Combining Intensity and Motion for Incremental Segmentation and Tracking Over Long Image Sequences

Michael J. Black

Research Report YALEU/DCS/RR-873
October 1991

Combining Intensity and Motion for Incremental Segmentation and Tracking Over Long Image Sequences*

Michael J. Black†

Department of Computer Science
Yale University
P.O. Box 2158 Yale Station
New Haven, CT 06520–2158
USA

Phone: +1 (203) 432–1223
Fax: +1 (203) 432–0593
Email: black-michael@cs.yale.edu

Abstract

This paper presents a method for incrementally segmenting images over time using both intensity and motion information. This is done by formulating a simple model of surface patches using local constraints on intensity and motion and then finding the optimal segmentation over time using an incremental stochastic minimization technique. The result is a robust and dynamic segmentation of the scene over a sequence of images. The approach has a number of benefits. First, discontinuities are extracted and tracked simultaneously. Second, a segmentation is always available and it improves over time. Finally, by combining motion and intensity, the structural properties of discontinuities can be recovered; that is, discontinuities can be classified as surface markings or actual surface boundaries.

*Submitted: Second European Conference on Computer Vision, May 1992.

†This work was supported by a grant from the National Aeronautics and Space Administration (NGT–50749), by the NASA Ames Research Center, Aerospace Human Factors Research Division, NASA RTOP 506-47, and by a grant from the Whitaker Foundation.

1 Introduction

Our goal is to efficiently and dynamically build useful and perspicuous descriptions of the visible world over a sequence of images. In the case of a moving observer or a dynamic environment this description must be computed from a constantly changing retinal image. Recent work in Markov random fields models [12], recovering discontinuities [5], segmentation [9, 11], motion estimation [3, 4, 18], motion segmentation [6, 10, 14, 21], and incremental algorithms [3, 13, 20, 24] makes it possible to begin building such a structural description of the scene over time by compensating for and exploiting motion information.

As an initial step towards the goal, this paper proposes a method for incrementally segmenting images over time using both intensity and motion information. The result is a robust and dynamic segmentation of the scene over a sequence of images. The approach has a number of benefits. First, discontinuities are extracted and tracked simultaneously. Second, a segmentation is always available and it improves over time. Finally, by combining motion and intensity, the structural properties of discontinuities can be recovered; that is, discontinuities can be classified as surface markings or actual surface boundaries.

By jointly modeling intensity and motion we extract those regions which correspond to perceptually and physically significant properties of a scene. The approach we take is to formulate a simple model of surface patches using local constraints on intensity and motion. The formulation of the constraints accounts for surface patch boundaries as discontinuities in intensity and motion. These constraints are modeled probabilistically using a Gibbs distribution. The segmentation problem is then represented as a Markov random field with line processes.

Scene segmentation is performed dynamically over a sequence of images by exploiting the techniques of *incremental stochastic minimization (ISM)* [3, 4] developed for motion estimation. The result is a robust segmentation of the scene into surface patches, an estimate of the intensity and motion of each patch, and a classification of the structural properties of the patch discontinuities.

Previous Work

Previous approaches to scene segmentation have typically focused on either static image segmentation or motion segmentation. Static approaches which attempt to recover surface segmentations from the 2D properties of a single image are usually not sufficient for a structural description of the scene. These techniques include the recovery of perceptually significant image properties; for example segmentation based on intensity [5, 8] or texture [9, 11], location of intensity discontinuities, and perceptual grouping of regions or edges. While there are serious limitations in using these techniques alone to recover structure, they can be used heuristically as cues to possible surface segmentations due the fact that different surfaces often have different material properties and hence may have different texture or intensity.

Structural information about image features can be gained by analyzing their behavior over time. Attempts to deal with image features in a dynamic environment have focused on the tracking of features over time [22, 29]. A notable exception to the tracking approach detects moving intensity edges over time by observing the space-time behavior of the edge

moving across a fixed detector array [16].

Motion segmentation, on the other hand, attempts to segment the scene into structurally significant regions using image motion. Early approaches focused on the segmentation and analysis of the computed flow field [26]. Other approaches have attempted to incorporate discontinuities into the flow field computation [3, 21], thus computing flow and segmenting simultaneously. There has been recent emphasis on segmenting and tracking image regions using motion, but without computing the flow field [6, 7, 10, 23]. While these approaches are promising since they provide structural information, they typically provide only a coarse segmentation of the scene.

In attempt to improve motion segmentation a number of researchers have attempted to combine intensity and motion information. Thompson [27] describes a region merging technique which uses similarity constraints on brightness and motion for segmentation. Heitz and Bouthemy [14] combine gradient based and edge based motion estimation and realize improved motion estimates and the localization of motion discontinuities. In the context of stereo reconstruction, Luo and Maître [19] use a segmented intensity image to correct and improve disparity estimates.

Advantages

The approach described here has significant advantages over single frame segmentation techniques. By extending segmentation over time the effect of noise in any single image is reduced. Only perceptual features which persist over an image sequence are recovered. Additionally, the cost of segmenting a scene is amortized over time.

Jointly modeling intensity and motion produces a more robust segmentation than can be achieved with either piece of information alone. Motion information is also used to distinguish between structural and intensity discontinuities. Furthermore, structural boundaries can be classified as occluding, disoccluding.

One of the major advantages of the approach is that it is incremental and dynamic. Feature-based approaches to motion estimation begin with static feature extraction and proceed to track the features over a number of frames. The advantage of the approach presented here is that the features themselves can be extracted dynamically over a sequence of images. Hence, we are *extracting* features over time as opposed to *tracking* them over time.

The following section formalizes the notion of a surface patch in terms of constraints on image motion and intensity. Section 3 describes the incremental minimization scheme used to estimate patch regions. Section 4 presents experimental results with real image sequences. Finally, before concluding, section 5 discusses issues with the approach and possible extensions of this work.

2 Joint Modeling of Discontinuous Intensity and Motion

To model our assumptions about the intensity structure and motion in the scene we adopt a *Markov random field (MRF)* approach [12]. MRF models have proved useful in representing

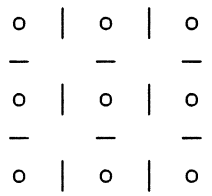


Figure 1: Arrangement of patch sites (o) and discontinuities (|, −).

the spatial properties of scenes for segmentation [9, 11], restoration [12], and motion estimation [3, 4, 18]. We formalize the prior model in terms of constraints, defined as energy functions over local neighborhoods in a grid. For an image of size $n \times n$ pixels we define a grid of *sites:*

$$S = \{s_1, s_2, \ldots, s_{n^2} \mid \forall w \; 0 \leq i(s_w), j(s_w) \leq n-1\},$$

where $(i(s), j(s))$ denotes the pixel coordinates of site s.

For the first order constraints employed here we define a *neighborhood system* $\mathcal{G} = \{\mathcal{G}_s, s \in S\}$ in terms of the nearest neighbor relations (North, South, East, West) in the grid:

$$\mathcal{G}_s = \{t \mid (i(t),\ j(t)) = (i(s) + \delta_i,\ j(s) + \delta_j)\ -1 \leq \delta_i, \delta_j \leq 1\}.$$

We define a *clique* to be a set of sites, $C \subseteq S$, such that if $s, t \in C$ and $s \neq t$, then $t \in \mathcal{G}_s$. Let $\mathcal{C}$ be a set of cliques.

We also define a "dual" lattice, $l(s,t)$, of connections between sites s and their neighboring sites $t \in \mathcal{G}_s$. This line process defines the boundaries of the surface patches. If $l(s,t) = 1$ then the sites s and t are said to belong to the same surface patch. In the case where $l(s,t) = 0$, the neighboring sites are disconnected and hence a discontinuity exists. Figure 1 illustrates the relationship between the patch sites and the boundary lattice.

Associated with each site s is a random vector $X(t) = [\mathbf{u}, i, l]$ which represents the horizontal and vertical image motion $\mathbf{u} = (u, v)$, the intensity i, and the discontinuity estimates l at time t. A discrete *state space* $\Lambda_s(t)$ defines the possible values that the random vector can take on at time t.

To model surface patches we formulate three energy terms, $E_{\mathcal{M}}$, $E_{\mathcal{I}}$, and $E_{\mathcal{L}}$ which express our prior beliefs about the motion field, the intensity structure, and the organization of discontinuities respectively. The energy terms are combined into an objective function which is to be minimized:

$$E(\mathbf{u}, \mathbf{u}^-, i, i^-, l, l^-) = E_{\mathcal{M}}(\mathbf{u}, \mathbf{u}^-, l) + E_{\mathcal{I}}(i, i^-, l) + E_{\mathcal{L}}(l, l^-). \tag{1}$$

The terms $\mathbf{u}^-$, i^-, and l^- are predicted values given the history of the sequence, and are used to express temporal continuity. In this section, we assume that these values are available. The following section will address the prediction and propagation of these values in the context of incremental minimization.

We convert the energy function, E, into a probability measure Π by exploiting the equivalence between *Gibbs distributions* [12, 17, 21] and MRF's:

$$\Pi(X(t)) = Z^{-1} e^{-E(X(t))/T(t)}, \tag{2}$$

where Z is the normalizing constant:

$$Z = \sum_{X(t)\in\Lambda(t)} e^{-E(X(t))/T(t)}, \tag{3}$$

and where $T(t)$ is a *temperature* constant at time t which serves to sharpen (or flatten) the distribution. Minimizing the objective function is equivalent to finding the maxium of Π.

2.1 The Intensity Model

We adopt a *piecewise constant*, or *weak membrane*, model of intensity [5]. This first order approximation to image intensity can easily be extended to higher order approximations [5] or to more complex texture models [11]. The current formulation differs from previous formulations in that we add a temporal continuity term to express the expected change in the image over time.

The prior model of image intensity is formulated as the energy term:

$$E_I(I, i, i^-, l, s) = \omega_{D_I} D_I(I, i, s) + \omega_{T_I} T_I(i, i^-, s) + \omega_{S_I} S_I(i, l, s), \tag{4}$$

where the ω_* are constant weights which control the relative importance of the constraints, and where the *data consistency* term is defined as:

$$D_I(I, i, s) \;=\; (I(s) - i(s))^2. \tag{5}$$

This expresses the constraint that the current estimate i should be close to the current intensity image I.

The *temporal coherence* term expresses the notion that the current estimate is related to the predicted value i^-:

$$T_I(i, i^-, s) \;=\; (i(s) - i^-(s))^2. \tag{6}$$

Finally, the *spatial coherence* term expresses an expectation of piecewise constant image patches with discontinuities:

$$S_I(i, l, s) \;=\; \sum_{n\in\mathcal{G}_s} l(s, n)(i(s) - i(n))^2. \tag{7}$$

When no discontinuity is present between sites s and n ($l(s, n) = 1$) we expect the differences in neighboring intensity values to be similar. If, however, a discontinuity is present ($l(s, n) = 0$) the difference between neighbors does not contribute to the energy term.

2.2 The Boundary Model

We want to constrain the use of discontinuities based on our expectations of how they occur in images. For example, we expect discontinuities to be rare and particular combinations to be more likely than others. Hence, we will penalize discontinuities which do not conform to

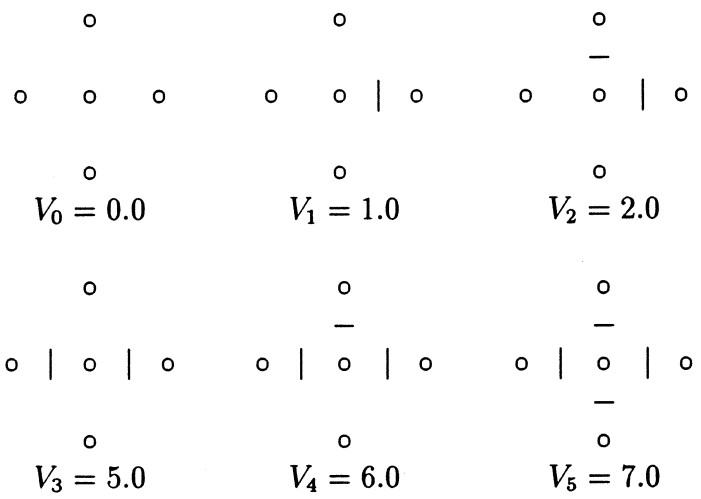


Figure 2: Examples of local surface patch discontinuities. Configuration V_0 (no discontinuities) is preferred to the situation, V_1, where a discontinuity is introduced. A corner V_2 is deemed less likely than a single discontinuity. Cliques V_3, V_4, and V_5, are highly penalized as they do not admit plausible physical interpretations.

expectations. The boundary model can then be expressed as the sum of a temporal coherence term and a penalty term defined as the sum of clique potentials V_C over a set of cliques $\mathcal{C}$:

$$E_{\mathcal{L}}(l, l^-, s) = \omega_{T_{\mathcal{L}}} \sum_{n \in \mathcal{G}_s} (l(s,n) - l^-(s,n))^2 + \omega_{P_{\mathcal{L}}} \sum_{C \in \mathcal{C}} V_C(l), \tag{8}$$

where $\omega_{T_{\mathcal{L}}}$ and $\omega_{P_{\mathcal{L}}}$ are constant weights.

One component of the penalty term expresses our expectation about the local configuration of discontinuities about a site. Figure 2 shows the possible local configurations up to rotation. We also express expectations about the local organization of boundaries; for example we express notions like "good continuation" and "closure" which correspond to assumptions about surface boundaries (figure 3). The values for these clique potentials were

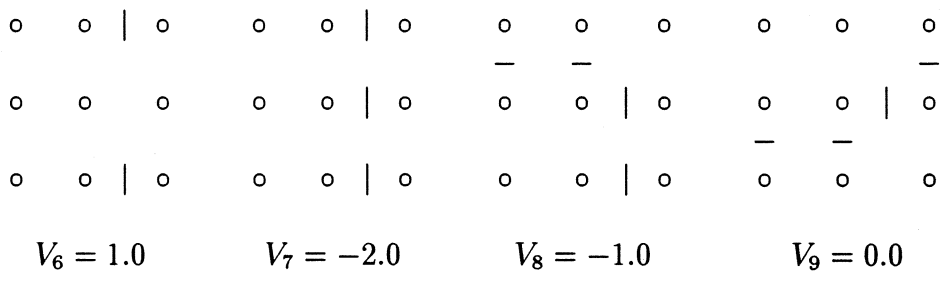


Figure 3: Examples of local organization of discontinuities based on continuity with neighboring patches. The lack of continuation in V_6 is penalized, while good continuation, V_7 is rewarded. Corners, V_8, and steps, V_9, are also rewarded.

determined experimentally and are similar to those of previous approaches [8, 21].

2.3 The Motion Model

As with the intensity model, we express our prior assumptions about the motion in terms of three constraints: data consistency, temporal coherence, and spatial coherence. The *data consistency* constraint $D_{\mathcal{M}}$ states that the image measurements corresponding to an environmental surface patch change slowly over time. The *spatial coherence* constraint $S_{\mathcal{M}}$ is derived from the observation that surfaces have spatial extent and hence neighboring points on a surface will have similar motion. Finally, the *temporal coherence* constraint $T_{\mathcal{M}}$ is based on the observation that the velocity of an image patch changes gradually over time.

This prior motion model is formulated as an energy term:

$$E_{\mathcal{M}}(I_n, I_{n+1}, \mathbf{u}, \mathbf{u}^-, l, s) = \omega_{D_{\mathcal{M}}} D_{\mathcal{M}}(I_n, I_{n+1}, \mathbf{u}, s) + \omega_{T_{\mathcal{M}}} T_{\mathcal{M}}(\mathbf{u}, \mathbf{u}^-, s) + \omega_{S_{\mathcal{M}}} S_{\mathcal{M}}(\mathbf{u}, l, s), \quad (9)$$

where the ω_* are constant weights, and where the spatial term is analogous to that of the intensity model:

$$S_{\mathcal{M}}(\mathbf{u}, l, s) \;=\; \sum_{t \in \mathcal{G}_s} l(s,t) \|\mathbf{u}(s) - \mathbf{u}(t)\|. \quad (10)$$

The temporal term, as with the other terms, is formulated in the image plane. Assuming constant acceleration, the term is formulated as:

$$T_{\mathcal{M}}(\mathbf{u}, \mathbf{u}^-, s) \;=\; \|\mathbf{u}(s) - (\mathbf{u}^-(s) + \Delta\mathbf{u}^-(s))\|, \quad (11)$$

where, at time t:

$$\Delta\mathbf{u}_t^-(s) = \mathbf{u}_t^-(s) - \mathbf{u}_{t-1}^-(s). \quad (12)$$

The data conservation constraint embodies the assumption that the intensity of a surface element remains constant over time, although its image location may change. We adopt a correlation based approach in which a correlation *surface* at a site s is defined over the space of possible displacements (u, v) with the height of the surface corresponding to an estimate of the data error of that displacement. The minimum of this surface corresponds to the best motion estimate with respect to the data conservation assumption.

Let s and t denote image locations, or sites, in S. We define a neighborhood for the data conservation constraint as :

$$\mathcal{G}_s^D = \{t \mid (i(t), j(t)) = (i(s) + \Delta i,\ j(s) + \Delta j),\ -c \le \Delta i, \Delta j \le c\},$$

which defines a square "window" of size $(-2c + 1) \times (2c + 1)$.

Data error is defined as the the difference between predicted and measured intensity values. Given image intensity functions I_n and I_{n+1} between two successive frames, the local contribution to the data conservation constraint is defined as:

$$D_{\mathcal{M}}(I_n, I_{n+1}, \mathbf{u}, s) = \sum_{t \in \mathcal{G}_s^D} \phi_D(I_n(i(t), j(t)) - I_{n+1}(i(t) + u, j(t) + v)). \quad (13)$$

where, if $\phi_D(x) = x^2$, we have the standard quadratic *sum of squared difference* surface [1]. Instead, following [3, 4], we adopt the following estimation function:

$$\phi_D(x) = \frac{-1}{1 + (x/\Delta_D)^2}, \tag{14}$$

where Δ_D is a constant scale factor ($\Delta_D = 5.0$ in our experiments). This measure is more robust in the presence of noise and outliers resulting from multiple motions within the correlation window.

3 The Computational Problem

The objective function defined in the previous section will typically have many local minima, making the task of minimizing it difficult. One possible approach to solving this minimization problem is to exploit stochastic techniques like *simulated annealing* [12, 17]. These techniques (in this case a *Gibbs Sampler* [12]) can be used to find the minimum $X(t)$ by sampling from the state space Λ according to the distribution Π with logarithmicly decreasing temperatures. As the temperature is lowered, the probability distribution Π becomes concentrated about the minimum while the stochastic nature of the process prevents the estimate from getting trapped in local minima. The result is that at high temperatures the sampling process freely chooses values of $X(t)$, but as the temperature is lowered, the minimum is chosen with increasing probability. In the limit, this process converges to the correct solution when a logarithmic cooling schedule is used. In practice, a sufficiently slow linear cooling schedule appears to provide acceptable convergence.

As mentioned earlier, each site contains random vector $X(t) = [\mathbf{u}, i, l]$ which represents the motion, intensity, and discontinuity estimates at time t. The discontinuity component of this state space is taken to be binary, so that $l \in \{0, 1\}$. While this works well in practice, it does not allow sub-pixel localization of the discontinuities. We are currently exploring ways of representing and recovering sub-pixel estimates by allowing real valued connections between sites [2, 25].

The intensity component i can take on any intensity value in the range $[0, 255]$. For efficiency, we can restrict i to take on only integer values in that range. This, however, still results in a large state space. We make the further approximation that the value of i at site s is taken from the union of intervals of intensity values about $i(s)$, the neighbors $i(t)$ of s, and the current data value $I_n(s)$. Small intervals result in a smaller state space without any apparent degradation in performance.

The motion component $\mathbf{u} = (u, v)$ is defined over a continuous range of displacements u and v. *Continuous annealing* techniques [3, 28] allow accurate sub-pixel motion estimates by making the state space for the flow component adapt to the local properties of the function being minimized.

Simulated annealing has a number of desirable properties. First, due to the local nature of the constraints, the algorithm is highly parallel, and the current Connection Machine implementation fully exploits this parallelism with a processor at each site in the MRF. More importantly, simulated annealing has the ability to cope with non-convex objective functions. Unfortunately, stochastic algorithms remain expensive, particularly without parallel hardware. For reasonable results, hundreds, or thousands, of iterations of the annealing

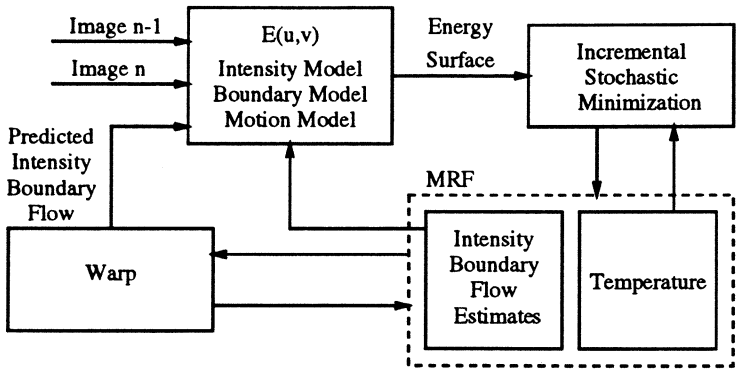


Figure 4: Incremental Stochastic Minimization.

algorithm may be necessary to settle into the global minimum energy configuration. This has a decidedly non-dynamic flavor. Ideally a motion algorithm should involve fast simple computations between a pair of frames, and exploit the fact that tremendous amounts of data are available over time.

By tracking small patches of a scene over an image sequence, we will modify the basic annealing concept to work on changing data over time. The strict convergence results of simulated annealing will be lost, but the result will be an incremental algorithm which produces good empirical results and meets many of the requirements of a truly dynamic motion algorithm.

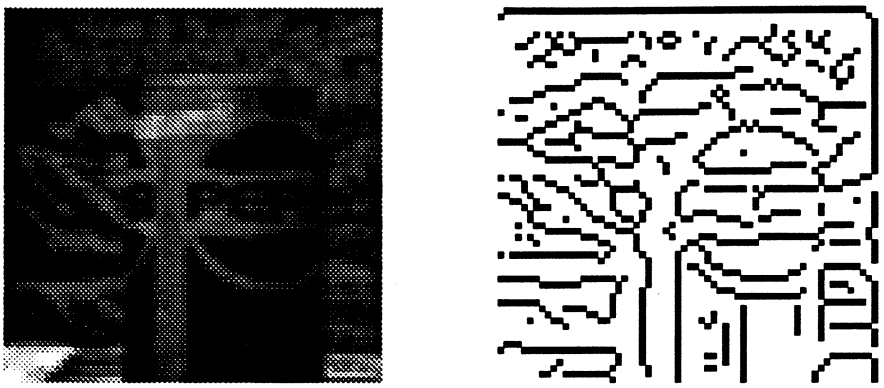
3.1 Incremental Minimization

In the context of optical flow, Black and Anandan [3] describe an *incremental stochastic minimization (ISM)* algorithm (figure 4) that has the benefits of simulated annealing without many of the shortcomings. As opposed to minimizing the objective function for a pair of frames, the ISM approach is designed to minimize an objective function which is *changing slowly over time.* The assumption of a slowly changing objective function is made possible by exploiting current motion estimates to compensate for the effects of the motion on the objective function. Estimates are propagated using the current optic flow estimate and refined with each new frame resulting in an incremental minimization algorithm. The cost of computing the motion estimate is spread over an entire sequence of images.

When a new image is acquired, the current predicted values $[\mathbf{u}^-, i^-, l^-]$ at a given site are used as the starting point for the annealing process. The current temperature at that site is used as the initial temperature, and is then lowered according to the annealing schedule.

After a fixed (usually small) number of iterations of the annealing process, each site has new estimates ($[\mathbf{u}, i, l]$) and a new temperature. The various properties of the associated surface are then propagated to the new site where the patch has moved. These properties include the patch's motion, intensity, discontinuities, temperature, and state space. This propagation can be viewed as *warping* the sites according to the motion estimate [3, 13]. Since motion is not discrete, the field is resampled using a weighted bi-linear interpolation, where the weighting reflects the confidence in the motion estimates.

During the warping process, the total flow into a site is measured. This allows us to



a *b*

Figure 5: **Can and Canny:** *a)* First image in the soda can sequence. *b)* Edges in the image extracted with the Canny edge operator.

classify motion discontinuities as occluding or disoccluding [3, 4]. In the absence of motion discontinuities the in-flow should be approximately unity. However, a site located at an occlusion boundary will have multiple sites projecting to it, thereby increasing the total in-flow. Similarly, at a disocclusion, we may expect the total flow to be less than unity. Hence, sites can be classified as locations of occlusion or disocclusion using two thresholds, one above and one below unity respectively.

A disoccluded site indicates a new patch of the environment which was previously hidden from view. For this new patch, there is no prior motion estimate, hence the annealing process should be initially uncommitted about the motion. This is achieved by initializing the site to have a high temperature.

Unlike standard annealing, the incremental algorithm uses different temperatures for the different sites and dynamically modifies the temperature according to the information available at a site. As a patch is tracked, its temperature will decrease over time. Hence, the temperatures of patches that have been tracked over many frames and whose motion is precisely known tend to be lower than those of more recently disoccluded patches.

4 Experimental Results

A number of experiments have been performed using real image sequences. For these experiments, the parameters of the model were determined empirically. The intensity model parameters were: $\omega_{D_\mathcal{I}} = \omega_{T_\mathcal{I}} = 1/40^2$ and $\omega_{S_\mathcal{I}} = 1/20^2$. For the boundary model, we set the weights as follows: $\omega_{T_\mathcal{L}} = 0.5$ and $\omega_{P_\mathcal{L}} = 1.0$. Finally, for the motion model, we have: $\omega_{D_\mathcal{M}} = 0.5$, $\omega_{T_\mathcal{M}} = 0.1$, and $\omega_{S_\mathcal{M}} = 1.5$, with a 3×3 correlation window. An initial temperature of $T(0) = 0.3$ was chosen with a linear cooling rate of $T(t+1) = T(t) - 0.0025$.

The Pepsi Sequence[1]

[1]This image sequence was provided by Joachim Heel.

Figure 6: **Feature extraction.** *a)* Intensity based segmentation without motion. *b)* Segmentation using joint intensity and motion model. *c)* Structural features in the scene.

The first sequence consists of ten 64 × 64 square images; the first image in the sequence is shown in figure 5*a*. The images contain a soda can in the foreground; the motion of which is slightly less than one pixel to the left between each frame. The can is moving in front of a textured background which is also undergoing a slight motion to the left; there is no vertical motion.

As an example of traditional, intensity-based, segmentation techniques, the Canny edge operator was applied to the image. The edges are shown in figure 5*b*. For comparison, figure 6*a* shows an intensity based segmentation using a piecewise constant intensity model with no motion information. The figure shows the estimate for a single static image after 25 iterations of the annealing algorithm. As with the Canny edges, the results correspond to intensity markings.

Figure 6*b* shows the results for the same image when a joint intensity and motion model is used. The results are from a two image sequence after 25 iterations. Compare the boundaries corresponding to the right and left edges of the can. In figure 6*a* the similarity of intensity between the can and the background results in smoothing across the object boundary. When motion information is added in figure 6*b* the object boundary is detected and smoothing does not occur across it.

Not only does the joint intensity and motion model improve the segmentation process, it provides additional information about the scene. In particular, it allows us to classify discontinuities as structural properties of the scene or purely surface markings. Figure 6*c* shows the motion boundary detected with the joint model.

The power of the approach does not lie in the ability to segment a scene using one or two frames, but rather in the ability to perform the segmentation incrementally over an image sequence. Figure 7 shows the results of processing the full ten image sequence. Figure 7*a* shows the last image in the sequence.

The horizontal and vertical motion is shown in figures 7*b* and *c* respectively. Dark areas indicate leftward or upward motion and similarly, bright areas indicate motion to the right and down. Notice that motion estimates are available in homogeneous areas where motion estimates are typically poor. Also, the modeling of discontinuities allows sharp motion

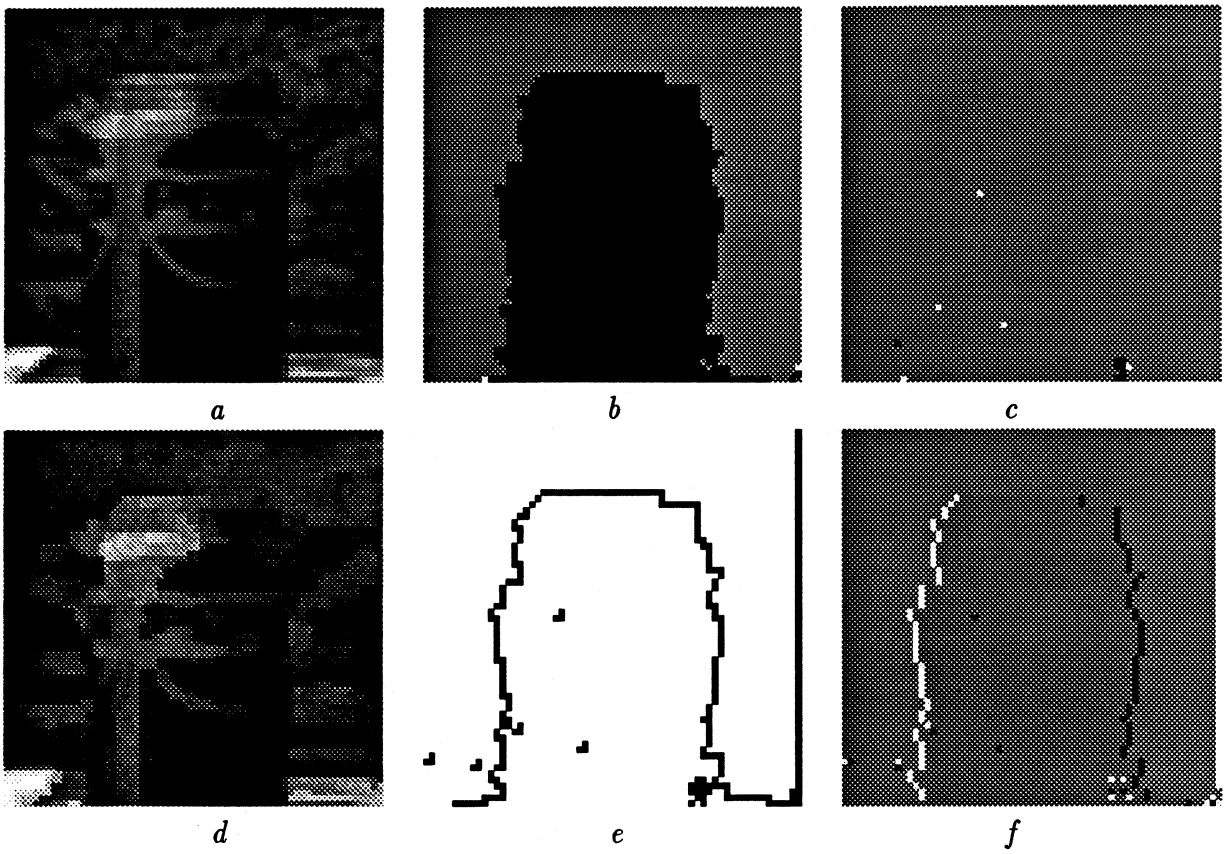


Figure 7: **Incremental Feature Extraction.** Results for a ten image sequence. *a)* Last image in the sequence. *b)* Horizontal component of image motion. *c)* Vertical component of image motion. *d)* Reconstructed intensity image. *e)* Motion boundaries. *f)* Occlusion and disocclusion boundaries.

boundaries and prevents over-smoothing.

Figure 7*d* shows the reconstructed intensity image which reflects the intensity estimates in the patches. Figure 7*e* shows the detected motion boundaries, while figure 7*f* shows the classification of the boundaries as occluding (bright areas) or disoccluding (dark areas).

Figure 8 shows the evolution of the features over the ten image sequence. The estimates start out noisy and are refined over time. Only five iterations of the annealing algorithm were used between each pair of frames. By carrying out the minimization over the sequence, the amount of computation between frames is kept small without sacrificing segmentation quality.

The knowledge of motion boundaries along with the first order flow estimates may provide enough information for many purposive vision tasks. If more detail is required, the scene can be reconstructed from the patch data. Figure 9 illustrates such a reconstruction. In figure 9*a* the disparity data and patch boundaries are used to reconstruct a segmented version of the $2\frac{1}{2}$ dimensional scene. Motion discontinuities correspond to depth discontinuities, while intensity discontinuities appear as surface markings. In figure 9*b* the intensity of the patches is used to construct a realistic rendering of the original scene.

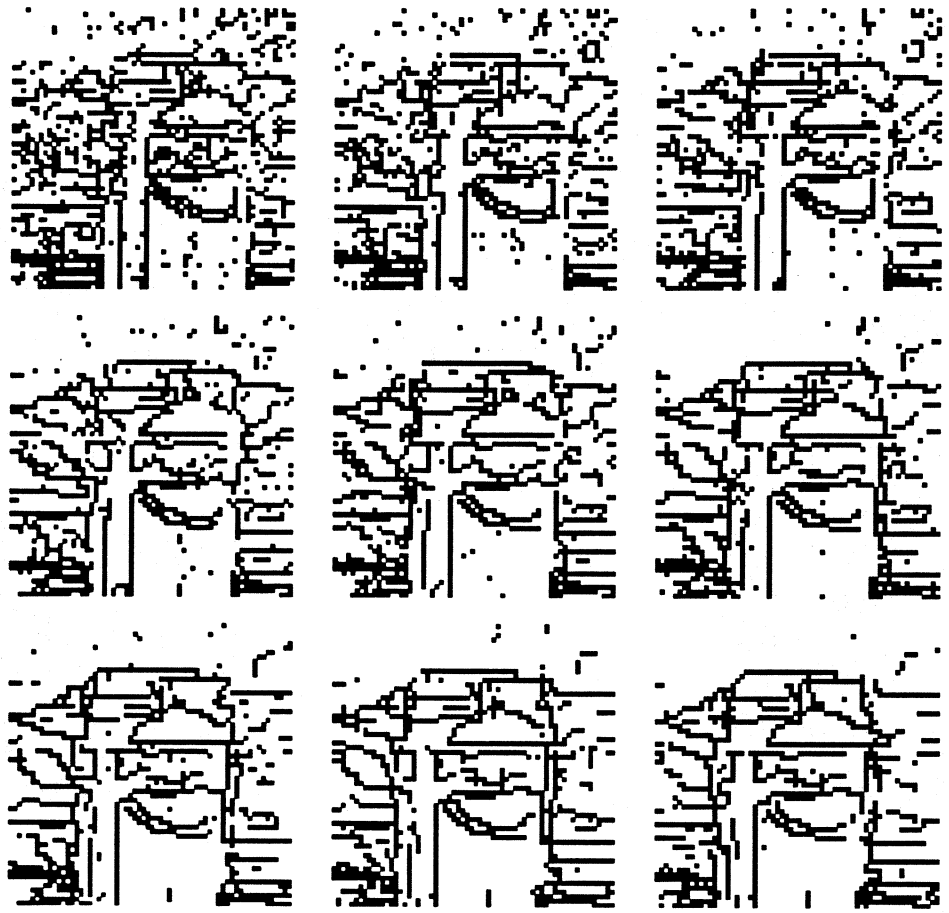


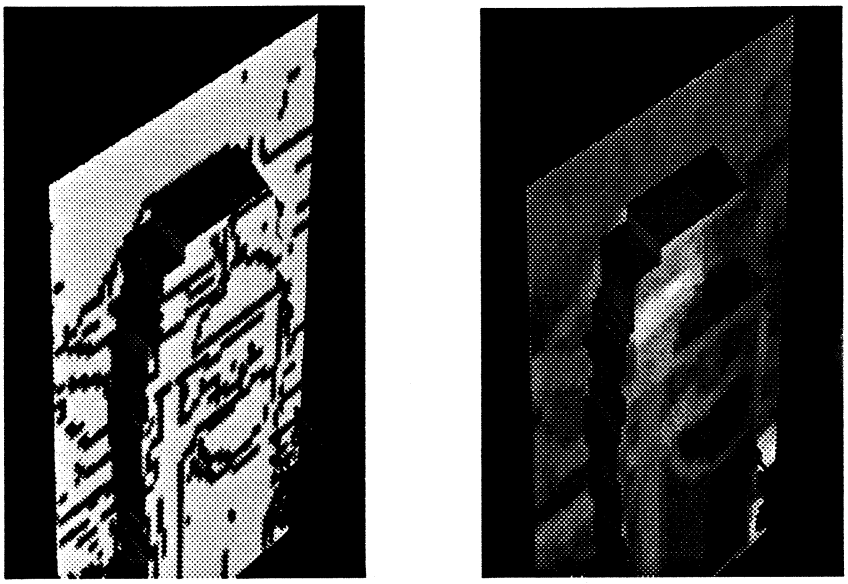
Figure 8: **Incremental Feature Extraction.** The images show the evolution (left to right, top to bottom) of features over a ten image sequence.

Equal Time for Coke[2]

The second image sequence contains 38 images of size 128 × 128 pixels. The camera is translating along the camera axis with the focus of expansion centered on the can. Figures 10*a* and *b* show the first and last images in the sequence. Figure 10*c* shows the image features at the end of the image sequence. Unlike standard segmentation, these features have been tracked over the length of the sequence. Figure 10*d* shows only features which are likely to correspond to surface boundaries. The pencils and metal bracket are correctly interpreted as physically significant while the sweater is interpreted as purely surface marking. Notice that the Coke can boundary is incorrectly interpreted as surface marking. This is a result of small interframe displacements; the motion of the can boundary is not significant enough to classify it as structural with the current scheme. This suggests the use of a different classification scheme which takes into account the behavior of features over time; we are currently exploring alternate schemes.

Figure 11 shows the evolution of the image features over time. Ten iterations of the

[2]This sequence was collected at the NASA Ames Research Center and is provided courtesy of Banavar Sridhar.



a *b*

Figure 9: Reconstructed views of the scene: *a)* intensity discontinuities, *b)* estimated intensity.

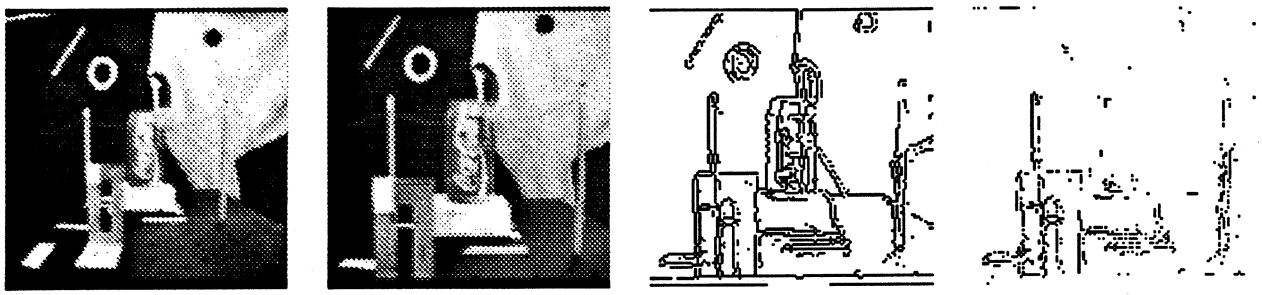
annealing algorithm were used between frames. The segmentation improves as the features are tracked over the image sequence. Due to the relatively large homogeneous regions in the image, the dense motion estimates are poor. Accurate dense flow however is not required for incremental segmentation. All that is required is that the motion estimates at the discontinuities be accurate; in fact, only accurate normal flow estimates are required.

5 Issues and Future Work

There are a number of issues to be addressed regarding the approach described. First, the current implementation employs only simple first order models of intensity and motion. While such a model may produce useful qualitative results in many situations, it is clearly not sufficient. In particular, to cope with textured surfaces more complicated image segmentation models will be required. Such an extension is straightforward as texture segmentation has been formulated by many authors in the MRF framework (see [9, 11]).

A second issue which must be addressed is one shared by many minimization approaches; that is the parameter estimation problem. The construction of an objective function with weights controlling the importance of the various terms is often based on intuition, empirical studies or, in the worst case, "tweaking" the parameters until the desired output is obtained. The problem of parameter estimation becomes more pronounced as the complexity of the model increases. In the model proposed here there are eight weights, ten clique energies, one scaling factor, an initial temperature and a cooling rate which must be determined. Experiments with the current model indicate that it is relatively insensitive to changes in the parameters. The general problem, however, remains open.

Finally, the local optimization approach to recovering surface patches is only the first step



a *b* *c* *d*

Figure 10: **The Coke Sequence.** Figures *a* and *b* are the first and last images in the sequence respectively. Figure *c* shows the image features at the end of the sequence. Figure *d* shows only those features which are likely to have a physical interpretation.

in recovering the structure of the scene. If our goal is to recover environmental structure then we must recover the *surfaces* present, their properties, and their relationships to each other. For this a more accurate modeling of surfaces will be required; for example 3D parameterized surface models [15]. Non-local properties of the patches will need to be computed and additional perceptual organization processes will likely be needed to group patches which are consistent with the surface models.

6 Conclusion

We have presented an incremental approach to extracting stable perceptual features over time. The approach formulates a model of surface patches in terms of constraints on intensity and motion while accounting for discontinuities. An incremental minimization scheme is used to segment the scene over a sequence of images.

The approach has advantages over traditional segmentation and motion estimation techniques. In particular, it is incremental and dynamic. This allows segmentation and motion estimation to be performed over time, while reducing the amount of computation between frames and increasing robustness.

Additionally, the approach provides information about the structural properties of the scene. While intensity based segmentation alone provides information about the spatial structure of the image, motion provides information about object boundaries. Motion segmentation alone, however, provides fairly coarse information. Combining the two types of information provides a richer description of the scene.

References

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

[2] Ballard, D. H., "Interpolation coding: A representation for numbers in neural models," *Biol. Cybern., 57,* pp. 389–402, 1987.

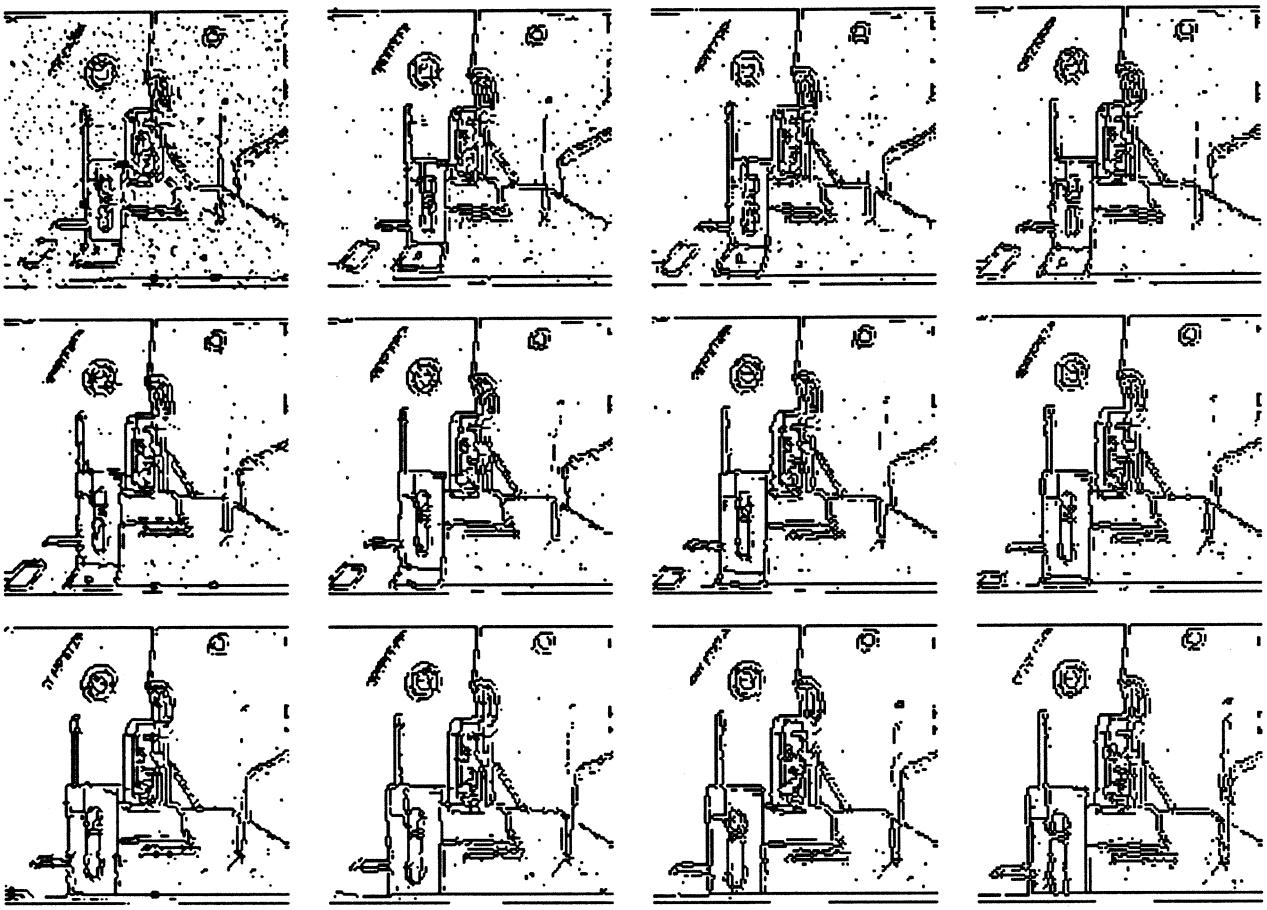


Figure 11: **Incremental Feature Extraction.** The sequence shows the evolution (left to right, top to bottom) of features at every third image in the 38 image sequence.

[3] Black, M. J., and Anandan, P., "Robust dynamic motion estimation over time," *Proc. Comp. Vision and Pattern Recognition*, CVPR-91, Maui, Hawaii, June 1991, pp. 296–302.

[4] Black, M. J., and Anandan, P., "A model for the detection of motion over time," *Proc. Int. Conf. on Comp. Vision*, ICCV-90, Osaka, Japan, Dec. 1990, pp. 33–37.

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

[6] Bouthemy, P. and Lalande, P., "Detection and tracking of moving objects based on a statistical regularization method in space and time," *Proc. First European Conf. on Computer Vision*, ECCV-90, Antibes, France, April 1990, pp. 307–311.

[7] Bouthemy, P. and Rivero, J. S., "A heirarchical likelihood approach for region segmentation according to motion-based criteria," *Proc. First Int. Conf. on Computer Vision*, ICCV-87, June 1987, pp. 463–467.

[8] Chou, P. B., and Brown, C. M., "The theory and practice of bayesian image labeling," *Int. Journal of Computer Vision,* Vol. 4, No. 3, 1990, pp. 185–210.

[9] Derin, H. and Elliott, H., "Modeling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-9, No. 1, January 1984, pp. 39–55.

[10] François, E. and Bouthemy, P., "Multiframe-based identification of mobile components of a scene with a moving camera," *Proc. Comp. Vision and Pattern Recognition,* CVPR-91, Maui, Hawaii, June 1991, pp. 166–172.

[11] Geman, D., Geman, S., Graffigne, C., and Dong, P., "Boundary detection by constrained optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 12, No. 7, July 1990, pp. 609–628.

[12] Geman, S. and Geman, D., "Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-6, No. 6, November 1984.

[13] Heel, J., "Temporally integrated surface reconstruction," *Proc. Int. Conf. on Comp. Vision,* ICCV-90, Osaka, Japan, Dec. 1990, pp. 292–295.

[14] Heitz, F. and Bouthemy, P., "Multimodal motion estimation and segmentation using Markov random fields," *Proc. IEEE Int. Conf. on Pattern Recognition,* June, 1990, pp. 378–383.

[15] Hung, Y., Cooper, D. B., Cernuschi-Frias, B., "Bayseian estimation of 3-D surfaces from a sequence of images," *Proc. IEEE Int. Conf. on Robotics and Automation,* April, 1988, pp. 906–911.

[16] Kahn, P., "Integrating moving edge information along a 2D trajectory in densely sampled imagery," *IEEE Proc. Comp. Vision and Pattern Recognition,* CVPR-88, June, 1988, pp. 702–709.

[17] Kirkpatrick, S., Gelatt, C. D. Jr., and Vecchi, M. P., "Optimization by simulated annealing," *Science,* Vol. 220, No. 4598, May 1983, pp. 671–680.

[18] Konrad, J., "Bayesian estimation of motion fields from image sequences," *Ph.D. Dissertation,* McGill University, Montreal, Canadan, June 1989.

[19] Luo, W. and Maître, H., "Using surface model to correct and fit disparity data in stereo vision," *Proc. IEEE Int. Conf. on Pattern Recognition,* June 1990, pp. 60–64.

[20] Matthies, L., Szeliski, R., Kanade, T., "Kalman filter-based algorithms for estimating depth from image sequences," *Int. J. of Computer Vision,* 3(3), Sept. 1989, pp. 209–236.

[21] Murray, D. W. and Buxton, B. F., "Scene segmentation from visual motion using global optimization," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. PAMI-9, No. 2, March 1987, pp. 220–228.

[22] Navab, N., Deriche, R., and Faugeras, O. D., "Recovering 3D motion and structure from stereo and 2D token tracking cooperation," *Proc. Int. Conf. on Comp. Vision*, ICCV-90, Osaka, Japan, Dec. 1990, pp. 513–516.

[23] Peleg, S. and Rom, H., "Motion based segmentation," *Proc. IEEE Int. Conf. on Pattern Recognition,* June 1990, pp. 109–113.

[24] Singh, A., "Incremental estimation of image-flow using a Kalman filter," to appear, *Proc. IEEE Workshop on Visual Motion,* Princeton, NJ, Oct. 1991.

[25] Szeliski, R. S., "Bayesian modeling of uncertainty in low-level vision," Ph.D. Thesis, Carnegie Mellon University, 1988.

[26] Thompson, W. B., Mutch, K. M., and Berzins, V. A., "Dynamic occlusion analysis in optical flow fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-7, No. 4, July 1985, pp. 374–383.

[27] Thompson, W. B., "Combining motion and contrast for segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-2 1980, pp. 543–549.

[28] 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.

[29] Viéville, T. and Faugeras, O., "Feed-forward recovery of motion and structure from a sequence of 2D-lines matches," *Proc. Int. Conf. on Comp. Vision*, ICCV-90, Osaka, Japan, Dec. 1990, pp. 517–520.