))$ is the motion vector at site s. Once again, there are a variety of choices for ϕ_S. Taking $\phi_S(x) = (x/\Delta)^2$ results in the standard quadratic smoothness term [2].

The Bayesian interpretation of the standard smoothness error term corresponds to a particular prior model of the motion field. In this case it is a Gaussian random variable model (white noise, or as Szeliski notes, an MRF with a fractal prior [8]) with a particular covariance structure reflected in the smoothness error term.

Once again, the smoothness assumption and its standard (quadratic) formulation are invalid in areas containing multiple motions. To deal with this we use the weak continuity constraint. The property of becoming non-committal as the differences increase amounts to a weakening of the smoothness assumption. For our current implementation, we have chosen the same weak continuity function ϕ for both the data-conservation and the smoothness terms with a possibly different Δ.

The Temporal Coherence Constraint

The temporal coherence constraint is intended to capture the idea that the motion of a particular surface element changes gradually over time. Implementing the constraint requires maintaining a correspondence

between sites and moving patches of the environment. The obvious solution is to use the estimated motion field itself to determine the correspondence of points over time. This amounts to tracking points over time. The details of a tracking scheme are described later in this document.

Assuming that tracking is achieved, even if only imperfectly, the temporal coherence constraint can be formulated as an error function defined over the temporal change in the motion field. In order to achieve a degree of robustness, we introduce the notion of the *history* of a point. At a minimum the history of a point should include its past motion information. (Additionally, it can also include the length of time the point has been visible, or even more complex information such as to what surface the point belongs.)

In our current formulation, we maintain a moving average of the past motion information associated with a point. Let s denote the site of that point in the current image frame, and $(u_{avg}(s), v_{avg}(s))$ be the average motion. Then the temporal coherence constraint is formulated as a error term:

$$E_T^s(u,v) = \phi_T(u(s) - u_{avg}(s)) + \phi_T(v(s) - v_{avg}(s)) \quad (4)$$

where ϕ_T is the same function used in the data conservation and the smoothness error terms, with a possibly different Δ.

The Computational Model

Survival imposes strong requirements on the visual system of a mobile agent. Computation must be fast; this leads to a computational model which is highly parallel and in which computations are simple and local. Information about motion must always be available even if is only a rough estimate. In a dynamic world, off-line processing of motion is unacceptable. While rough motion estimates may be useful, they are not enough. A mobile agent should be able to improve its knowledge about the environment over time.

The model of motion processing we propose has the flavor of MRF approaches in that the probability of a surface patch having a particular displacement is determined by its relationship to its neighbors in space and time. A separate processor is allocated to each site $s \in S$ described in the previous section and can communicate with some set of its neighbors G_s.

Each processor represents a small environmental surface patch and maintains information about the motion of that patch over time; where it originated and where it is going. Associated with processor s is a random vector (u_s, v_s) which represents the current image displacement of the corresponding surface patch. In our current implementation, the vector (u_s, v_s) can take on any value from within a discrete set Λ, which is defined as,

$$\Lambda = \{(u,v) \mid -m \leq u, v \leq m\}$$

where m is the maximum expected displacement.

The constraints described in the previous section are defined in terms of nearest neighbor relations. Our goal is to find the values of (u_s, v_s) which minimize the function:

$$H(u,v,t) = \beta_D E_D(u,v) + \beta_S E_S(u,v) + \beta_T E_T(u,v),$$

where each of the error terms is evaluated at the current time instance t, and $\beta_\star$ are constant weights which control the relative importance of the constraints.

Each of the individual terms constituting the the objective function H is a non-convex function with multiple local minima. Our method for the minization of H is derived from simulated annealing. For each site, a probability density function Π is defined over Λ as follows:

$$\Pi(u,v,t) = Z^{-1} e^{-H(u,v,t)/T(t)},$$

where:

$$Z = \sum_{u,v} e^{-H(u,v,t)/T(t)}.$$

The quantity $T(t)$ can be viewed as a *temperature* and serves to sharpen (or flatten) the distribution.

At each time instance, each processor randomly selects a displacement vector according to its probability density function Π. This sampling method is repeated with decreasing values of the temperatures. At high temperatures our sampling process freely chooses among displacements, but as the temperature is lowered, the minimum is chosen with increasing probability as the probability distribution Π becomes concentrated about the minimum.

The similarities between our formulation and simulated annealing techniques for MRF models [5] are obvious. There are some important differences, however. For instance, in our formulation, different sites may have distinct temperatures. We also do not try to find a global minimum of the objective function H at a single time instance. The standard stochastic relaxation techniques used for this purpose require a large number of iterations with fixed data, which makes them undesirable for dynamic motion processing.

By exploiting slowly varying imagery and keeping track of the motion of patches, we attempt to replace iteration over a pair of frames with iteration over a sequence. Thus, while our approach raises the potential for true dynamic processing, we do not make any claims to the applicability of the the rigorous convergence results associated with the standard stochastic relaxation methods.

Each patch is tracked in the image by updating the grid of processors to reflect the predicted motion. At any time instance, after all processors have determined the motion of their corresponding patches, the grid is updated to determine what patch projects to a particular processor at the next time instance. This can be viewed as *warping* the previous image according to the current motion field in order to produce an expectation for the following image.

At *occlusion* boundaries, patches on both the occluding and the occluded surfaces may project to the same site. Hence, each site receiving multiple projections marks itself as a possible occlusion boundary and chooses to take on the motion of one of the projecting neighbors.

Analogously, a site may have no other patches projecting to it. This is the case at a *disocclusion* boundary where new surfaces become visible. These sites correspond to, and are assigned, new patches in the world which have yet to be tracked. These sites are

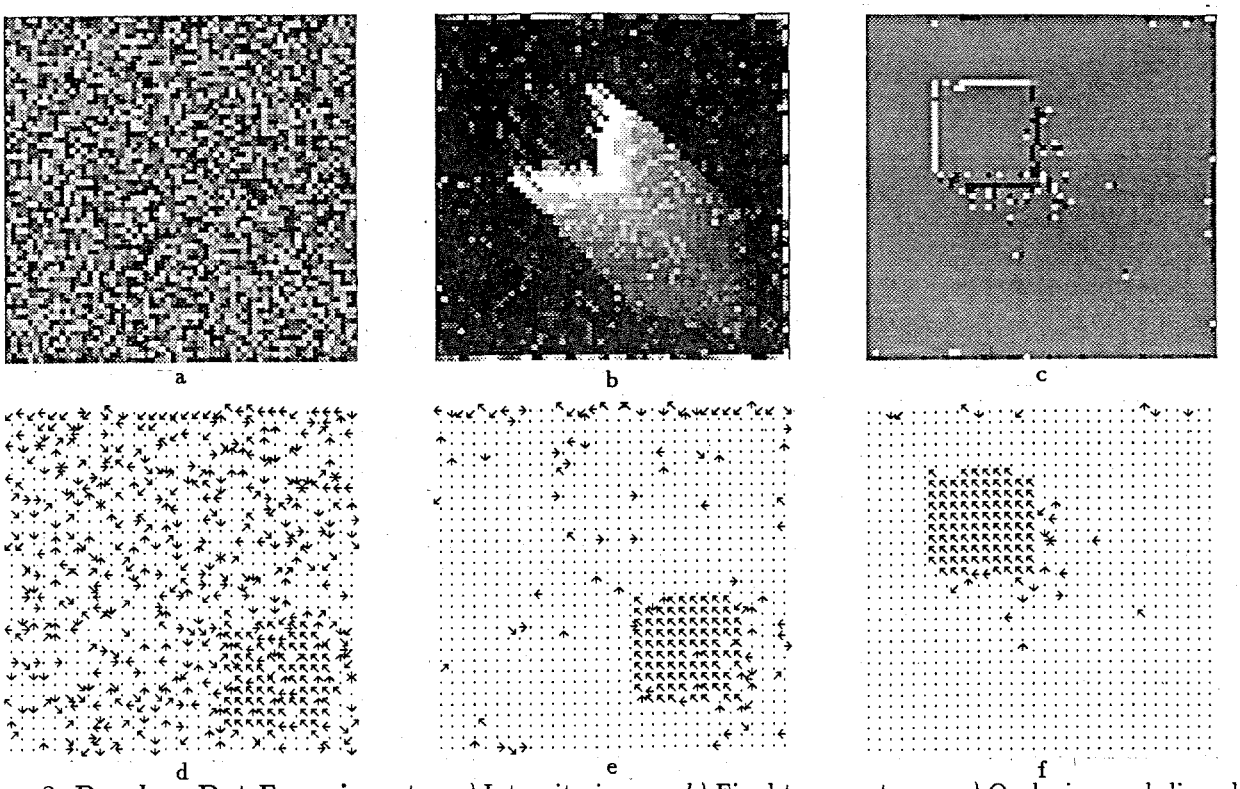

Figure 2: **Random Dot Experiments.** *a)* Intensity image. *b)* Final temperatures. *c)* Occlusion and disocclusion boundaries. *d)* Initial flow field. *e)* Flow field after 6 images. *f)* Final flow field after 20 images.

marked as a possible disocclusion boundaries and assigned an initial high temperature to reflect their lack of history. The temperatures in recently disoccluded areas tend to be high, and gradually decrease as their corresponding surface patches persist in the image.

In addition to the current motion of a patch, each site records information about the history of the patch. When a site is updated at time t, a weighted average of the patches motion (u_{avg}, v_{avg}) is computed, where:

$$u_{avg} = (1 - \delta)u_t + \delta u_{avg},\ 0 < \delta < 1.$$

The value of δ determines the rate at which old information decays and is overridden by the recent motion of a patch.

After the grid is updated, a new image in the sequence is examined and the annealing process is restarted. The number of annealing iterations between any particular pair of frames is a parameter of the system.

Experimental Results

The model described in the previous section has been implemented on the Connection Machine. The architecture of the machine is ideally suited to our computational model.

The first example involves a synthetic image sequence with a textured square moving across a stationary textured background. The random dot texture of the foreground and background patches is uniformly distributed between 0 and 255 (figure 2*a*). The sequence consists of thirty frames; in each frame the foreground patch moves one pixel up and to the left. Uniform random noise over the range $[-\eta, \eta]$ was added to each image in the sequence. For the example in figure 2*a*, η is taken to be five percent of the intensity range; so $\eta = 12.75$.

The results of the motion algorithm applied to the sequence are shown in figure 2. Only a single iteration of the temporal annealing algorithm was performed for each pair of images in the sequence.

Figure 2*b* shows the temperature at each site at the end of the image sequence. Lighter areas correspond to higher temperatures. The stationary background and the patch itself are dark, indicating that by the end of the sequence, the motion of these areas is known accurately. The brightest areas correspond to recently disoccluded portions of the background.

Occlusion and disocclusion estimates are shown in figure 2*c*; occlusion boundaries are displayed as white, and disocclusion boundaries appear black.

Figure 2*d* shows the initial flow field. The incremental algorithm continually improves the estimate over the 30 frame sequence, resulting in the flow field in figure 2*f*. The motions of the patch and background are known with fairly high accuracy while the recently disoccluded areas with high temperatures have not yet settled into a stable interpretation. Increasing the number of iterations per frame would permit these ar-

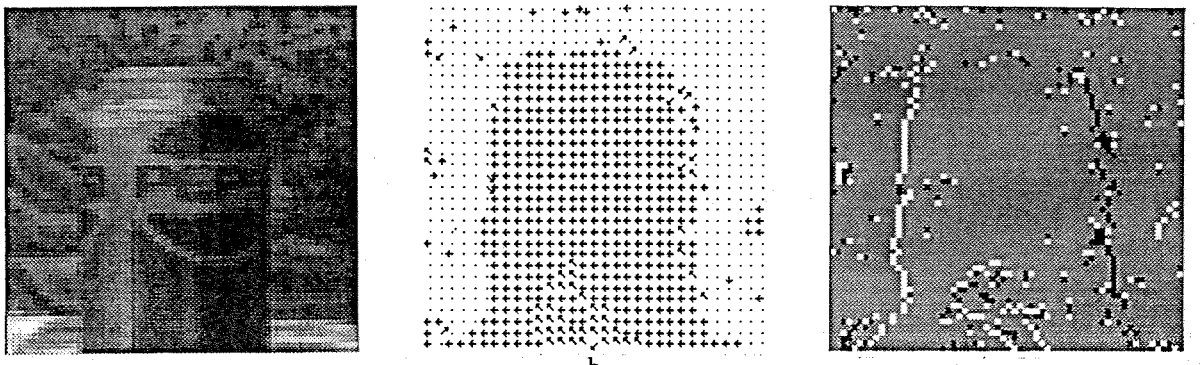

Figure 3: **Pepsi Sequence,** *a)* Intensity image. *b)* Flow field. *c)* Discontinuities.

eas to settle more quickly to the correct interpretation, but would diminish the dynamic nature of the processing.

The final results of the model applied to a real image sequence are shown in figure 3. This sequence[1] consists of ten images. The images contain a can in the foreground, moving approximately one pixel to the left each frame, in front of a textured background which is undergoing a very slight leftward motion.

Since the motion sequence is relatively short, good motion estimates cannot be achieved with only a single iteration of the annealing algorithm for each frame. For this reason, ten iterations were used for this example.

Figure 3*b* shows the final flow field. Figure 3*c* shows the occlusion and disocclusion estimates, Errors in the flow field appear in areas of the image where there is little texture.

Conclusion

We have reformulated data conservation, spatial coherence, and temporal coherence constraints to take into account the possibility of multiple motions. We have also presented a new computational framework for exploiting these constraints. The model has some desirable properties: it is parallel, computation is local, occlusion/disocclusion boundaries are estimated, and its incremental nature means that a motion estimate is always available and improves over time.

Much work remains to be done, however; for example, while our updating scheme is derived from MRF approaches, we have not attempted to extend the of the mathematical foundation of MRFs to this temporal framework. We make no claims about the convergence of our new incremental scheme; a theoretical analysis of the model is required.

Our preliminary implementation assumes that all motion is approximately one pixel per frame. While larger motions can be dealt with by expanding the state space, the result is a loss of efficiency. Instead, our current work is extending the computational model to include a multi-resolution processing scheme [2, 7].

The greatest weakness of this initial implementation is that only discrete motions are detected. To extend the model to general fractional pixel motion requires extending the data constraint to provide sub-pixel accuracy and extending the minimization scheme to deal with a continuous state space. The former problem has been addressed in [2]. Continuous minimization can be realized by exploiting results in continuous annealing [9]. Finally, new warping and discontinuity detection schemes are being developed to cope with arbitrary motions.

Acknowledgements

We wish to thank George Reynolds for his insightful comments on MRFs and weak continuity constraints.

References

[1] Anandan, P., "Measuring visual motion from image sequences," *Ph.D. dissertation*, COINS TR 87-21, U. of Mass., Amherst, MA, 1987.

[2] Anandan, P., "A computational framework and an algorithm for the measurement of visual motion," *Int. J. of Comp. Vision*, 2, 1989, pp. 283–310.

[3] Besl, P.J., Birch, J.B., Watson, L.T., "Robust window operators," *Int. Conf. on Comp. Vision*, ICCV-88, 1988, pp. 591–600.

[4] Blake, A. and Zisserman, A., *Visual Reconstruction*, The MIT Press, Cambridge, MA, 1987.

[5] Geman, S. and Geman, D., "Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images," *IEEE PAMI*, Vol. PAMI-6, No. 6, Nov. 1984.

[6] Geman, D. and Reynolds, G., "Constrained restoration and the recovery of discontinuities," unpublished manuscript.

[7] Konrad, J., and Dubois, E., "Miltigrid Bayesian estimation of image motion fields using stochastic relaxation," *Int. Conf. on Comp. Vision*, ICCV-88, pp. 354–362, 1988.

[8] Szeliski, R. S., "Bayesian modeling of uncertainty in low-level vision," Ph.D. Thesis, CMU, 1988.

[9] Vanderbilt D., and Louie S. G., "A monte carlo simulated annealing approach to optimization over continuous variables," *J. of Comp. Physics*, **56**, pp. 259–271, 1984.

[1] We would like to thank Joachim Heel for providing the motion sequence.