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

Abstract. *This paper presents a method for incrementally segmenting images over time using both intensity and motion information. This is done by formulating a model of physically significant image resgions 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.*

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 field models [7], recovering discontinuities [2], segmentation [6], motion estimation [1], motion segmentation [3, 5, 8, 10], and incremental algorithms [1, 9] 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 image regions using local constraints on intensity and motion. These regions correspond to the location of possible surface patches in the image plane. The formulation of the constraints accounts for surface patch boundaries as discontinuities in intensity and motion. The segmentation problem is then modeled as a Markov random field with line processes.

* This work was supported in part by a grants from the National Aeronautics and Space Administration (NGT–50749 and NASA RTOP 506–47), by ONR Grant N00014–91–J–1577, and by a grant from the Whitaker Foundation.

Scene segmentation is performed dynamically over a sequence of images by exploiting the technique of *incremental stochastic minimization (ISM)* [1] developed for motion estimation. The result is a robust segmentation of the scene into physically meaningful image regions, an estimate of the intensity and motion of each patch, and a classification of the structural properties of the patch discontinuities.

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 [2, 4] or texture [6], location of intensity discontinuities, and perceptual grouping of regions or edges. 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 [11].

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. Other approaches have attempted to incorporate discontinuities into the flow field computation [1, 10], 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 [3, 5].

In attempt to improve motion segmentation a number of researchers have attempted to combine intensity and motion information. Thompson [12] describes a region merging technique which uses similarity constraints on brightness and motion for segmentation. Heitz and Bouthemy [8] combine gradient based and edge based motion estimation and realize improved motion estimates and the localization of motion discontinuities.

The following section formalizes the notion of a surface patch in the image plane 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 regarding the approach.

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 [7]. 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. 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* [7] defines the boundaries of the image patches. If $l(s,t) = 1$ then the sites s and t are said to belong to the same image patch. In the case where $l(s,t) = 0$, the neighboring sites are disconnected and hence a discontinuity exists.

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.

We convert the energy function, E, into a probability measure Π by exploiting the equivalence between *Gibbs distributions* [7, 10] and MRF's:

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

where Z is the normalizing constant, and where $T(t)$ is a *temperature* constant at time t. Minimizing the objective function is equivalent to finding the maximum of Π.

The constraints are summarized in figure 1 and described briefly below:

The Intensity Model: We adopt a *weak membrane* model of intensity [2]. The data consistency term $D_{\mathcal{I}}$ keeps the estimate close to the data while the term $S_{\mathcal{I}}$ enforces spatial smoothness. The current formulation differs from previous approaches in that we add a temporal continuity $T_{\mathcal{I}}$ term to express the expected change in the image over time.

The Boundary Model: We want to constrain the use of discontinuities based on our expectations of how they occur in images. Hence, we will penalize discontinuities which do not conform to expectations. The boundary model is 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 $\mathcal{C}$.

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 determined experimentally and are similar to those of previous approaches [4, 10].

The Motion Model: As with the intensity model, we express our prior assumptions about the motion in terms of three constraints. 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.

Intensity Model

$$E_{\mathcal{I}}(I,i,i^-,l,s) = \omega_{D_{\mathcal{I}}} D_{\mathcal{I}}(I,i,s) + \omega_{T_{\mathcal{I}}} T_{\mathcal{I}}(i,i^-,s) + \omega_{S_{\mathcal{I}}} S_{\mathcal{I}}(i,l,s) \tag{3}$$
$$D_{\mathcal{I}}(I,i,s) = (I(s) - i(s))^2 \tag{4}$$
$$T_{\mathcal{I}}(i,i^-,s) = (i(s) - i^-(s))^2 \tag{5}$$
$$S_{\mathcal{I}}(i,l,s) = \sum_{n \in \mathcal{G}_s} l(s,n)(i(s) - i(n))^2 \tag{6}$$

Boundary Model

$$E_{\mathcal{L}}(l,l^-,s) = \omega_{T_{\mathcal{L}}} \sum_{n \in \mathcal{G}_s} (l(s,n) - l^-(s,n))^2 + \omega_{P_{\mathcal{L}}} \sum_{C \in \mathcal{C}} V_C(l) \tag{7}$$

Motion Model

$$E_{\mathcal{M}}(I_n, I_{n+1}, \mathbf{u}, \mathbf{u}^-, l, s) =$$
$$\omega_{D_{\mathcal{M}}} D_{\mathcal{M}}(I_n, I_{n+1}, \mathbf{u}, s) + \omega_{T_{\mathcal{M}}} T_{\mathcal{M}}(\mathbf{u}, \mathbf{u}^-, s) + \omega_{S_{\mathcal{M}}} S_{\mathcal{M}}(\mathbf{u}, l, s) \tag{8}$$
$$D_{\mathcal{M}}(I_n, I_{n+1}, \mathbf{u}, s) = \sum_{t \in \mathcal{G}_s} \phi_D(I_n(i_t, j_t) - I_{n+1}(i_t + u, j_t + v)) \tag{9}$$
$$S_{\mathcal{M}}(\mathbf{u}, l, s) = \sum_{t \in \mathcal{G}_s} l(s,t) \|\mathbf{u}(s) - \mathbf{u}(t)\| \tag{10}$$
$$T_{\mathcal{M}}(\mathbf{u}, \mathbf{u}^-, s) = \|\mathbf{u}(s) - (\mathbf{u}^-(s) + \Delta\mathbf{u}^-(s))\| \tag{11}$$
$$\Delta\mathbf{u}_t^-(s) = \mathbf{u}_t^-(s) - \mathbf{u}_{t-1}^-(s) \tag{12}$$

Miscellaneous

$$\mathcal{G}_s^D = \{t \mid (i_t, j_t) = (i_s + \Delta i,\ j_s + \Delta j),\ \ -c \le \Delta i, \Delta j \le c\} \tag{13}$$
$$\mathcal{G}_s = \{t \mid (i_t,\ j_t) = (i_s + \delta_i,\ j_s + \delta_j),\ \ -1 \le \delta_i, \delta_j \le 1\} \tag{14}$$
$$\phi_D(x) = \frac{-1}{1 + (x/\Delta_D)^2} \tag{15}$$

Fig. 1. Robust constraints on image motion.

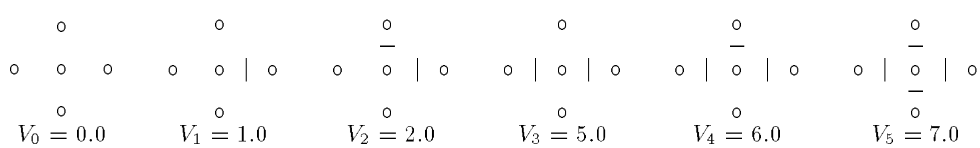

Fig. 2. Examples of local surface patch discontinuities; (sites: ($\circ$), discontinuities: ($|, -$)).

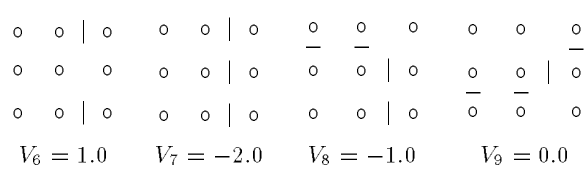

Fig. 3. Examples of local organization of discontinuities based on continuity with neighboring patches.

3 The Computational Problem

The objective function defined in the previous section will typically have many local minima. Simulated annealing (in this case a *Gibbs Sampler* [7]) 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 mentioned earlier, each site contains a 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 $l \in \{0, 1\}$.

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. 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 [1] 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.

3.1 Incremental Minimization

Unfortunately, stochastic algorithms remain expensive, particularly without parallel hardware, making them ill-suited to dynamic problems. 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.

In the context of optical flow, Black and Anandan [1] describe an *incremental stochastic minimization (ISM)* algorithm (figure 4) that has the benefits of simulated annealing

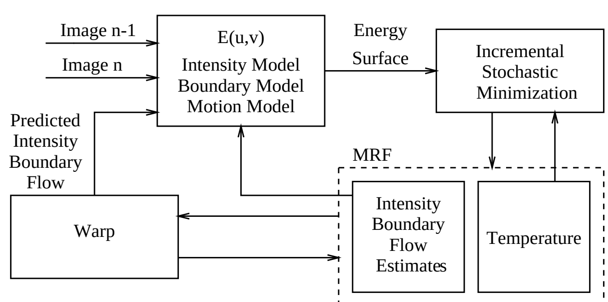

Fig. 4. Incremental Stochastic Minimization.

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. With each new image, current estimates are propagated by *warping* the grid of sites using the current optic flow estimate. The estimates are then refined using traditional stochastic minimization techniques. Additionally, during the warping process motion discontinuities are classified as occluding or disoccluding.

4 Experimental Results

The system is implemented in *Lisp on a 8K node Connection Machine (CM-2). 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, the weights were: $\omega_{T_\mathcal{L}} = 0.5$ 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 cooling rate of $T(t+1) = T(t) - 0.0025$ and Δ_D was set to 5.0.

The Pepsi Sequence[1] The first sequence consists of ten 64×64 square images; the first image in the sequence is shown in figure 5*a*. The Canny edge operator was applied to the image and the edges are shown in figure 5*b*. For comparison, figure 5*c* 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 5*d* 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 5*c* the similarity of intensity between the can and the background results in smoothing across the object boundary. When motion information is added in figure 5*d* the object boundary is detected (figure 5*e*) and smoothing does not occur across it.

Figures 5*f*–5*l* show the results of incrementally processing the full ten image sequence. Figure 5*f* shows the last image in the sequence. The horizontal and vertical motion is shown in figures 5*g* and 5*h* respectively. Dark areas indicate leftward or upward motion and similarly, bright areas indicate motion to the right and down. Figure 5*i* shows the intensity estimates of the patches and figure 5*j* shows the discontinuities. Figure 5*k* shows the detected motion boundaries, while figure 5*l* shows the classification of the boundaries as occluding (bright areas) or disoccluding (dark areas). Figure 6 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. The processing time for each frame was approximately 30 seconds.

The Coke Sequence[2] The second image sequence contains 38 images of size 128×128 pixels. Figures 7*a* and *b* show the first and last images in the sequence respectively. Figure 7*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 7*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. Figure 8 shows the evolution of the image features over time. The segmentation improves as the features are tracked over the image sequence. Five iterations of the annealing algorithm were used between frames with a processing time of approximately one minute per frame.

[1] This image sequence was provided by Joachim Heel.
[2] This sequence was provided by Dr. Banavar Sridhar at the NASA Ames Research Center.

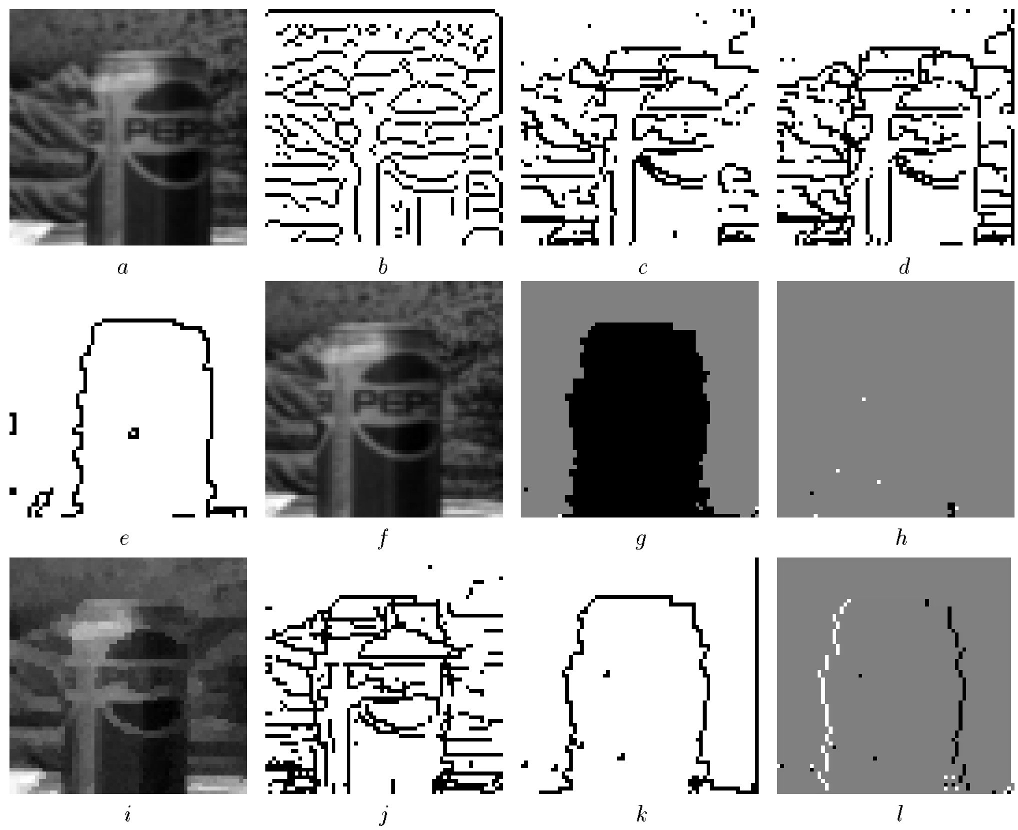

Fig. 5. Feature Extraction: *a)* First image in the Pepsi can sequence. *b)* Edges in the image extracted with the Canny edge operator. *c)* Intensity based segmentation without motion. *d)* Segmentation using joint intensity and motion model. *e)* Structural features in the scene. *f)* Last image in the sequence. *g)* Horizontal component of image motion. *h)* Vertical component of image motion. *i)* Reconstructed intensity image. *j)* Final patch boundaries. *k)* Motion boundaries. *l)* Occlusion and disocclusion boundaries.

Fig. 6. Incremental Feature Extraction. The images show the evolution (left to right, top to bottom) of features over a ten image sequence.

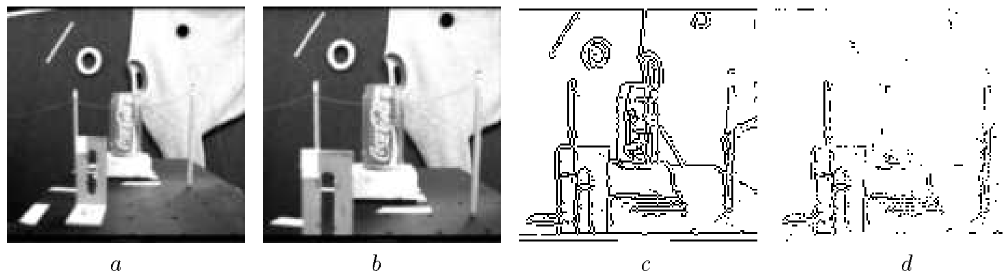

Fig. 7. **The Coke Sequence.** *a, b)* first and last images in the sequence, *c)* image features at the end of the sequence, *d)* those features which are likely to have a physical interpretation.

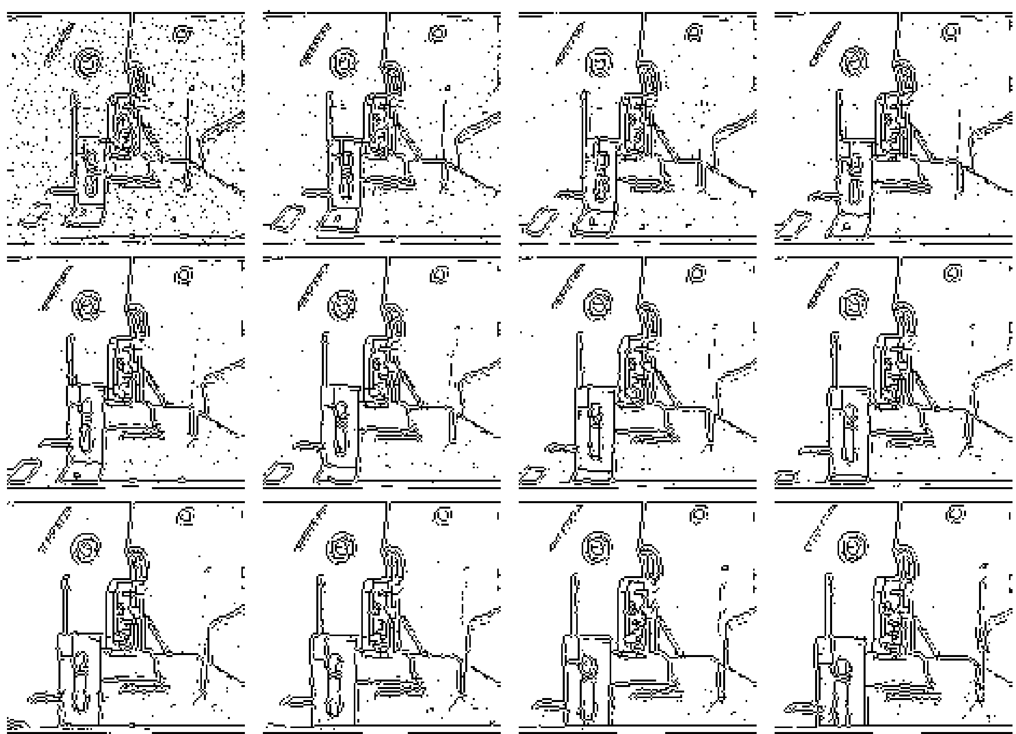

Fig. 8. Incremental Feature Extraction. The sequence shows the evolution (left to right, top to bottom) of features at every third image in the 38 image sequence.

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. To cope with textured surfaces more complicated image segmentation models will be required.

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 or empirical studies. The problem becomes more pronounced as the complexity of the model increases. Experiments with the current model indicate that it is relatively insensitive to changes in the parameters.

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. Combining the two types of information provides a richer description of the scene.

References

1. 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.
2. Blake, A. and Zisserman, A., *Visual Reconstruction,* The MIT Press, Cambridge, Massachusetts, 1987.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Thompson, W. B., "Combining motion and contrast for segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-2 1980, pp. 543–549.

This article was processed using the LaTeX macro package with ECCV92 style