Scene Segmentation from Visual Motion Using Global Optimization

DAVID W. MURRAY AND BERNARD F. BUXTON, MEMBER, IEEE

Abstract—This paper presents results from computer experiments with an algorithm to perform scene disposition and motion segmentation from visual motion or optic flow. The maximum a posteriori (MAP) criterion is used to formulate what the best segmentation or interpretation of the scene should be, where the scene is assumed to be made up of some fixed number of moving planar surface patches. The Bayesian approach requires, first, specification of prior expectations for the optic flow field, which here is modeled as spatial and temporal Markov random fields; and, secondly, a way of measuring how well the segmentation predicts the measured flow field. The Markov random fields incorporate the physical constraints that objects and their images are probably spatially continuous, and that their images are likely to move quite smoothly across the image plane. To compute the flow predicted by the segmentation, a recent method for reconstructing the motion and orientation of planar surface facets is used. The search for the globally optimal segmentation is performed using simulated annealing.

Index Terms—Global optimization, MAP criterion, Markov random fields, optic flow, segmentation, simulated annealing, structure from motion.

I. INTRODUCTION

The visual motion, or optic flow, derived from a sequence of time-varying two-dimensional images is a valuable source of information on the three-dimensional nature of the scene being viewed. At the most basic level, visual motion can be used simply to detect and flag motion in the viewed scene, whereas from more elaborate processing it is possible to derive detailed information about the motion, position, and orientation of a visible structure or surface. This capability has been demonstrated in the human visual system in various psychological experiments (e.g., [1]-[3]) and in recent years exploited in several computational vision schemes (cf. [4]-[16]).

If this detailed type of processing were extended to allow the distinction of several structures or surfaces in the scene, it would provide a low-level method for the segmentation of the scene into primitives labeled by disposition and, probably more importantly, motion. Such a scheme is attractive, as visual motion appears to be a powerful cue for segmentation in the human visual system. This is illustrated, for example, by the ease with which we see an otherwise successfully camouflaged creature as soon as it moves and by simple demonstrations with interpenetrating dot pictures. In these, the rough dotted outline of some object is superimposed on a randomly dotted background. When the pictures are stationary, no outline is apparent, but when they are moved relative to each other the outline is immediately and surprisingly obvious. This test, moreover, illustrates a further aspect of motion: that it provides a strong cohesive or "fusing" influence between disjoint portions of the image and scene that are moving similarly.

Although the importance of visual motion as a cue for segmentation is widely accepted [4], [6], [17], the emphasis of much of the recent work on optic flow has been on methods of reconstructing the three-dimensional scene from its two-dimensional flow field, and it has usually been assumed that the scene contains only one moving structure or surface. In contrast, more attention was given to the segmentation problem in earlier research [4], [5], [18], but then the flow-to-scene reconstructions used involved severe simplifying assumptions, such as that in [5] where the only scene motion considered was translation parallel to the image plane, or in [18] where the image was constructed under orthographic projection. A recent exception where realistic reconstruction and the motions of several surfaces have been considered is the work of Adin [19]. In this, the flow field is decomposed into small areas, in each of which the field approximates that expected from a moving planar facet in the scene. This method however suffers the philosophical disadvantage that the resulting mosaic of many pieces must then be grouped or fused, on the basis of similar motion and structure, into larger surfaces using a multipass Hough transform technique.

In this paper we present some results from experiments with an algorithm which includes this "fusion" stage at an earlier level by taking a more global view of the flow field. It builds on some of the ideas expressed in a recent paper [20] in which scene and motion segmentation is discussed in terms of regularizing an ill-posed problem within the framework of maximizing conditional likelihood. (Poggio, Torre, and Koch [21], [22] have propounded the regularization approach for a wide range of vision problems.) Our segmentation involves a search for the maximum of the a posteriori probability of an interpretation of a field of optic flow data. By interpretation here, we mean a collection of facts like "this datum of optic flow arose
from that surface in the scene.' The segmentation is thus scene-based and occurs at the same time as the flow-to-scene reconstruction, and is neither a precursor nor postlude to it.

II. FORMULATING A SOLUTION USING THE MAP CRITERION

The fundamental step in constructing a solution is to specify what the "best" segmentation or interpretation should be. We have adopted the maximum a posteriori (MAP) criterion. This requires us to state both how well the current interpretation explains the measured data, and how well the interpretation conforms to our prior expectations of a sensibly organized flow field. We shall discuss these two requirements in Sections III and IV, but, in brief, we model the flow data on the assumption that the scene comprises moving planar facets (although the segmentation method per se is not restricted to any particular scene model) and expect the interpretation to be a spatio-temporal Markov random field (MRF). The search for the maximum in the a posteriori distribution is made using the global optimization technique as described in the introductory section. This approach was stimulated by the work of Geman and Geman [23], who modeled image intensities as a spatial MRF and used stochastic relaxation to restore noisy gray-level images.

Within the MAP criterion we wish to maximize the probability of the scene-based interpretation \( X \) of the flow field given the optic flow data \( D \); that is we search the interpretation set \( \Omega \) for the interpretation \( X = \omega \) which maximizes

\[
(P(X = \omega | D)).
\]

This may be rewritten using Bayes’ theorem as

\[
P(X = \omega | D) = \frac{P(D | X = \omega) P(X = \omega)}{P(D)},
\]

where \( P(D | X) \) is the probability of the data given the interpretation, \( P(X) \) is the prior probability of the interpretation, and \( P(D) \) is the prior probability of the data. This last term is constant as far as the maximization is concerned, and may be ignored.

In a noise-free system the data would be related exactly to the interpretation by some transformation \( \phi \) so that \( D = \phi(X) \), and there would of course be only one interpretation for which \( P(D | X) \) is nonzero. In the presence of noise \( n \), however, the joint probability is related to the noise distribution \( P_n(n) \) simply by

\[
P(D | X) = P_n(n).
\]

Assuming that the noise is Gaussian, has zero mean, and is independent at each of the \( M \) data points, and has the same standard deviation \( \sigma \), a univariate distribution may be used so that

\[
P_n(n) = \exp \left( -\frac{\|n\|^2}{2\sigma^2} \right) / (2\pi \sigma^2)^{M/2},
\]

where

\[
\|n\|^2 = \sum_{i=1}^{M} (n_i)^2.
\]

The noise contribution for an individual point is

\[
n_i = D_i - [\phi(X)]_i
\]

which, in the present problem, depends on the way the optic flow data are distributed among the various scene facets.

\( P(X) \), the prior probability of the interpretation, describes by how much the interpretation meets our expectations. Here we model the interpretation as a Markov random field, which effectively introduces local constraints on the interpretation. The constraints express physical expectations and will be discussed in Section IV.

It may be shown that if \( X \) is an MRF with respect to some local neighborhood then it is also a Gibbs distribution with respect to the neighborhood (e.g., [24]), and thus \( P(X) \) can be expressed in terms of a potential function \( U(\omega) \) as:

\[
P(X = \omega) = \exp \left( -U(\omega) \right) / Z,
\]

where \( Z \) is the partition function

\[
Z = \sum_{\omega \in \Omega} \exp \left( -U(\omega) \right).
\]

The potential \( U(\omega) \) is given by the sum of local potentials whose specification will be deferred until Section IV.

Ignoring \( P(D) \) and taking logarithms of (II.2), our task has become that of minimizing the function

\[
(U(\omega) + \|n\|^2/2\sigma^2 + M \log \sigma + \log Z).
\]

The second term of (II.9) expresses the familiar least-squares condition and the first is like a regularizing potential for an ill-posed problem. The third and fourth are less familiar, because they are usually omitted on the assumption that \( Z \) and \( \sigma \) are constant—that is, not part of the interpretation. We make this assumption here, which means that we assume the noise level is fixed, that there are no interpretable parameters in the potential and finally that the definition of "interpretation" is itself fixed. The roles of these two terms have been discussed in [20], and although the \( M \log \sigma \) term presents no particular difficulties, computing the partition function is impractical because the number of possible interpretations in \( \Omega \) is enormous. Furthermore, such partition functions are typically very difficult, if not impossible, to derive analytically.

III. COMPUTING \( P(D | X) \) USING THE FLOW-TO-SCENE RECONSTRUCTION

At any stage in the search for the maximum of the a posteriori distribution, the current segmentation or interpretation \( X \) of the data \( D \) consists of a collection of assignments of optic flow data to the various surface facets in the scene. Taking the set of flow vectors assigned to one facet, we can perform a flow-to-scene reconstruction in a least squares manner to obtain the facet’s structure and motion parameters and then back-transform to find the predicted flow vectors at the appropriate sites in the image. The deviations of the predicted flow from the observed flow \( D \) give a contribution to the noise term. Here we take the facets to be planar and use a recent method
of reconstructing their structure and motion parameters [14], but an alternative would be, for example, to use Waxman and Wohlin's curved surface patch model [15]. Before explaining the planar facet model in more detail, we first indicate the exact nature of the flow data we assume.

Two major paradigms for computing visual motion can be distinguished. The first embraces token-matching or correspondence schemes [25]-[27], in which significant irradiance features in the image are identified, matched, and followed over time. One characteristic of these methods is that, if the tokens are highly distinctive, they give a full flow field, but at a sparse set of image points. The other paradigm includes gradient-based techniques [28]-[30], in which spatio-temporal changes in the image irradiance are used to obtain information about the flow locally. These methods yield information more uniformly throughout an image, but only provide a partial optic flow vector, namely the component along the direction of the spatial gradient of the intensity, that is, perpendicular to an image edge-feature. In the present context, gradient-based schemes have a distinct advantage over token matching. Many of the latter schemes are not local, and so frequently have implicit in them an image segmentation, which is precisely one of the things for which we want to use the optic flow itself. Thus we assume in these experiments a dense field of edge normal flow vectors.

Following Buxton et al. [14] we assume that the scene points have coordinates

\[
R = (X, Y, Z) \tag{III.1}
\]

and lie on a planar facet in space, described by the equation

\[
R \cdot N = 1, \tag{III.2}
\]

which moves with rectilinear and angular velocities \(V\) and \(\Omega\) so that

\[
\dot{R} = V + \Omega \times R. \tag{III.3}
\]

All these vectors are relative to the sensor optics, which obey perspective geometry, have a pinhole at \((0, 0, 0)\), have focal length \(l\), and have their optic axis along the \(z\) axis. The scene point is imaged at

\[
r = -IR/Z = (x, y, -l). \tag{III.4}
\]

The full optic flow \(\dot{r}\) is the time differential of (III.4), but the edge motion is the projection of this onto the normal to the edge contour. Denoting this edge motion, or vernier velocity, by \(v\), it follows from equations (III.1)-(III.4) that at each edge point \(i\) [14]:

\[
|v_i| = \sum_{k=1}^{8} c_k \beta_k. \tag{III.5}
\]

The quantities \(c_k\) and the eight parameters \(\beta_k\) are written out in the Appendix. The important point to note is that the \(c\) values depend only on measured image quantities \((x, y, v)\), and that the \(\beta\) parameters depend only on the scene structure and motion parameters \((V, N, \Omega)\). We can therefore solve for \(\beta\) either exactly for \(M = 8\) points, or in the classical least squares sense for \(M > 8\) points [31] from:

\[
[e^T][v] = [e^T][\beta], \tag{III.6}
\]

where \([v]\) is a column vector of length \(M\), \([e]\) is an \(M \times 8\) matrix, and \([\beta]\) is a column vector of length eight.

As we shall see, the \([\beta]\) vector alone suffices to compute the predicted data, but the structure and motion parameters which can be found from \([\beta]\) are the direction of the rectilinear motion \(V\) (where \(V\) is a unit vector in the direction of \(V\)), the angular velocity \(\Omega\), the direction of the surface orientation \(N\), and the so-called relative depth \(V_N\). It is not possible of course to find both \(V\) and \(N\) fully because of the inherent depth/speed scaling in visual motion under perspective projection. There is sometimes also a further ambiguity involving interchange of \(V\) and \(N\) which has been described by Longuet-Higgins [13], but again we stress that ambiguity in the interpretation of \(\beta\) does not affect the predicted flow.

Given \([\beta]\), the predicted full flow \(\dot{r}'\) at position \((x, y)\) in the image is found [32] from

\[
\dot{x}' = \beta_1 x + \beta_2 y - \beta_3 l + x(\beta_4 x + \beta_5 y)/l
\]

\[
\dot{y}' = \beta_4 x + \beta_5 y - \beta_6 l + y(\beta_7 x + \beta_8 y)/l. \tag{III.7}
\]

and the predicted vernier velocity \(v'\) is then

\[
v' = ((v' \cdot v)/v^2) v. \tag{III.8}
\]

The square of its deviation from \(v\) is

\[
\delta_i^2 = (v' - v) \cdot (v' - v), \tag{III.9}
\]

and is summed over all the \(m_i\) points assigned to facet \(k\) to give a contribution to the noise norm from that facet of

\[
\|n\|^2_k = \sum_{i=1}^{m_i} \delta_i^2. \tag{III.10}
\]

The form of this noise term is rather different from that used, say, in the work of Geman and Geman [23]. Our term is not local, in that the predicted flow at a site depends on all the measured flows assigned to a particular surface facet by the segmentation. The effect of this is felt on the optimization process, as we shall discuss.

IV. PRIOR EXPECTATIONS FOR ORGANIZED FLOW FIELDS

We have chosen to model our expectation for the interpretation of the flow field as an MRF. What this means for our problem is that the conditional probability that flow at some site be labeled as coming from a particular facet depends only on the labels attached to the sites in some finite neighborhood. The MRF model thus imposes local constraints on the segmentation which if they are to be useful should reflect real physical expectations. Two constraints suggest themselves. First, it is likely that a surface patch in the scene will be continuous or cohesive and, if the facets are opaque, we should expect flow in the same
spatial region of the image to originate from the same surface patch. Secondly, in a sequence of time varying images we expect the image of a particular patch to be only slightly displaced from frame to frame, provided the frame rate is commensurate with the speed/distance ratio in the scene. It is clear that, within the stated restrictions, our expectations conform to spatial and temporal MRF's.

Whereas gradient-based schemes to obtain optic flow (e.g., [33], [34]) can yield optic flow at image positions defined to subpixel accuracy and therefore not on a regular grid, we have here assumed for simplicity that the flow is available on a regular square lattice throughout the image. (If the data were irregular and sparse, the MRF model could be used to interpolate and extrapolate the field; but we have not explored this further.)

One possible neighborhood system for the spatial MRF is shown in Fig. 1. It consists of the four nearest and four next nearest neighbors of site $i$. Given such a neighborhood system a set of "cliques" may be defined, where each clique is a set of sites which are mutual neighbors, and the total potential is given by a sum over individual clique potentials:

$$U(\omega) = \sum_c V_c(\omega). \quad (IV.1)$$

We have considered only pair interactions, so only cliques with two sites contribute to the potential. Hence, for the spatial MRF the total potential is

$$U = \sum_i \left( \sum_j V_{ij}(ij) \right). \quad (IV.2)$$

where $j$ runs over the neighbors of $i$, and $i$ runs over the entire set of sites. To generate piecewise constant interpretation regions we set

$$V_{ij}(ij) = \begin{cases} -a, & \text{if } X_i = X_j \\ +a, & \text{otherwise}, \end{cases} \quad (IV.3)$$

where $X_i$ denotes the interpretation at site $i$ and where $a_i$ is a positive number.

This potential favors spatial continuity of the interpretation. Geman and Geman introduced another process in their picture restoration algorithm, that of the line or discontinuity [23], which allows neighboring sites to have different interpretations at no cost other than that of introducing the line itself. In this way interpretation boundaries should appear more naturally. The discontinuities are situated on an interstitial lattice and are either active ("on") or inactive. A new neighborhood system incorporating the discontinuities is shown in Fig. 2(a). For simplicity, the extent of the neighborhood has been reduced to the four nearest neighbors.

The potential is modified to

$$U = \sum_i \left( \sum_j V_{ij}(ij) \right) + \sum_r V_r(L). \quad (IV.4)$$

where

$$V_{ij}(ij) = \begin{cases} 0, & \text{if } L_{ij} \text{ is ON} \\ V_{ij}(ij), & \text{otherwise}. \end{cases} \quad (IV.5)$$

The $V_r$ are the spatial line clique potentials, describing the cost of introducing the various line configurations, such as corners and tee-junctions. The line configurations are taken from Geman and Geman [23], but their associated potentials differ somewhat. They are shown in Fig. 2(b).

Finally, we consider how to include the temporal MRF constraint which provides an interpretation memory. The temporal neighborhood should be less extensive than the spatial neighborhood because it must not conflict either with expansion or contraction of the images of facets approaching or receding from the camera. Thus for the spatial neighborhood of Fig. 2, we are restricted to the spatio-temporal neighborhood shown in Fig. 3. The potential of equation (IV.4) is augmented by an extra term
choose an initial interpretation: for the first frame of a sequence this is random, but for later frames we can use the final annealed interpretation of the previous frame. Starting at a high temperature $T$, sites are visited in a random sequence and at each a random change in status proposed, say from facet $p$ to $q$. The local change in potential $\Delta U$ is computed from (IV.2), or (IV.2) with (IV.6), if the temporal site is included and the change in noise contributions from facets $p$ and $q$ derived using the planar facet algorithm:

$$\Delta(\|n\|^2) = \Delta(\|n\|^2_p) + \Delta(\|n\|^2_q).$$

(V.2)

The total change in cost is

$$\Delta C = \Delta U + \alpha \Delta(\|n\|^2)/2\sigma^2.$$  

(V.3)

If $\Delta C$ is negative or zero the proposed change is confirmed immediately, but if $\Delta C$ is positive it is accepted with probability equal to $\exp(-\Delta C/T)$. After one cycle of site visits the temperature is lowered according to the schedule proposed by Geman and Geman [23]:

$$T = \tau / \log(K + 1),$$

(V.4)

where $K$ is the iteration cycle. Geman and Geman determined a theoretical value of $\tau$ which guarantees convergence, but in practice found that much less conservative schedules gave good results.

The addition of line processes introduces another set of changes at each temperature. The line sites are visited in random order and a status change accepted or rejected according to the change in cost

$$\Delta C = \Delta U,$$

(V.5)

where $U$ is given by (IV.4). Note that a change in line status may change the site potentials and the line configuration potentials, but does not involve the data term.

In addition we have used the ergodicity theorem [23] to predict the line site status. A record of each line site status is kept over a number of cycles, and if the line has been "on" for more than 50 percent of the time it is turned on at the end of the period, and otherwise turned off.

VI. RESULTS FROM COMPUTER EXPERIMENTS

Fig. 4 shows the performance of the segmentation algorithm on a synthetic flow field of $16 \times 16$ data sites. The field was produced from the motion of three planar facets in the scene, each with different orientation and velocities. Fig. 4(a) shows the original configuration used to set up the flow field. The gray levels 1, 2, and 3 indicate which plane was visible from that point in the image, and so also indicate from which plane the flow at that site arose. Planes 1, 2, 3 are in the back-, middle-, and foreground, respectively, with plane 1 visible in two disjoint areas of the image. The initial guessed segmentation was random [Fig. 4(b)]. The segmentation produced by the spatial process without lines is shown in Fig. 4(c). The annealing was carried out over 1200 cycles with $\tau = 8.5$ and $a_i = 1$. The scale factor $\alpha$ was $8 \times 10^{-2}$, with 0.5 percent noise added to the flow vectors.
Fig. 4. (a) The original (unobserved) segmentation used to synthesize the flow fields. The gray levels correspond to image areas arising from different surface facets in the scene. (b) The initial random guess for the segmentation. (c) The segmentation achieved using the spatial process without lines (after 1200 iterations).

In most respects the results speak for themselves, but we have noticed one interesting effect in the annealing process. Its progress does not appear to be one of steady gradual improvement but rather typified by periods of slow progress followed by periods of more rapid reorganization. We believe this is caused by the nonlocal nature of the data term. As more sites are correctly assigned to a facet, the predicted values of flow from that facet improve. Thus, as the majority of sites become correctly assigned, the global configuration moves into the domain of attraction of a deep (though not necessarily global) minimum.

In Fig. 5(a) and (b) we show the results from annealing with line processes, using the configuration potentials of Fig. 2(b). The annealing was performed over 1200 cycles and the other parameters were unchanged. Again the results are convincing.

The performance of the full spatio-temporal annealing with lines is shown in Fig. 6. The true configuration [Fig. 6(a)] was obtained by allowing the facets to execute their motions in space for a short time and then recomputing their visible portions and the flow they contribute. The initial guessed segmentation was not random, but the final annealed segmentation from the previous frame [i.e., Fig. 5(a)] and the results of annealing are shown in Fig. 6(b) and (c). To prevent unnecessary reorganization the starting temperature was lowered to $\tau = 4.5$. The strength of the memory potential was $\alpha_{r} = 0.4$ and the other parameters were unchanged. The process converged after 600 cycles.

In Fig. 7 we show the segmentations achieved for surfaces which differ in orientation, but have the same motion with respect to the camera, and are at similar depths. One can imagine the camera being moved in a room, with the back and side walls and floor visible. The flow field and original configuration are shown in Fig. 7(a) and (b), respectively. The segmentation after spatial annealing of the first frame and spatio-temporal annealing of the second frame (2500 and 1000 cycles, respectively) is shown in Fig. 7(c). The cost function parameters used were the same as in the earlier experiments, although the starting
temperature of the spatial annealing was raised to $\tau = 10$. Because of the absence of different motions and depth discontinuities the segmentation is more difficult to achieve, with the edges running into each other. This may be contrasted with the results when the back wall is changed to, say, the back of the vehicle moving with respect to the side wall and floor as in Fig. 8. Here the boundaries of the moving portion are more cohesive, but again the interpretations of the side wall and floor, having the same motion, have run into each other.

The line configuration potentials used for the experiments described here are given in Fig. 2(b). Two more sets of values were tested, those of Geman and Geman [23] and Marroquin [37]. No catastrophic change in performance of the segmentation was found, although for our problem the former set stinted line formation somewhat and the latter set was too liberal. Our values lie between these sets, but are not otherwise optimized. The most difficult parameter to estimate was $\alpha$. Obviously a relatively large value of $\alpha$ requires the interpretation to fit the data, whereas a small value tends to produce a smooth interpretation. Empirically, we have observed that if at the start
of annealing the typical changes in the data term [the second term of (V.1)] are a few tens larger than the typical potential changes a good segmentation will be achieved. Too high a value of $\alpha$ in the presence of noisy flow data causes motting of the interpretation; there is some hint of this in Fig. 8.

**VII. CONCLUDING REMARKS**

The main result from this work is that the formulation of scene segmentation from visual motion as an optimization problem weakly constrained, or guided, by prior physical expectations is not only plausible but feasible.

In general terms, we have found that the approach to convergence is less predictable than that found by Geman and Geman in their picture restoration algorithm. There seems to be no invariable point by point improvement in the interpretation until a substantial fraction of points are correctly assigned, quite unlike a version of the picture restoration algorithm coded within the same optimization shell. This behavior appears to be a consequence of the nonlocal nature of the data or noise term: the assignment of one flow datum to a surface affects the contribution to the data term of all the other flow data assigned to that surface. In state space one can imagine very many high energy disordered states and a relatively small number of ordered states in which most of the flow is correctly interpreted. In these ordered regions the minimization problem appears easier, in that the global configuration lies inside a well of attraction of a deep cost minimum. The role of the MRF potentials is to encourage the state changes towards a configuration where the collective noise term can take over and drive the process to this deep near-global, or indeed global, minimum.

Notwithstanding the success of these simple experiments, there are some remaining difficulties with technique.

Perhaps the least important of these at present is the contrived geometry of the flow field. It is of course unrealistic to expect flow data at every site on a regular lattice, but, as we have suggested above, irregularly spaced data could be managed by incorporating some distance term into the potentials and the MRF could be used to interpolate the segmentation in areas where data are missing.

A more serious unsatisfactory aspect of the current implementation of the segmentation process is that the number of objects likely to be found has to be specified. One way to overcome this may be simply to overestimate and let objects disappear, but a more promising method may be to include a minimum entropy constraint in the cost function (e.g., [20]) which tries to reduce the number of different objects and different object attributes (that is, it prefers to coalesce two objects whose velocities, say, differ by only a small fraction into one object) but which would allow the number to increase if, say, the fit to the data were much better.

Without doubt, however, the most severe difficulty with the algorithm is computational inefficiency: the product of the many stochastic state changes the optimization demands and the nontrivial computation (i.e., the planar facet algorithm) at each proposed site change. One way of alleviating this burden might have been to make the state changes in parallel. This was suggested for the image restoration algorithm by Geman and Geman [23], and an approximation to their method was implemented recently on an SIMD machine, the ICL DAP, with a speedup of several hundred [38]. However, although the planar facet algorithm can itself be executed in parallel [31], there is a fundamental difficulty in making any site changes in parallel, because the computation of the noise or data term is not a local process. Another possibility to reduce the number of state changes for convergence would be of course to make a better than random initial "guess" for the segmentation. For this one could certainly use other image cues such as texture or color, but there is a further method of exploiting visual motion. One can detect the boundaries in the flow field between the moving planar patches locally (and so in parallel) by taking advantage of the ill-conditioning inherent in the planar-facet algorithm itself [39] and this segmentation could in turn be passed to the more global process for improvement.

A final possibility, yet to be addressed, might be to exploit in some way the particular form of the cost function to find a global minimum by a more deterministic process. Although this remains an attractive proposition, we believe that the nonlocal properties of the data term will make the problem difficult for any optimization scheme.

**APPENDIX**

In full, the quantities $c$ and $\beta$ are given by

$$
c_{11} = x_i \cos \theta_i, \quad \beta_1 = V_1 N_1 - V_3 N_3,
c_{12} = y_i \cos \theta_i, \quad \beta_2 = V_1 N_2 - \Omega_1,
c_{13} = -l \cos \theta_i, \quad \beta_3 = V_1 N_3 + \Omega_2,
c_{14} = x_i \sin \theta_i, \quad \beta_4 = V_2 N_1 + \Omega_1,
c_{15} = y_i \sin \theta_i, \quad \beta_5 = V_2 N_2 - V_3 N_3,
c_{16} = -l \sin \theta_i, \quad \beta_6 = V_2 N_3 - \Omega_1,
c_{17} = x_i (c_{11} + c_{12})/l, \quad \beta_7 = V_3 N_1 - \Omega_2,
c_{18} = y_i (c_{11} + c_{12})/l, \quad \beta_8 = V_4 N_2 + \Omega_1.
$$
with \( \cos \theta = \mathbf{v}_1 \cdot \mathbf{\hat{x}} / \| \mathbf{v}_1 \| \) and \( \sin \theta = \mathbf{v}_1 \cdot \mathbf{\hat{y}} / \| \mathbf{v}_1 \| \). The vectors \( \mathbf{\hat{x}} \) and \( \mathbf{\hat{y}} \) are unit vectors in the \( x \) and \( y \) directions.

**Acknowledgment**

The authors wish to thank the referees for useful criticism of an earlier version of this work.

**References**


David W. Murray graduated in physics from the University of Oxford, Oxford, England, in 1977 and continued there to receive the Doctorate in nuclear physics in 1980. From 1980 until 1982 he was a Research Fellow in Physics at the California Institute of Technology, Pasadena. Since joining the Long Range Research Laboratory at GEC Research in 1982, his research interests have centered on computer vision, especially the measurement and use of visual motion from early vision through to model-based recognition.

Bernard F. Buxton (M’84) studied theoretical physics at the Cavendish Laboratory, Cambridge, England, and the H. H. Wills Physics Laboratory at the University of Bristol, Bristol, England. He worked at both institutions before joining the Long Range Research Laboratory at GEC Research in 1981. His current main research interest is in the theory of computer vision, especially the development and demonstration of techniques and concepts required to enable machine vision systems to function in a wide variety of environments where 3-D perception is required.