Biol. Cybern. 62, 113–116 (1989)

Biological
Cybernetics

Structure-From-Motion by Tracking Occlusion Boundaries*

W. B. Thompson

Computer Science Department, 4-192 EE/CS Building, University of Minnesota, Minneapolis, MN 55455, USA

Abstract. Active visual tracking of points on occlusion boundaries can simplify certain computations involved in determining scene structure and dynamics based on visual motion. Tracking is particularly effective at surface boundaries where large, discontinuous changes in depth are occurring. Two such techniques are described here. The first provides a measure of ordinal depth by distinguishing between occluding and occluded surfaces at a surface boundary. The second can be used to determine the direction of observer motion through a scene.

1 Introduction

A flurry of recent activity has been directed at investigations of ways in which the interpretation of visual motion can be simplified in those perceptual systems which have the ability to visually track environmental surface points (e.g., Aloimonos 1987; Ballard 1987; Ballard and Ozcandarli 1988). Active tracking constrains eye/camera rotation in a particular way that aids in removing some of the complications normally associated with the computational analyses of the perception of scene structure and dynamics. This paper shows that these simplifications can be significant when the surface feature being tracked is an occlusion boundary involving large changes in depth. Two examples are presented. The first technique determines local depth orderings by recognizing which side of a boundary corresponds to an occluding surface. The second technique is able to estimate the direction of observer motion in a simpler manner than most other, previously proposed approaches.

The methods described below are most effective when the following three assumptions hold: *(1)* Occlusion boundaries involving significant changes in depth commonly occur. Both techniques *require* that depth discontinuities be present. This is almost always true for real scenes. It should be noted that it is commonly not true in the experimental displays used to probe biological perception of visual motion. *(2)* The observer is able to keep a selected edge element centered in the field of view. This is at least plausible in most natural situations where boundaries are not straight and/or surfaces are visually textured. Biological vision systems are in fact quite good at this task. Computer vision systems, on the other hand, have not yet demonstrated competence in tracking sufficient to effectively support the methods described below. Thus, actual implementation in engineered system of the techniques described in this paper is not yet possible. *(3)* Estimation of observer motion is done in situations in which at most a relatively small portion of the visual field corresponds to moving objects. Note that other techniques exist for recognizing the presence of moving objects (e.g., Thompson and Pong 1987).

2 Analysis

Visual motion depends on the instantaneous translational and rotational velocities of the eye/camera and the range to surface points in the scene. In our case, rotational velocities are constrained by the need to track a particular scene point. Analysis will be based on optical flow in the image near the tracked edge element. Note that in biological terms, this corresponds to retinal flow, not the Gibsonian idea of flow in the "optic array". Figure 1 illustrates the situation in the neighborhood of a boundary when no tracking is occurring. S_n corresponds to a near surface, which has associated optical flow f_n. S_n is occluding a more distant surface S_d, with associated flow f_d. The bound-

* A preliminary version of this material appeared in the Proceedings of the Workshop on Visual Motion, 1989. This work was supported by NSF Grant IRI-8722576

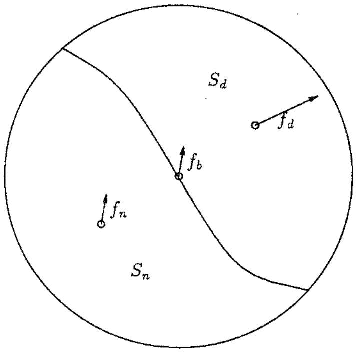

Fig. 1. Optical flow near a surface boundary

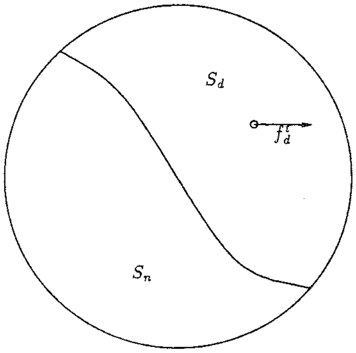

Fig. 2. Optical flow with edge tracking

ary itself moves in the image with flow f_b. The possible flow values in Fig. 1 are related by the *boundary flow constraint* (Thompson et al. 1985):

The image of an occluding contour moves with image of the occluding surface immediately adjacent to the contour.

Thus, over a small neighborhood, $f_b = f_n$. Figure 2 describes the situation when the edge is being accurately tracked. Tracking is effected by introducing an eye/camera rotation of velocity $\omega = (A, B, 0)^T$ which exactly compensates for f_b.[1] Because of the boundary flow constraint, this also has the effect of nulling out f_n. The only visible flow left, $f_d^t = f_d - f_b$, is associated with the more distant surface.

A simple set of equations defines the relationship between optical flow, motion, and scene structure (Horn 1986). Using a planar imaging system with image position normalized by focal length, perspective projection, and a coordinate system with origin at the optical center and z axis along the line of sight:

$$u = u_t + u_r, \qquad v = v_t + v_r, \tag{1}$$

where u and v are the x and y components of flow, z is the distance to the surface point imaged at (x, y), translational velocity is $\mathbf{T} = (U, V, W)^T$, and

$$u_t = \frac{-U + xW}{z}, \qquad v_t = \frac{-V + yW}{z}, \tag{2}$$

$$u_r = Axy - B(x^2 + 1), \qquad v_r = A(y^2 + 1) - Bxy. \tag{3}$$

The optical flow equations simplify considerably at the center of the field of view:

$$\lim_{x,y \to 0} u_t = \frac{-U}{z}, \qquad \lim_{x,y \to 0} v_t = \frac{-V}{z}, \tag{4}$$

$$\lim_{x,y \to 0} u_r = -B, \qquad \lim_{x,y \to 0} v_r = A. \tag{5}$$

If the tracked boundary element is centered within the field of view and if surface flow is measured sufficiently near this center, then f_b, f_n, and f_d are all determined by (4) and (5).

Utilizing the fact that $z_n < z_d$, we can now compute f_d^t.

$$f_d^t = f_d - f_b = f_d - f_n \tag{6}$$

$$= \left(\frac{-U}{z_d} - B - \frac{-U}{z_n} + B, \right.$$
$$\left. \frac{-V}{z_d} + A - \frac{-V}{z_n} - A \right) \tag{7}$$

$$= \left(\left(\frac{1}{z_n} - \frac{1}{z_d} \right) U, \left(\frac{1}{z_n} - \frac{1}{z_d} \right) V \right) \tag{8}$$

$$= (aU, aV), \quad a = \left(\frac{1}{z_n} - \frac{1}{z_d} \right) > 0. \tag{9}$$

It is easily shown that the image plane location indicating the line of sight corresponding to the direction of translational motion is:

$$(x, y)_{\text{trans}} = \left(\frac{U}{W}, \frac{V}{W} \right). \tag{10}$$

While we don't know the values of z_n or z_d, (9) and (10) show that f_d^t is a scaled version of the projection of the translation vector onto the image plane. That is, f_d^t points towards the image plane location of the direction of motion.[2] (Two notes are in order: First of all, (10) is still meaningful for purely lateral motion when $E = \infty$. In such cases, the "line of sight" corresponding to the direction of motion is parallel with the image plane and thus the corresponding image plane location is at infinity. Secondly, if there is a backwards component of motion, $W < 0$ and (10) give the direction from which the observer is coming. f_d^t still points in the projected direction of observer motion.)

We can now summarize the two algorithms for analyzing visual motion using edge tracking:

[1] We have simplified the presentation by not allowing for "spin" rotations around the z axis

[2] This location is commonly called the "focus of expansion", but the term is only strictly correct for purely translational motion

Identification of Occluding Surface

When a boundary element is visually tracked, the region to the side of the boundary corresponding to the occluding surface will have near-zero image flow. The region to the side of the boundary corresponding to the occluded surface will in general be associated with significant visual motion.

Determination of Direction of Observer Motion

When a boundary element is visually tracked, optical flow due to the more distant surface indicates the direction of observer motion. The flow vectors point in the direction of the image location corresponding to the line of sight coincident with the direction of translational motion. Multiple fixations over the field of view can be used to solve for the actual direction of translation.

3 Discussion

Both algorithms offer significant computational simplifications over alternate approaches. The few previously reported optical flow based techniques for differentiating between occluding and occluded surfaces require reasonably accurate flow estimates on either side of the boundary (Thompson et al. 1985; Clocksin 1980). The method reported here only requires that regions of significant image motion be differentiated from regions with little or no motion. It is far easier to determine whether or not image motion is occurring than it is to estimate the specific characteristics of that motion. When eye/camera rotations are possible, the determination of observer motion is difficult because of the complex manner in which translational and rotational motion interact to generate an optical flow field (see Horn 1986). Edge tracking of occlusion boundaries eliminates the complexity associated with rotation.

The effectiveness of these two algorithms is limited by the accuracy with which boundaries can be tracked and by the possible absence of visual texture adjacent to the boundaries. While biological systems are capable of tracking environmental points with relatively high precision, the computer vision community has only recently begun to study the engineering difficulties involved in tracking features in complex scenes. Aperture effects are a further consideration. It is generally felt that only the component of motion perpendicular to an edge can be determined. In fact, this ambiguity is usually resolvable due to the curvature and end points of contours (e.g., see Hildreth 1983). Reasonably reliable two-dimensional tracking should be possible for most realistic scenes, though sufficient experimentation has not yet been done. Both algorithms depend on recognizing aspects of image motion in the neighborhood of the tracked edge. This is most easily accomplished if surfaces on either side of the boundary are visually textured. This will hold in many but not all scenes. We do know that human vision is capable of "filling in" the motion of homogeneous portions of surfaces. However, we do not as yet have good computational models of how this is done.

Open questions remain as to whether or not biological vision systems actually use methods of this sort to simplify the determination of scene structure and motion trajectories. To answer these questions, we need to know more about fixation patterns in realistic dynamic environments and about how fixation and eye tracking affect the perception of relative depth.

Finally, it is important to examine carefully the notion that tracking is actually "simplifying" the problem. For example, Ballard and Ozcandarli (1988) argue

"Programmed eye movements [tracking] reduce the degrees of freedom in a given computation and thus lead to simple solutions."

Tracking does not in fact reduce the *conceptual* difficulties associated with interpreting visual motion. Eye tracking provides neither additional constraints nor other sorts of new information. This is easily seen by recognizing that all of the information in the tracking image is available in an image of the same scene without tracking. Tracking is accomplished by generating a rotation of the eye/camera system based on estimates of image drift such as optical flow at the image center. Once this rotational velocity is determined, a non-tracking image sequence can trivially be converted into the equivalent tracking sequence using (3). This conversion does not require that tracking actually take place and does not require any additional information. (For example, no knowledge of scene structure is required.) Since the "tracked" image flow can be predicted from the "untracked" flow, there can be no additional information gained by actually doing the tracking. This is illustrated in Figs. 3–5. Figure 3 is a (simulated) flow field resulting from motion through a simple scene with surfaces at three distinct depths. We can exactly predict the rotation needed to null the flow in Fig. 3 at the image center. The flow corresponding to this rotation is shown in Fig. 4. Adding together the flow fields in Figs. 3 and 4 we get the field in Fig. 5, which is exactly what would have arisen if tracking had actually occurred.

Phrased in terms of Marr's description of information processing tasks (1982), tracking does not simplify the computational theory of structure-from-motion problems, but it can simplify the algorithms and implementation. In fact, both of the algorithms described above are really special cases of methods

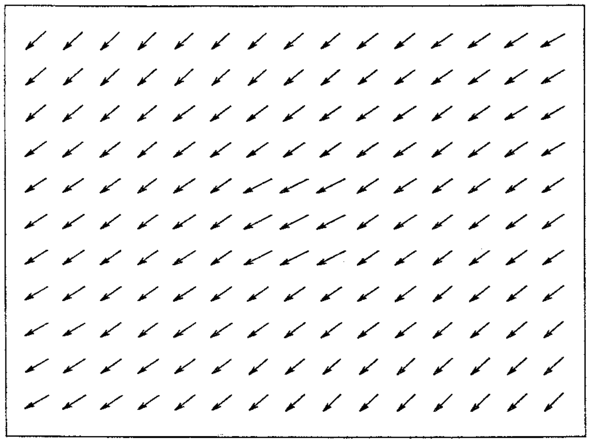

Fig. 3. An "arbitrary" flow pattern with no tracking

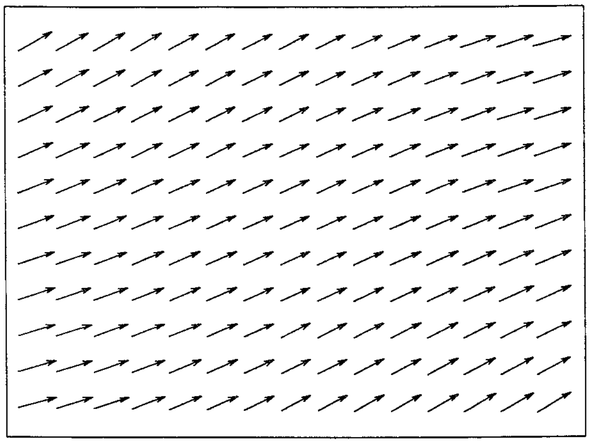

Fig. 4. Compensating rotational flow field

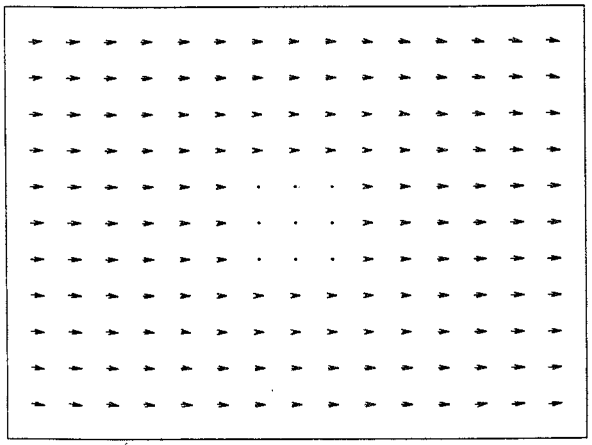

Fig. 5. Simulation of tracking for Fig. 3

already presented in the literature. Occlusion analysis is described in (Thompson et al. 1985). The method for determining direction of motion is essentially equivalent to that described in (Reiger and Lawton 1983). What is different are the simplifications in actual algorithms, not the underlying computational theory. Even these simplifications, however, occur only because hard parts of the problem are transferred elsewhere. The interpretation of flow is easier, but a difficult tracking problem must also be solved. What we actually end up with is an example of coarse-grain parallelism. Part of the computational effort is in the tracking system, part in the simultaneously executing flow interpretation component. Since tracking must be done anyway, this represents an efficient decomposition of the problem.

Acknowledgement. Dana Ballard first got me thinking about the role of tracking in visual motion analysis.

References

Aloimonos J, Weiss I, Bandyopadhyay A (1987) Active vision. Proc First Int Conf Comput Vision: 35–54

Ballard DH (1987) Eye movements and spatial cognition. Technical Report 218, University of Rochester

Ballard DH, Ozcandarli A (1988) Eye fixation and early vision: kinetic depth. Proc Second Int Conf Comput Vision: 524–531

Clocksin WE (1980) Perception of surface slant and edge labels from optical flow: a computational approach. Perception 9:253–269

Hildreth EC (1983) The measurement of visual motion. MIT Press, Cambridge, Mass

Horn BKP (1986) Robot vision, MIT Press, Cambridge, Mass

Marr DA (1982) Vision. Freeman, San Francisco

Reiger JH, Lawton DT (1983) Sensor motion and relative depth from difference fields of optic flows. Proc Eight Int Joint Conf Artif Intell: 1027–1031

Thompson WB, Pong TC (1987) Detecting moving objects. Proc First Int Conf Comput Vision: 201–208

Thompson WB, Mutch KM, Berzins VA (1985) Dynamic occlusion analysis in optical flow fields. IEEE Trans PAMI-7:374–383

Received: December 30, 1988
Accepted: June 27, 1989

Dr. William B. Thompson
Computer Science Department
4-192 EE/CS Building
University of Minnesota
Minneapolis, MN 55455
USA