Qualitative Detection of Motion by a Moving Observer

Randal C. Nelson

Technical Report 341
April 1990

UNIVERSITY OF
ROCHESTER
COMPUTER SCIENCE

Qualitative Detection of Motion by a Moving Observer

Randal C. Nelson
nelson@cs.rochester.edu
Department of Computer Science
University of Rochester
Rochester, NY 14627 USA

Abstract

Two complementary methods for the detection of moving objects by a moving observer are described. The first is based on the fact that, in a rigid environment, the projected velocity at any point in the image is constrained to lie on a 1-D locus in velocity space whose parameters depend only on the observer motion. If the observer motion is known, an independently moving object can, in principle, be detected because its projected velocity is unlikely to fall on this locus. We show how this principle can be adapted to use partial information about the motion field and observer motion that can be rapidly computed from real image sequences. The second method utilizes the fact that the apparent motion of a fixed point due to smooth observer motion changes slowly, while the apparent motion of many moving objects such as animals or maneuvering vehicles may change rapidly. The motion field at a given time can thus be used to place constraints on the future motion field which, if violated, indicate the presence of an autonomously maneuvering object. In both cases, the qualitative nature of the constraints allows the methods to be used with the inexact motion information typically available from real image sequences. Implementations of the methods that run in real time on a parallel pipelined image processing system are described.

Key Words and Phrases: motion analysis, movement detection, qualitative vision, real-time vision.

This research was supported by the U.S. Army Engineering Topographic Laboratories under Grant DACA76-85-C-0001 and by the National Science Foundation under Grant CCR-8320136.

Qualitative Detection of Motion by a Moving Observer

Abstract:

Two complementary methods for the detection of moving objects by a moving observer are described. The first is based on the fact that, in a rigid environment, the projected velocity at any point in the image is constrained to lie on a 1-D locus in velocity space whose parameters depend only on the observer motion. If the observer motion is known, an independently moving object can, in principle, be detected because its projected velocity is unlikely to fall on this locus. We show how this principle can be adapted to use partial information about the motion field and observer motion that can be rapidly computed from real image sequences. The second method utilizes the fact that the apparent motion of a fixed point due to smooth observer motion changes slowly, while the apparent motion of many moving objects such as animals or maneuvering vehicles may change rapidly. The motion field at a given time can thus be used to place constraints on the future motion field which, if violated, indicate the presence of an autonomously maneuvering object. In both cases, the qualitative nature of the constraints allows the methods to be used with the inexact motion information typically available from real image sequences. Implementations of the methods that run in real time on a parallel pipelined image processing system are described.

Key words and phrases: Motion Analysis, Movement Detection, Qualitative Vision, Real-time Vision.

I. Introduction

The ability to rapidly detect moving objects seems to be almost universal in animals with eyes. An obvious reason is that some of the most pressing issues in the survival of an organism involve objects that move (e.g. predators, prey, and falling rocks). Robotic systems that interact with real-world environments face similar issues. They too are frequently critically concerned with objects that move. For example, an autonomous vehicle must avoid hitting people or animals that wander into its path; a surveillance system must identify intruders; and a smart weapon may pursue a moving target. A reasonable heuristic for interaction with the real world is if it is moving, you should probably pay attention. A method of detecting independent motion is thus valuable as a method for directing more sophisticated (and costly) processing to areas where it can be most effectively utilized.

For a stationary observer, a simple approach is to difference images obtained a short time apart, and mark the non-zero regions of the resulting image. This strategy has formed the basis for a number of movement alerting systems (e.g. [Ande85, Dins88]). Unfortunately, a system cannot always keep still. For a moving observer, the problem is much harder, since everything in the image may be undergoing apparent motion; and the overall pattern of motion may be quite complex. In principle, independently moving objects can be identified by a general-purpose shape-from-motion analysis. However, an entirely satisfactory solution to the structure from motion problem has not yet been demonstrated, even under the assumption of global rigidity. The main problem is that entirely different observer motions can produce very similar apparent motion fields, especially for narrow fields of view. Consequently, the accuracy of the solutions tends to be very sensitive to the accuracy of the underlying motion data [Tsai84] (though see [Nels89]), often requiring a precision which is difficult to attain in practice. The presence of multiple moving objects in a scene further complicates the picture since the rigid world constraints, which constitute the usual device for combining data from different portions of the image, no longer completely valid. The above notwithstanding, several authors have proposed methods of embedding moving object detection in general purpose motion analysis systems [Heeg88, Burt89]. These methods, however, tend to be computationally expensive, and suffer from the same limitations as techniques which assume a rigid world. A few studies have concentrated specifically on the detection of moving objects by a moving observer. Thompson [Thom90] describes some basic principles that can be used to discriminate moving objects when various aspects of the observer motion are known, but leaves open some questions about how these principles might be applied in practice. In our discussion of the constraint ray filter, we show how one of these principles can be adapted to allow the use of imprecise and partial motion information. Bhanu et al. [Bhan89] propose a method of detecting moving targets based on the identification of a fuzzy focus of expansion and a qualitative analysis of the motion of scene points. This also has some aspects in common with our proposals but it utilizes motion information derived from point correspondences, and invokes a rule-based system of qualitative reasoning, making it considerably higher level (and more expensive) than the methods described here.

This paper argues that the movement detection problem can be solved (at least in most cases) without having to solve the entire structure-from-motion problem. In particular, we present two complementary qualitative measures that can be used to tag motion that is inconsistent with an interpretation of global rigidity. The first method, which we term *constraint ray filtering*, makes use of knowledge about the observer's motion. It is based on the fact that, in a rigid environment, the projected 3-D velocity at any point in the image is constrained to lie on a 1-D locus in velocity space whose parameters depend only on the observer motion. Thus in principle, if the motion field and observer motion are known, an independently moving object can be detected because its projected velocity is unlikely to fall on this locus. In practice, quantitative estimates of the motion field and observer motion are both difficult and computationally expensive to obtain. We show how the basic principle can be adapted to use partial information about

the motion field and observer motion that can be rapidly computed from real image sequences. The second method makes use of qualitative knowledge about the motion of the object to be detected. It takes advantage of the fact that the apparent motion of a fixed point due to smooth observer movement changes slowly while the apparent motion of moving objects such as animals or maneuvering vehicles often changes rapidly. We term such movement *animate motion*. Such motion can be detected by using the motion field at a given time to constrain the future motion field under smooth continuation, and then looking for violations of these constraints.

The techniques presented here reflect our conviction that vision in general and motion in particular is better suited for recognition than reconstruction. This position is best clarified by examining how the two paradigms are distinguished. A major distinction is that of specificity. Reconstruction can be viewed as a general transformation of information in one form into another (presumably more useful) modality (e.g., time varying imagery into depth maps). Recognition, on the other hand, serves to identify a specific situation of interest to the system, for instance, the approach of a fly if you are a frog, or a bird if you are a fly. A reconstructed world model contains a lot of information, possibly enough to find a fly if you are a frog, but it also contains a lot of information that a frog has no interest in, and that was expensive to obtain. More specifically, a characteristic of reconstructive vision is that information is transformed without regard for its intended use, following a policy of least commitment. The usual justification is that since you never know what information you will need, you should preserve as much as possible. The disadvantage is that, since most of the information is never needed, such a policy can result in a huge amount of wasted effort, especially if attempted at higher levels. We advocate instead, what might be termed a policy of most commitment; that is, compute only what is necessary to solve the problem of interest. It might be argued that such a policy is poor science because it will never produce generalizable systems. On the contrary, we believe that the world is so structured that what is useful for one purpose will, perhaps in slightly modified form, prove useful for another. Such a statement is, of course, impossible to prove; we can only point at the history of science which is rife with examples of one structure being built on another, or at evolution, which also seems to operate in this manner.

A second distinction is the one between qualitative and quantitative methods. Reconstruction is, in essence, a quantitative procedure, and consequently dependent for its success on the numerical accuracy of the algorithms employed. This has been a problem in shape-from-x analyses in general. Recognition, on the other hand, can make use of qualitative distinctions (moving up, moving down, rotating, expanding) and relative relationships (faster, slower, in front, behind), which can be computed from much less exact information. What we feel has been overlooked is a wide variety of applications in which robustly computable motion information can be used for identification directly, and much more efficiently, than than via traditional 3-D reconstruction. The movement detection techniques presented here are one example. For others and for further discussion see [Nels88a, Nels88b, Nels89].

2. Background and Notation

2.1 Structure-from-motion

The techniques described here have roots in the research that has been done in the context of the structure-from-motion problem, and in methods developed to obtain local motion information from image sequences. Although we propose a somewhat different use of the available information, this work has motivated and provided foundations for our approach, and it is thus appropriate to review the field.

A camera moving within a three dimensional environment produces a time-varying image that can be characterized at any time t by a two dimensional vector-valued function f known as the *motion field*. The motion field describes the two dimensional projection of the three

dimensional motion of scene points relative to the camera. Mathematically, the motion field is defined as follows. For any scene point (x,y) in the image, there corresponds at time t a three dimensional scene point (x',y',z') whose projection it is. At time $t+\Delta t$, the world point (x',y',z') projects to the image point $(x+\Delta x,y+\Delta y)$. The motion field at (x,y) at time t is given by

$$f(x,y,t) = \lim_{\Delta t \to 0} [\frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t}].$$

The motion field depends on the motion of the camera, the three dimensional structure of the environment, and the three dimensional motion (if any) of objects in the environment. If all these components are known, then it is relatively straightforward to calculate the motion field. In the traditional approach to motion analysis, the question has been whether the process can be inverted to obtain information about camera motion and structure of the environment. This is the basis of the structure-from-motion problem. The solution is not easy, and if arbitrary shapes and motions are permitted in the environment, there may not be a unique solution. However, it can be mathematically demonstrated that, in many situations, a unique solution exists.

The existence of such solutions has inspired a large body of work on the mathematical theory of extracting shape and/or motion information from the motion field. There have been two basic approaches to the problem. The first utilizes point correspondences in one or more images, generally under the assumption of environmental rigidity [Ullm79, Tsai81]. This is equivalent to knowing the motion field at isolated points of the image. Several authors have obtained closed form solutions to the shape from motion problem in this formulation, obtaining a set of linearized equations [Long81, Tsai84]. The second approach uses information about the motion field and its derivatives in a local neighborhood under some assumption about the structure of environmental surfaces (e.g., they are planar) [Praz81, Boll87, Waxm87]. In this case, the end result is a set of equations relating the motion field derivatives to the camera motion and the three-dimensional structure of the environment. Most of these studies, however, have started with the assumption that detailed and accurate information, either in the form of point correspondences or dense motion fields, is available. Unfortunately, the solutions to the equations are frequently inordinately sensitive to small errors in the motion field. In the case of point correspondences, Tsai and Huang [Tsai84] report 60% error for a 1% perturbation in input for some instances using their method. This error sensitivity is due both to inherent ambiguities in the motion fields produced by certain camera motions, at least over restricted fields of view, and (in the second approach) to the reliance on differentiation which amplifies the effect of any error present in the data.

The two approaches to obtaining shape from motion utilize somewhat different methods for extracting motion information from image sequences. The methods using point correspondences rely on matching techniques similar to those employed in stereo vision [Moro79, Marr79, Barn80]. This process is well known to be difficult since features may change from one image to the next, and even appear and disappear completely.

Techniques for computing dense motion fields have relied heavily on differential methods, which attempt to determine the motion field from local computations of the spatial and temporal derivatives of the gray scale image. The first derivative methods originally proposed by [Horn81] must deal with what is known as the *aperture problem*, which refers to the fact that only the component of image translation parallel to the gradient can be recovered locally from first order differential information. Intuitively, the aperture problem corresponds to the fact that for a moving edge, only the component of motion perpendicular to the edge can be determined. This effect is responsible for the illusion of upward motion produced by the rotating spirals of a barber pole where either vertical or horizontal motion could produce the local motion of the edges, and the eye chooses the wrong one. In order to obtain an approximation of the full motion field vector, information must be combined over regions large enough to encompass significant variations in the gradient direction. The most common method of doing this involves some form of

regularization [Horn81, Anan85, Nage86]; however such methods often result in blurring of motion discontinuities. A non-blurring method known as *constraint line clustering* has been proposed by Schunck [Schu84]. Techniques using higher order derivatives to avoid the aperture problem have been proposed [Nage83, Uras88]; however these suffer from stability problems due to multiple differentiation and typically require extensive smoothing to produce clean results. Other methods include spatio-temporal energy methods [Heeg87], Fourier methods based on phase correlation [Burt89], and direct correlation of image patches [Barn80, Litt88]. Recent work by Anandan [Anan89] provides a common framework into which many of these methods can be incorporated.

A potential problem with most of the above approaches is the assumption that the motion field manifests itself locally as a rigid 2-D motion of an image patch. Unfortunately, the local apparent motion of the image, known as the *optical flow*, does not necessarily correspond to the 2-D motion field. The most obvious demonstrations are pathological examples. For instance, a spinning, featureless sphere under constant illumination has zero optical flow, but a non-zero motion field. Conversely, a stationary sphere under changing illumination has non-zero optical flow, but zero motion field. Image patches also undergo various non-rigid deformations such as expansion and skewing. Verri and Poggio [Verr87] have shown that only under special conditions of lighting and movement do the motion field and the optical flow correspond exactly. They also show, however, that for sufficiently high gradient magnitude, the agreement can be made arbitrarily close. This corresponds to the intuition that for strongly textured images the motion field and the optical flow are approximately equal. A few authors have attempted to explicitly include some of these effects (e.g. [Burt89]), but it is not clear that any great advantage has been obtained thereby.

On the whole, despite a great deal of effort expended in devising motion invariants, regularization methods, and matching techniques, neither correspondence nor dense field methods have yielded data sufficiently accurate to allow the theoretical structure-from-motion results to be reliably applied. Adiv [Adiv85] argues that inherent near ambiguities in the 3-D structure-from-motion problem may make unfeasible the extraction of information sufficiently precise to allow uniform application of the theoretical solutions. Verri and Poggio [Verr87] make essentially the same point, arguing that the disagreement between the motion field and the optical flow makes the computation of sufficiently accurate quantitative values impractical.

An alternative is to devise qualitative applications that can make use of inaccurate motion field information [Thom86, Nels88a, Nels89]. The movement detection strategies described here represent one such application. Specifically, they utilize motion features derived from qualitative descriptions of optical flow direction and magnitude rather than accurate measurements of the motion field. Thus they can directly utilize partial information such as the approximation of the gradient parallel component computed in the first step of the Horn and Schunck procedure.

2.2 Notation: spherical images and the local frame.

We consider the image formed by spherical projection of the environment onto a sphere of radius ρ termed the *image sphere*. The use of spherical projection makes all points in the image geometrically equivalent with respect to the observer, which considerably simplifies some of the analyses. In particular, we can define local coordinate systems with respect to an arbitrary image point p. The locations of points in the environment are expressed in terms of coordinates (X,Y,Z) where $(0,0,0)$ coincides with the center of projection, and the positive Z axis passes through p. Image positions in the neighborhood of p are expressed in terms of coordinates (x,y) where $(0,0)$ coincides with p and the x and y axes with the local projection of the X and Y axes respectively. This is permissible because the image sphere is locally Euclidean. The Euclidean neighborhood of p will be referred to as the *local projective plane* (Figure 1). Since all points in the image are geometrically equivalent under spherical projection, we can notationally simplify much of our

local analysis by carrying it out in terms of these local coordinate systems.

Ordinary cameras do not utilize spherical projection, but if the field of view is not too wide, the approximation is reasonably close. Since the distortion is purely geometric in origin, it could be corrected should it prove to be a problem for any particular camera. In experiments we performed using a camera with a field of view of approximately 20×30 degrees, no correction was necessary in order to obtain good results.

3. Movement Detection via Constraint Ray Filtering

3.1 Theoretical basis

The first method of detecting an independently moving object is based on the observation that the projected motion at any point on the image sphere is constrained to lie on a half line (ray) in local velocity space whose parameters depend only on the observer motion and the location of the image point. In other words, despite the fact that objects at different depths typically display different apparent motion, the possibilities are constrained to a one dimensional locus in a two dimensional space. On the other hand, the projected motion for an independently moving object is unconstrained and is unlikely to fall on this locus. Thus testing the motion field to determine whether it is consistent with the local constraint ray provides a means of detecting non-rigid motion. The basic nature of the constraint is fairly well known, and it has been discussed as a means of movement detection [Thom90]; however its adaptation to partial and inexact motion information for use in a fast, practical system does not appear to have been much developed.

As a simple motivating example, consider the case of an observer translating to the right while looking straight ahead. The apparent motion of imaged objects rigidly attached to the world is horizontal and to the left with magnitude inversely proportional to the distance to the projecting world point. Any point of the motion field that contains a vertical component must thus arise from an independently moving object. Constraint ray filtering is a generalization of this idea.

To see how the constraint ray arises consider the local projective plane centered at point p on the image sphere. The projected velocity at point p is expressed by the vector (u,v) where u and v are the apparent (angular) velocities parallel to the local x and y axes respectively. The rotational motion of the observer can be decomposed into components ω_X , ω_Y, and ω_Z parallel to the local X, Y, and Z axes. The projected velocity due to this rotation is given by $V_\omega=(\omega_x,\omega_y)$ independent of the distance to the world point projecting to p. Similarly, the translational motion of the observer can be decomposed into components v_x, v_y, and v_z, again parallel to the local axes. In this case, the projected velocity is given by $V_t=(v_x/Z,v_y/Z)$ where Z is the distance from the origin to the world point that projects to p. The net projected velocity is the sum of the two pieces is

$$V = V_\omega + V_t = V_\omega + \frac{1}{Z} V_{XY}$$

where V_{XY} is (v_X,v_Y). For a given observer motion, both V_ω and V_{XY} are uniquely determined, and $1/Z$ runs from 0 to $+\infty$. The possible values for V thus lie on a ray in velocity space parallel to V_{XY} and with endpoint V_ω (Figure 2).

3.2 Practical considerations

As noted in Section 2, many methods for approximating the motion field utilize low-level computations that provide only the component of the field parallel to the local image gradient. These components are then combined by various methods to obtain an approximation to the complete motion field. Since such methods are often computationally expensive it is worthwhile to examine the constraints that can be placed on the gradient parallel component and to consider

whether it alone might provide information which could be used to identify independently moving objects. Consider a point x in velocity space representing the value of the motion field somewhere in the image. The gradient parallel component of the motion field can be represented by the vector describing the projection of x onto the line that is parallel to the gradient and that passes through the origin. Recall from elementary geometry that chords drawn from diametric points on a circle to a third point on the circle meet at a right angle. Thus all points on the circle whose diameter is the line segment $\overline{ox}$ represent possible gradient parallel projections of x (Figure 3). Conversely, all lines passing through the origin intersect the circle at a point representing the projection of x onto them. Hence this circle represents all the possible gradient parallel components consistent with the motion vector. Suppose that the 1/Z lies between 0 and α. Then the possible image motions lie on the line segment with endpoints V_∞ and $V_\infty+\alpha V_{XY}$. Each point on the line segment generates a circle as described above. The union of all these circles thus represents a constraint on the gradient parallel component of the motion field. It is easily seen that all these circles pass through both the origin and the projection of the origin on the constraint ray (or its extension). Thus the constraint region can be determined from the circles generated by the segment endpoints. In particular, the constraint region is the union of the two solid circles less their intersection (i.e. their exclusive OR). Figure 4 shows the partitions for several situations. In the limiting case of $\alpha=\infty$ ($Z=0$) the partitions are formed by the intersection of a circle and a half space.

The fraction of the plane representing gradient parallel components consistent with a rigid environment is frequently sufficiently small that an independently moving object has a good chance of generating gradient parallel components that fall outside of this region. This is particularly true if α can be bounded (e.g by knowing that the observer is at least a certain distance from the nearest object). Thus a movement detector can be can be constructed that utilizes local results of a differential motion computation. Such a detector would exhibit more false negatives than one utilizing the complete motion field, but on the other hand, the approximation of the gradient parallel component is far less computationally expensive.

The next issue is determining the motion of the observer. It was assumed in the above analysis that this motion was known. In some situations, such information might be available from external sources, for instance, from inertial sensors or from explicit knowledge of the observer motion (stationary robots). For many applications though, it is desirable to have a self-contained system that does not rely on outside sources of information. Unfortunately, determining the observer motion from an image sequence is, in the general case, tantamount to solving the structure from motion problem for static scenes, which can be hard to do. Assuming such information to be available might thus seem to be begging the question concerning the hardest part of movement detection.

It turns out, however, that the technique can be used in practical cases without the ability to determine observer motion exactly. Two facts make this possible. First, recall that a major difficulty of solving the ego-motion from motion problem arises from the fact that, in certain situations, the motion field arising from rotation and translation can be very similar, while the implications of each about the 3-D structure of the world are very different. In our case, however, the constraints arising from such similar fields are similar. Thus, unlike the case for structure from motion, it does not much matter if the two sources are confused. Second, for many practical problems, the system spends most of its time executing only a few types of motion. For instance, in the case of camera mounted in a car and stabilized against high frequency rotational jitter (much as the eyes of most mammals are stabilized by the reflexive VOR system), the motion is primarily straight ahead with slow rotations about the axes perpendicular to the forward motion as the car goes around corners and up and down hills. This suggests that the necessary information about observer motion can be obtained by matching against a relatively small set of prototype motion fields. The following are examples of canonical motion fields that would be useful

when utilizing images having angular extent small enough so that the distortion produced in projecting the spherical image onto a plane is relatively small. (i.e. < about 40 x 40 degrees).

1. Field that is all approximately in the same direction: This corresponds either to rotation about an axis approximately perpendicular to the direction of the gaze, translation roughly perpendicular to the direction of gaze, or a combination of rotation with a translation such that the directions of the flows align. Such motions frequently arise in systems navigating on approximately flat terrain. The constraints effectively exclude motion with a component parallel to the dominant direction but of opposite sign, or with a significant perpendicular component.
2. Field having a focus of expansion in center of image: This corresponds to translation in the direction of the gaze. Here the constraints exclude motion towards the origin or having a significant tangential component.
3. Field with a distinct focus of expansion anywhere in the image (2 above is a special case). This corresponds to pure translation, or to movement while fixating on a distant point.
4. Field having an expanding periphery with uniform components perpendicular to lines passing through the image origin. This corresponds to straight ahead motion with slow rotation about a perpendicular axis. The rotation can be obtained from the motion field components normal to a perpendicular pair of lines through the origin (e.g. the local x and y axes in the image) and used to set the constraints.
5. Field that is all in one of two directions 180 degrees opposed. This results from fixating an object in the scene while moving in a direction roughly perpendicular (e.g. +- 20 degrees) to the direction of gaze. Fixation on a point at infinity or the nearest point in the image will produce a field fitting the criteria for case 1. The constraints exclude motion with a significant perpendicular component.

These cases all have robust signatures that allow them to be identified from relatively sparse information using simple pattern classification techniques (e.g. nearest-neighbor methods) They also span a wide range of motions, covering many of the situations that occur in practice in moving systems.

There are a few situations that will cause trouble. The most common arises when an isolated nearby object appears in front of a distant background. Without external information about either the observer's movement or the distance to the object, there is no way to determine whether the object is stationary or undergoing uniform motion. This would be a problem with any movement detection system. Another, which might be called the moving moon illusion, results from an isolated distant object and a strong, flat foreground. In this situation, it is possible to fixate the foreground and interpret it as distant, whereupon the distant "moon" appears to move.

3.3 Implementation

We have implemented a movement detector based on the above principles that operates in real time (~.1s latency), and robustly detects independent motion in a wide variety of situations. The first step is the computation of local motion information. We use a differential method similar to the first step of the Horn and Schunck algorithm [Horn81], dividing the temporal derivative by the gradient magnitude to obtain an estimate of the gradient parallel component of the motion field. This operation is performed on a 512 x 512 video signal at 30 hertz using a collection of Datacube Maxvideo image processing boards, and provides usable values for angular velocities between about 20 and 200 pixels per second (see Appendix A). This array is then subsampled to 64 x 64 and downloaded to a Sun 360. A Hough transform technique is used to rapidly compute a coarse representation (here a 4 x 4 array quantized to one of 8 directions) of the true motion

field from the gradient parallel values. Figure 5 shows the coarse field produced by a combination of rotation and translation. This is normalized to form a feature vector which is then compared against a stored library of canonical motion fields in order to determine which of the known types of motion the observer is making. Currently the system recognizes motions in classes 1 and 2 described above. The canonical field is used to generate a filter image which specifies, for every point in the image, the range of gradient parallel components consistent with the presumed motion. The filter image is then compared with the measured estimates of the gradient parallel component, and regions exhibiting inconsistent motion are marked as potentially containing independently moving objects. By using bit encodings to coarsely represent the motion field, an update rate of about 10 Hertz was achieved.

The system has been tested using the Rochester Robot (Figure 6), which consists of a two-eyed (we used only one), three degree of freedom head attached to a six degree of freedom robot arm, to provide observer motion [Brown88a]. For the cameras we used, having a field of view approximately 20 x 30 degrees, a surprisingly large range of what might be considered "natural" movements produce image motion which matches case 1 above (all approximately in the same direction). This included rotations about and translations along not only axes parallel to the picture plane, but almost any axis which did not actually intersect the image. Even with the sacrifices in resolution and accuracy made in the interest of achieving real-time performance, the system proved quite successful both at detecting independently moving objects, and ignoring the apparent movement due to its own motion. The expected exception occurred when the independent movement was in the same direction and near the same velocity as the apparent motion, corresponding to landing on the constraint ray. Figure 7 shows the detector's response to a person walking across the field of view as the observer rotates and translates so that the entire scene appears to moving upward. The magnitude of the motion field due to the camera motion is of the same order as that due to the walker. Everything in the image is moving, yet the system reliably identifies those regions whose motion is inconsistent with a rigid interpretation of the world.

4. Detection of Animate Motion

4.1 Theoretical basis

The constraint ray filter described above depends on having some knowledge of observer motion. It is also possible to use knowledge about the motion of the object to be detected. In particular, we take advantage of the fact that, for an observer moving smoothly with respect to a rigid environment, the apparent motion of a world point projected on the image sphere is a relatively slowly changing function of time. Independently moving objects such as people, animals or rolling rocks, on the other hand, frequently *maneuver*, that is, they or their component parts follow trajectories for which the projected velocity changes rapidly compared to the apparent velocity change due to self motion. This suggests that high rates of change or temporal discontinuities in the the projected velocities of world points could provide a basis for distinguishing a wide variety of moving objects against an apparently moving background. Since the types of motion which would be detected by this method are characteristic of living creatures (though they are not the only source) we will use the term *animate motion* to refer to highly accelerated movement used in this context.

The intuitive argument presented above can be formalized as follows. Consider an observer translating with velocity (v_X, v_Y, v_Z) and rotating at $(\omega_X, \omega_Y, \omega_Z)$ with respect to the local Euclidean coordinate system established by the projection of world point p on the image sphere at time 0 (p projects to (0,0) at t=0). The apparent acceleration of the projection of p in terms of this local image plane contains two terms: a coriolis term arising from the interaction of the apparent translation of the projection of p with ω_Z, and a divergence term arising from the apparent expansion

of the image due to translation in the Z direction. In component form the (angular) acceleration in the local coordinate system is

$$a_x = \omega_Z(-\omega_X - \frac{v_Y}{Z}) - 2\frac{v_Z v_X}{Z^2}$$

$$a_y = \omega_Z(-\omega_Y + \frac{v_X}{Z}) - 2\frac{v_Z v_Y}{Z^2}.$$

The first term in each expression is the coriolis effect; the second is the divergence.

We can use these expressions to determine the conditions under which the method is usable. Examining the acceleration equation, we observe that the accelerations due to self motion are on the order of the (angular) velocity of points in the image squared. The accelerations of independently moving objects, on the other hand are on the order of their (projected) angular velocity times the characteristic frequency of their movement. Thus, for example, if the components of the motion field due to self motion and independent motion are of comparable magnitude, the accelerations due to autonomous motion will stand out if objects reverse themselves (180 degree phase shift) significantly faster than they traverse 180 degrees on the image sphere. This condition holds in a large variety of real-world situations.

The above analysis holds for a spherically projected image. In a planar image, there will be an additional acceleration induced by planar distortion at all points away from the image center. By twice differentiating the expression for planar projection we can show that the planar projection of a point moving away from the image center with apparent angular velocity v displays an apparent linear acceleration given (in one dimension) by

$$R\omega^2\frac{\sin(\theta)}{\cos^3(\theta)}.$$

Where R is the distance from the center of projection to the image plane, and θ is the angular distance of the projection of p from the image center. Since $\sin(x)/\cos^3(x)$ is less than unity for $x <$ 34 degrees, as long as the images are smaller than about 70 degrees square, this effect is smaller than those already mentioned.

4.2 Practical considerations

We next address the problem of identifying highly accelerated regions. Because of the effects of occlusion and depth discontinuities, simply differentiating the motion field with respect to time will not work. Conceptually, the problem can be solved by tracking the projections of world points from image to image, but actually doing this is often difficult as it requires solving a correspondence problem. Fortunately however, identifying regions where rapidly changing motion is present is simpler than obtaining quantitative values for the accelerations. The idea is to use the measured motion field at a point in the image to predict where associated world point might project in the next frame. Since the image motion due to non-maneuvering objects changes slowly, the candidate locations can be flagged to indicate that motion as a possible value. If the motion field were known precisely, then each point in the original image would flag a unique location in the next frame. In general, however, since the field is inexactly known, a "footprint" of non-zero area, whose exact shape depends on the nature of the available information, should be flagged (Figure 8). This makes even incomplete information, such as the gradient parallel component available from local differential measurements, usable. Carrying out this operation for each point in the original image produces a constraint map which lists possible values of the motion field for each point in the new image. Typically, since footprints from different antecedent points can overlap, a point in the map may contain more than one value. This constraint map can be compared to the computed field in the new image, and inconsistent points

marked. These represent potential regions of high acceleration.

The above approach can yield a false negative if, due to the local complexity of the original motion field, so many different directions occur close together that their overlapping footprints obscure genuinely new values due to changing motion. For most scenes, however, such regions constitute a small portion of the image if they occur at all, so this will generally not be a big problem. The approach can yield a false positive only at occluding boundaries when previously invisible points appear. Since they were not present in the original image to flag their future location, such points can produce spurious indications of changing motion. This problem can be greatly ameliorated by extending the footprint through and slightly to the counterflow side of its generating point (Figure 8). Thus an object which is partially visible and emerging from behind an occluding object will predict the appearance of similarly moving points at the boundary. The only case where this will break down is on the first appearance of such an object. Such events occur infrequently enough that they do not generally cause a problem and, in fact, represent situations which should be noticed, since a suddenly appearing object may very well be an independently moving one.

The animate motion method of has the advantage that it does not require any information about the observer motion, and is thus applicable for any smooth observer motion rather than just a subset. On the other hand, it can detect moving objects only when they maneuver. For animals, this is almost any time they move, since legs or wings must move back and forth to provide propulsion. Certain manmade targets such as ships and airplanes, on the other hand, may move at the same velocity for long periods. In this case, the technique would be inappropriate. The method would also be sensitive to jitter produced by small rotations of the observer, and thus requires some method of rotationally stabilizing the gaze. It is interesting to note in this connection, that animals which rely much on their eyes almost always possess some such system.

4.3 Real-time implementation and testing.

We have implemented a version of the animate motion detector described above. The first stage is similar to that utilized in our implementation of the constraint ray algorithm described in Section 3, with Datacube boards arranged to compute gradient parallel components of the motion field. This information is subsampled as before, and downloaded to the Sun, which computes the constraints at each pixel of the image using the footprint method described above. These constraints are encoded in an intrinsic image, which is used to filter the next image for motion which violates the temporal smoothness constraints. The algorithm runs in real time (about 10 hertz) robustly identifies animate (e.g. human) motion while the camera translates and rotates in a complicated 3-D environment. Figure 9 shows the detection of a moving hand from a moving camera. Unlike our implementation of the constraint ray filter, this method is not restricted to a limited set of observer motions, and seems to perform equally well under a very wide range of movements, the only criterion being that they not be too violent (in the sense quantified in section 4.1). The limitation of course, is that the system is insensitive to smoothly moving objects. Combining the systems could provide the best of both worlds, with the constraint ray algorithm operating providing detection of smooth independent movement when the observer motion is known.

5. Conclusions

We have described two methods for the detection of independently moving objects by a moving observer. The methods are robust in the sense that they are both resistant to error in the input, and can make use of motion information of low accuracy. This robustness results in large part from the use of matching and filtering techniques based on qualitative features of the motion field rather than numerical computations based on quantitative measurements. The methods are not infallible, in fact, as mentioned in section 3, there exist situations in which no method involving passive monocular observation can distinguish autonomous movement from apparent motion

due to observer egomotion. However, it is possible to characterize precisely the circumstances under which the techniques are effective, and such analysis indicates that they have a broad useful range. Moreover, the domains are somewhat complementary. The first uses information about the motion of the observer, while the second make use of information about the motion of the object of interest. The techniques are primarily useful because they are extremely fast, and can thus serve as interest indicators to direct more sophisticated (and expensive) processing to critical areas. We envisage such detectors being used as the first of three steps in a general purpose motion recognition system. The second step involves stabilization of the area of interest through active visual processes such as fixation and tracking [Brow88b, Brow89, Coom89]. This places the motion of interest in a canonical form that facilitates the final recognition procedure. The third step, is the recognition of the region of interest via a more detailed analysis of its motion. We are currently engaged in developing such a motion recognition system.

Appendix A: Datacube Processing.

The Datacube Maxvideo™ system is a real time, pipelined image processing system implemented as a set of single-board processing modules connected by high speed busses. Image data is transmitted over the busses as a byte stream, and processing modules act on the stream as it flows through them. The processing modules are also connected to a host machine (in our case a SUN) via a VME backplane. Each module has a set of control and status registers which are mapped to a block of VME address space, thus providing the host with the means of controlling and monitoring the activity of the modules. Some modules also contain local memory which is also mapped into the VME space, providing a means of transferring image data from the host to the maxvideo and vice-versa. Configuring a set of modules to perform a particular image-processing task involves setting the control registers via the host computer, and physically rerouting the image bus cables connecting the modules. Processing typically introduces a phase lag into the data stream of between 2 and several thousand pixels, depending on the operation, and dealing with these delays and the resultant synchronization problems is one of the major headaches in using the system. On the other hand, the fact that data is not delayed by a whole frame (as is the case with some other systems such as PIPE) means that multiple processing stages can be cascaded with the introduction of minimal latency. This is a valuable property in real-time applications. A number of different processing modules are available including DIGIMAX, which performs input/output A/D and D/A conversion for RS 170 video signals; VIFIR, which performs an 8 x 8 convolution; ROISTORE, which serves as a frame buffer where image frames can be temporarily held or transferred to and from the host; and MAX-MUX, which provides a 16 x 16 bit lookup table that can be used to implement general functions of two 8 bit variables.

We used the Maxvideo system in both of our prototype movement detectors to compute the gradient parallel component of the motion field. The computation was carried out at 512x512 resolution in real time (30 frames/sec). Our implementation utilized a DIGIMAX, 4 VIFIRS, 3 ROISTORES, and 3 MAX-MUX units, configured as shown in Figure 10. In brief, the output of the DIGIMAX is fed into two cascaded convolution units (VFIRS) which blur the input so that long-range motion can be detected. A ROISTORE is used to delay this signal by one frame. The delayed and undelayed signals are then fed into a MAX-MUX which computes their average and difference using its 16 x 16 lookup table. The difference serves as an estimate of the time derivative, and the average as an unbiased estimate of the gray level. This gray level signal is split and fed through two parallel convolution units which compute estimates of the x and y partial derivatives. A second MAX-MUX converts these values into an estimate of the gradient magnitude and direction. A third MAX-MUX combines the time derivative (difference) signal and the gradient magnitude to produce an estimate of the magnitude of the gradient parallel component of the motion field. Finally, the motion field magnitude and the gradient direction are fed into a second

ROISTORE which subsamples the signals to produce 64x64 frames; these are placed in VME addressable memory from which they are transferred to the Sun host for further processing. The third ROISTORE is used to transfer the results of the host processing to the DIGIMAX for visual display. Synchronization of the various data paths is achieved by judicious use of the delay lines which are provided for this purpose in most of the modules.

References

[Adiv85] - G. Adiv, Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1985, 70-77.

[Anan85] - P. Anandan and R. Weiss, Introducing a smoothness constraint in a matching approach for the computation of optical flow fields, *Proc. Third Workshop on Computer Vision: Representation and Control*, 1985, 186-194.

Anan89] - P. Anandan, A computational framework and an algorithm for the measurement of visual motion, *International Journal of Computer Vision*, **2**, 1989, 283-310.

[Ande85] - C. H. Anderson, P. J. Burt, and G. S. van der Wal, Change detection and tracking using pyramid transform techniques. *Proc. SPIE Conference on Intelligent Robots and Computer Vision*, Boston MA, 1985, 300-305.

[Barn80] - S. T. Barnard and W. B. Thompson, Disparity Analysis of Images, *IEEE Trans. PAMI* **2**, 4, 1980, 330-340.

[Bhan89] - B. Bhanu, P. Symosek, J. Ming, W. Burger, H. Nasr, and J. Kim, Qualitative motion detection and tracking. *Proc. DARPA Image Understanding Workshop*, 1989, 370-398.

[Boll87] - R. C. Bolles Epipolar Plane Analysis: an Approach to Determining Structure from Motion, *Proc. International Joint Conference on Artificial Intelligence*, 1987, 7-15.

[Brow88a] - C. M. Brown (Ed), with D.H. Ballard, T.G. Becker, R.F. Gans, N.G. Martin, T.J. Olson, R.D. Potter, R.D. Rimey, D.G. Tilley, and S.D. Whitehead, The Rochester robot, TR 257, Computer Science Dept., U. Rochester, August 1988.

[Brow88b] - C. M. Brown and R.D. Rimey, Coordinates, conversions, and kinematics for the Rochester Robotics Lab, TR 259, Computer Science Dept., U. Rochester, August 1988.

[Brow89] - C. M. Brown, Centralized and decentralized Kalman filter techniques for tracking, navigation and control, *Proc. DARPA Image Understanding Workshop*, May, 1989, 651-675.

[Burt89] - P. J. Burt, J. R. Bergen, R. Hingorani, R. Kolczynski, W. A. Lee, A. Leung, J. Lubin, and H. Shvayster, Object tracking with a moving camera, *Proceedings of IEEE Workshop on Motion*, Irvine CA., 1989.

[Coom89] - D. J. Coombs, Tracking objects with eye movements, *Proc. Topical Meeting on Image Understanding and Machine Vision*, Optical Society of America, 1989.

[Dins88] - I. Dinstein, A new technique for visual motion alarm, *Pattern Recognition Letters*, **8**, 5, December 1988, 347-.

[Heeg87] - D. Heeger, Optical flow from spatio-temporal filters, *Proc. 1st International Conference on Computer Vision*, 1987, 181-190.

[Heeg88] - D. Heeger and G. Hager, Egomotion and the stabilized world, *International Conference on Computer Vision*, Tampa, 1988, 435-440.

[Horn81] - B.K.P. Horn and B.G. Schunk, Determining optical flow, *Artificial Intelligence* **17**, 1981, 185-204.

[Litt88] - J. J. Little, H. H. Bulthoff, and T. Poggio, Parallel optical flow using local vote counting, *2nd International Conference on Computer Vision*, 1988, 454-459.

[Long81] - H.C. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, *Nature*, **293**, 1981.

[Marr79] - D. Marr and T. Poggio, A Computational Theory of Human Stereo Vision, *Proceedings of the Royal Society of London*, **B(204)**, 1979, 301-328.

[Moro79] - H. P. Morovec, Visual Mapping by a Robot Rover, Proc. IJCAI 1979, 598-600.

[Nage83] - H. H. Nagel, Displacement vectors derived from second order intensity variations in image sequences, *Computer vision, Pattern Recognition, and Image processing*, **21**, 1983, 85-117.

[Nage86] - H. H. Nagel and W. Enkelmann, An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Trans. PAMI*, **85**, Sept. 1986, 565-593.

[Nels88a] - R.C. Nelson and J. Aloimonos, Finding motion parameters from spherical flow fields (or the advantages of having eyes in the back of your head) *Biological Cybernetics*, **58**, 1988, 261-273.

[Nels88b] - R. C. Nelson *Visual Navigation*, Ph.D. Thesis, University of Maryland, 1988, also University of Maryland Computer Science Department TR 2087.

[Nels89] - R.C. Nelson and J. Aloimonos, Using flow field divergence for obstacle avoidance in visual navigation, *IEEE transactions on PAMI*, **11**, 10, Oct. 1989, 1102-1106.

[Praz81] - K. Prazdny, Determining the Instantaneous Direction of Motion from Optical Flow Generated by a Curvilinear Moving Observer. *Computer Vision Graphics and Image Processing*, **22** 1981, 238-248.

[Reic88] - W. Reichard and M. Engelhaaf, Movement detectors provide sufficient information for local computation of 2-D velocity field, *Naturwissenschaften*, 74, 1988 313-315.

[Schu84] - B. G. Schunck, Motion segmentation by constraint line clustering, *IEEE Workshop on Computer Vision: Representation and Control*, 1984, 58-62.

[Thom86] - W.B. Thompson and J. K. Kearney, Inexact vision, *Workshop on Motion, Representation, and Analysis*, May 1986, 15-22.

[Thom90] - W. B. Thompson and T. C. Pong, Detecting Moving Objects, *International Journal of Computer Vision* 4, 1, 1990 39-58.

[Tsai81] - R. Y. Tsai and T. S. Huang, Estimating 3-D motion parameters of a rigid planar patch I, *IEEE ASSP*, 30, 1981, 525-534.

[Tsai84] - R.Y. Tsai and T.S. Huang, Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces, *IEEE Trans. PAMI*, 6, 1984, 13-27.

[Ullm79] - S. Ullman, The interpretation of structure from motion, *Proceedings of the Royal Society of London*, B 203, 1979, 405-426.

[Uras88] - S. Uras, F. Girosi, and V. Torre, A computational approach to motion perception, *Biological Cybernetics*, 60, 1988, 79-87.

[Verr87] - A. Verri and T. Poggio, Against quantitative optical flow, *International Conference on Computer Vision*, June 1987, 171-180.

[Waxm87] - A. Waxman, Image Flow Theory: a Framework for 3-D Inference from Time Varying Imagery, *Advances in Computer Vision*, C. Brown (Editor), Lawrence Erlbaum Inc. 1987

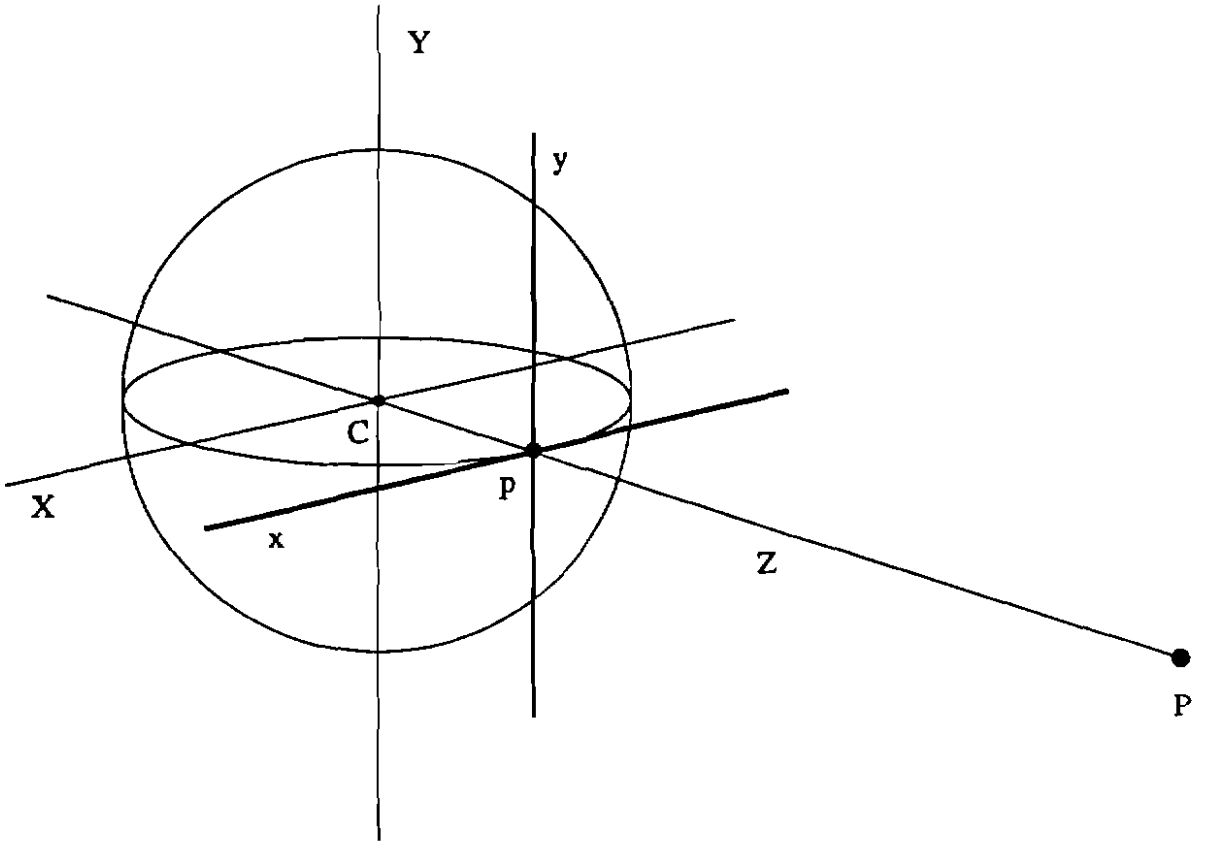

Figure 1: Spherical projection and the local coordinate systems induced by point P and its projection p.

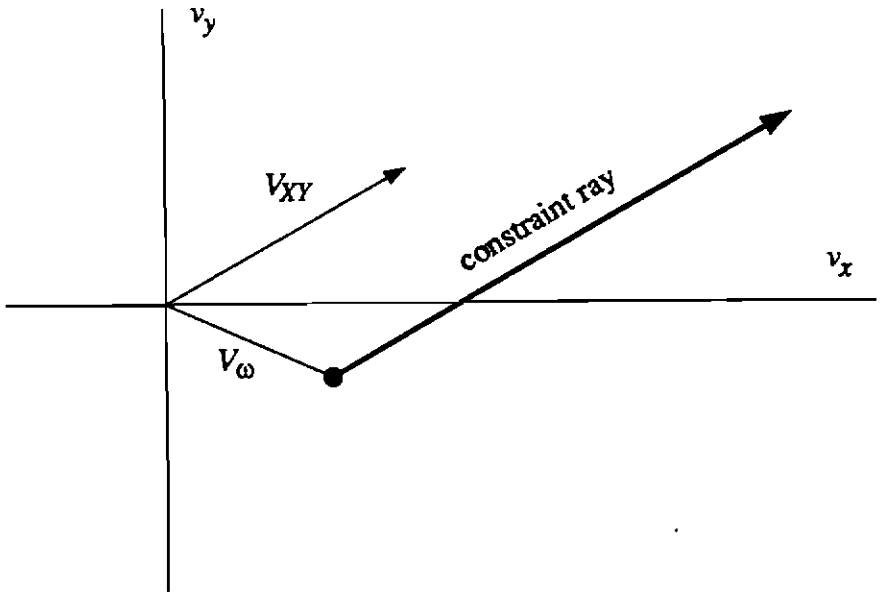

Figure 2: The constraint ray generated by the parametric equation $V=V_\omega+V_{XY}/Z$ for Z ranging from zero to infinity. Intuitively, the ray is generated by adding all positive multiples of the vector V_{XY}, whose direction is determined by the observer's translation, to the constant vector V_ω, which is produced by the observer's rotation. The axes v_x and v_y represent the components of the image velocity.

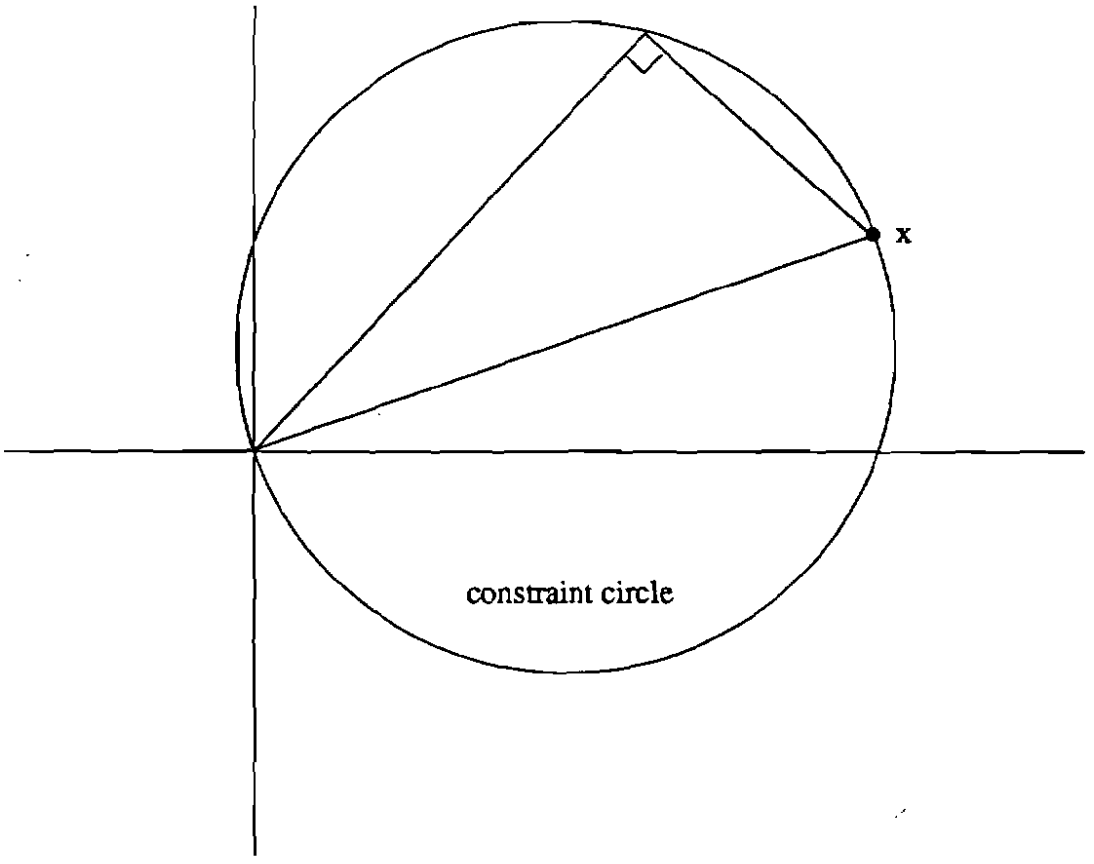

Figure 3: The constraint circle representing all possible projections of the point x onto lines passing through the origin, consequently all possible values for the gradient parallel flow component.

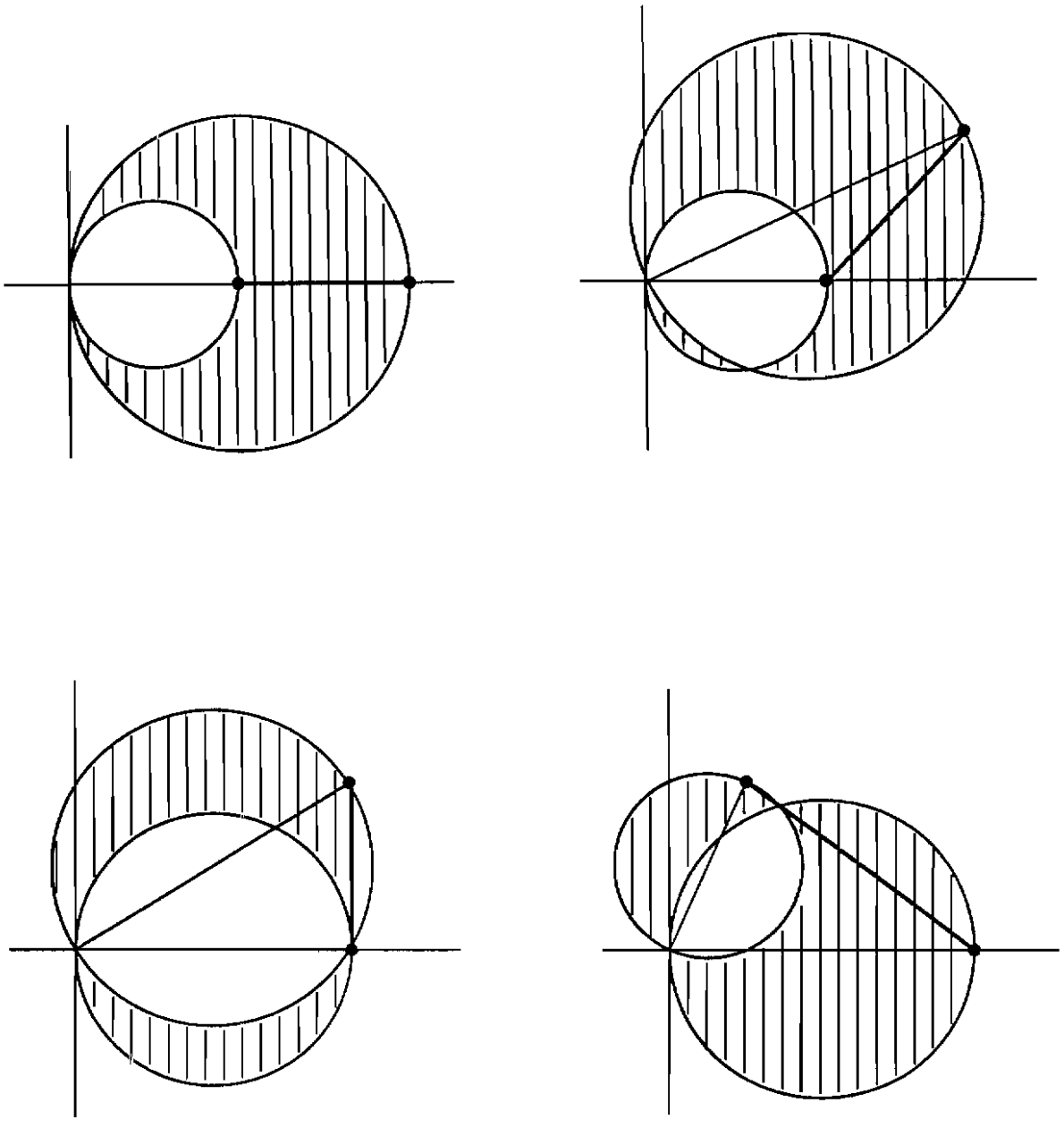

Figure 4: Constraint regions generated by segments of the constraint ray in several situations. The shaded areas represent the portion of velocity space in which the gradient parallel component of the motion field is constrained to lie when a bound can be placed on how close objects may be to the observer. If no such bound can be justified, the constraint ray extends to infinity, and the bounding disk associated with the far end becomes a half plane.

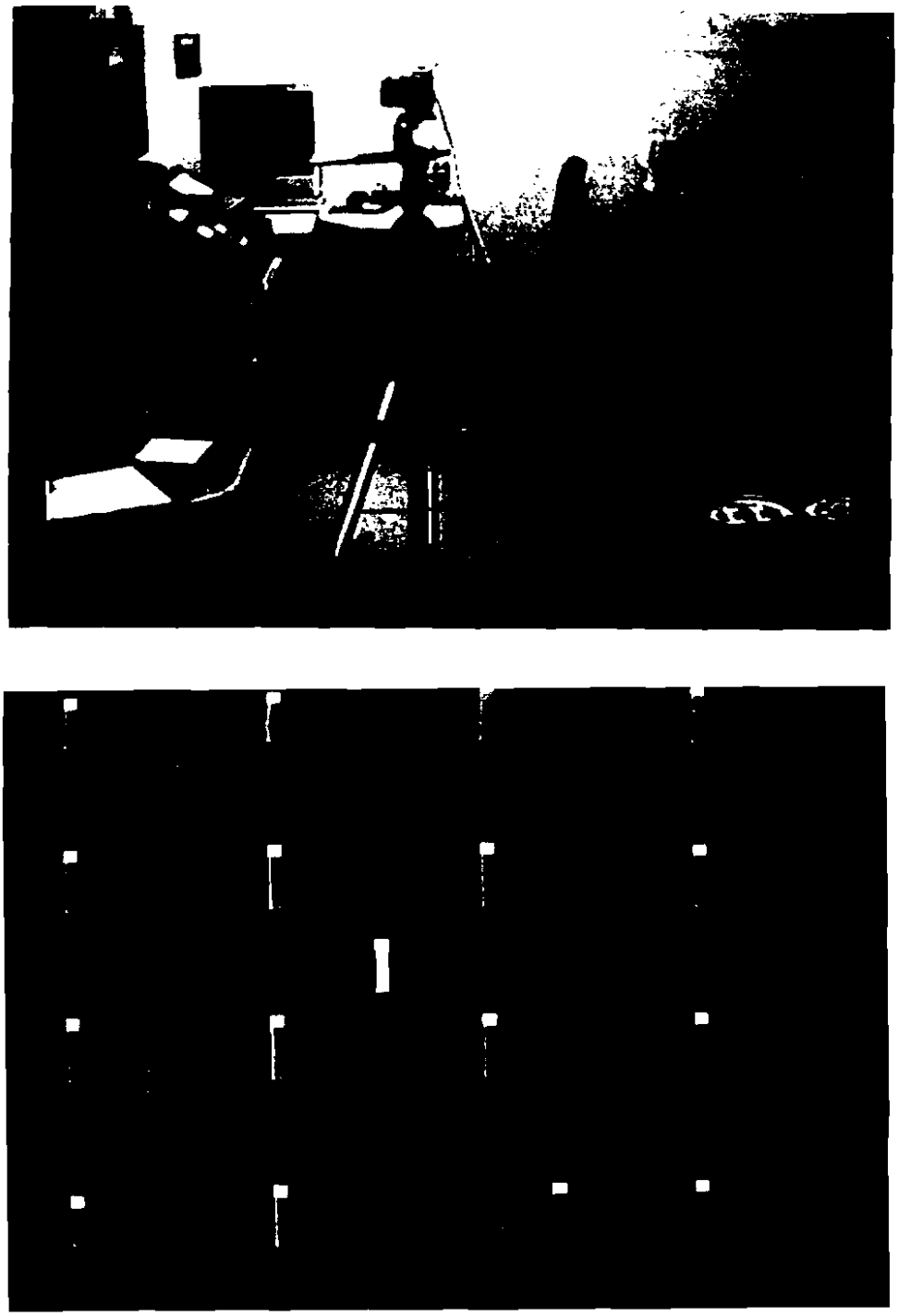

Figure 5: Coarse motion field used to infer class of observer motion. In this case, the camera is translating and rotating downward. The vector in the center represents the system's conclusion that the situation is one in which all motion is in approximately the same direction (Case 1 in the text).

Figure 6: The Rochester Robot consists of a two-eyed, three degree of freedom head attached to a 6 degree of freedom manipulator, and provides a testbed for developing applications involving visual motion, gaze control, and active recognition In the experiments described here, only one eye was used.

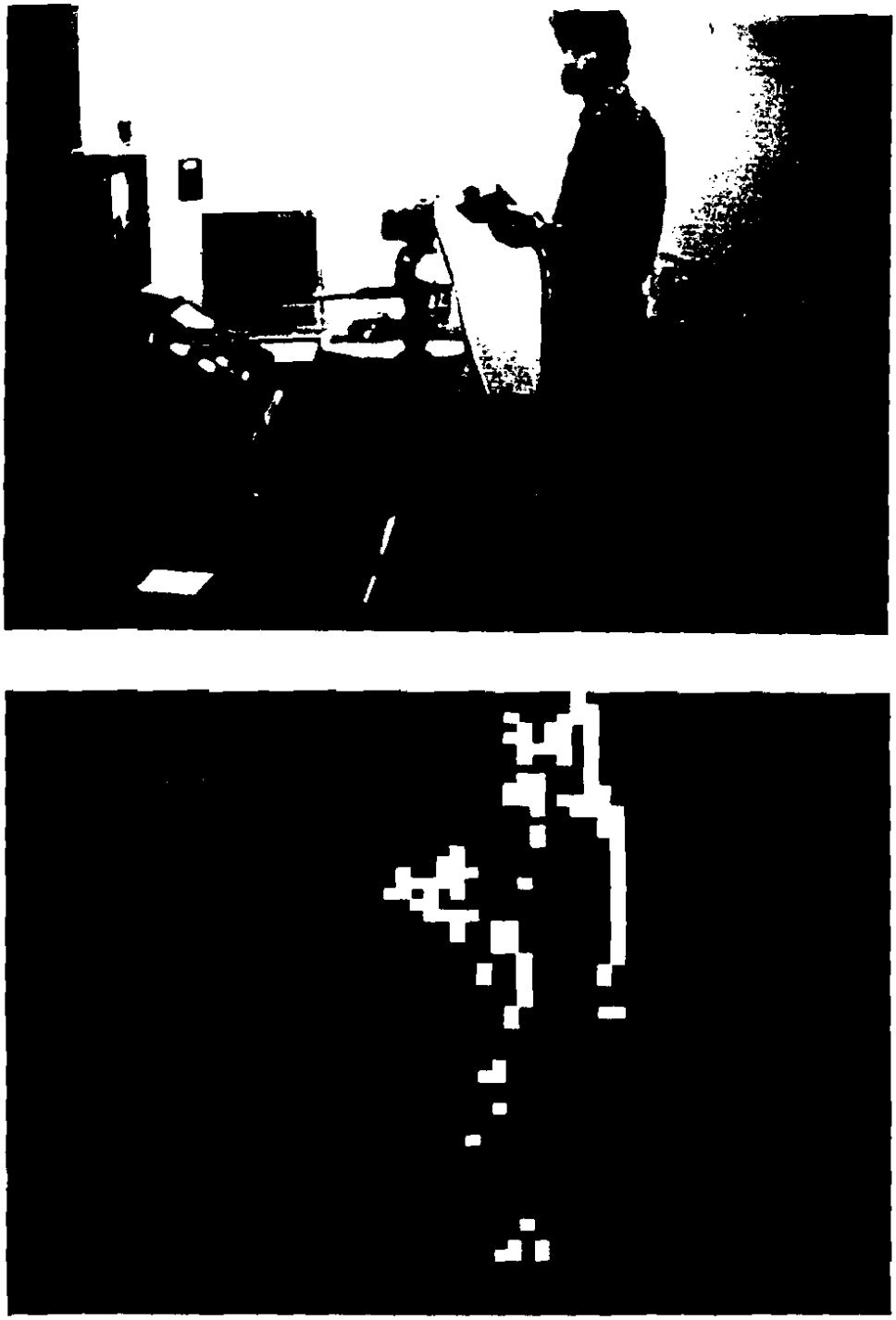

Figure 7: Detection of a walking figure in a moving scene by a movement detector based on constraint ray matching. The camera motion is a combination of rotation and translation whose net effect is to provide impart upward apparent motion (at velocities which depend on distance) to objects in the scene. The walking figure is detected as inconsistent with a rigid interpretation of the motion. The system operates in real time (10 frames/sec).

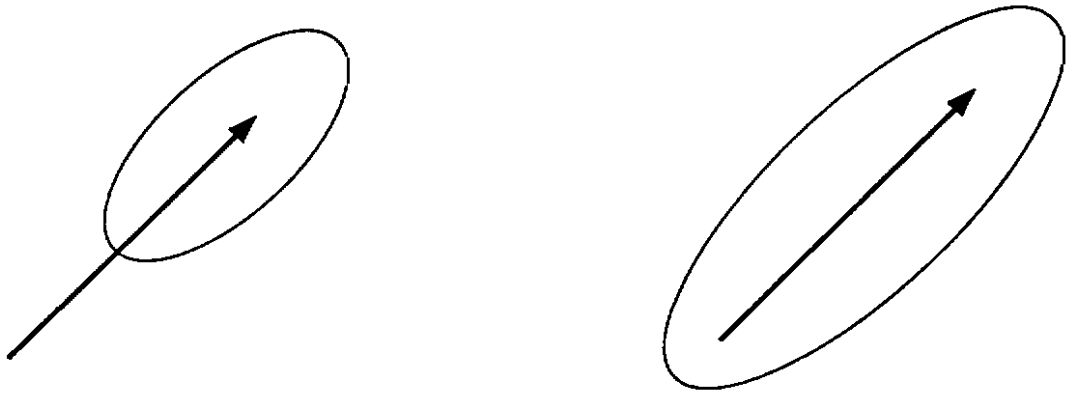

Figure 8: "Footprints" generated by the motion field at a point in the image. The footprint on the left represents an estimate of where the corresponding point will appear in the next image frame, and hence where similar motion can be expected. The footprint on the right has been enlarged to anticipate the continuous appearance of an object from behind an occluding obstacle.

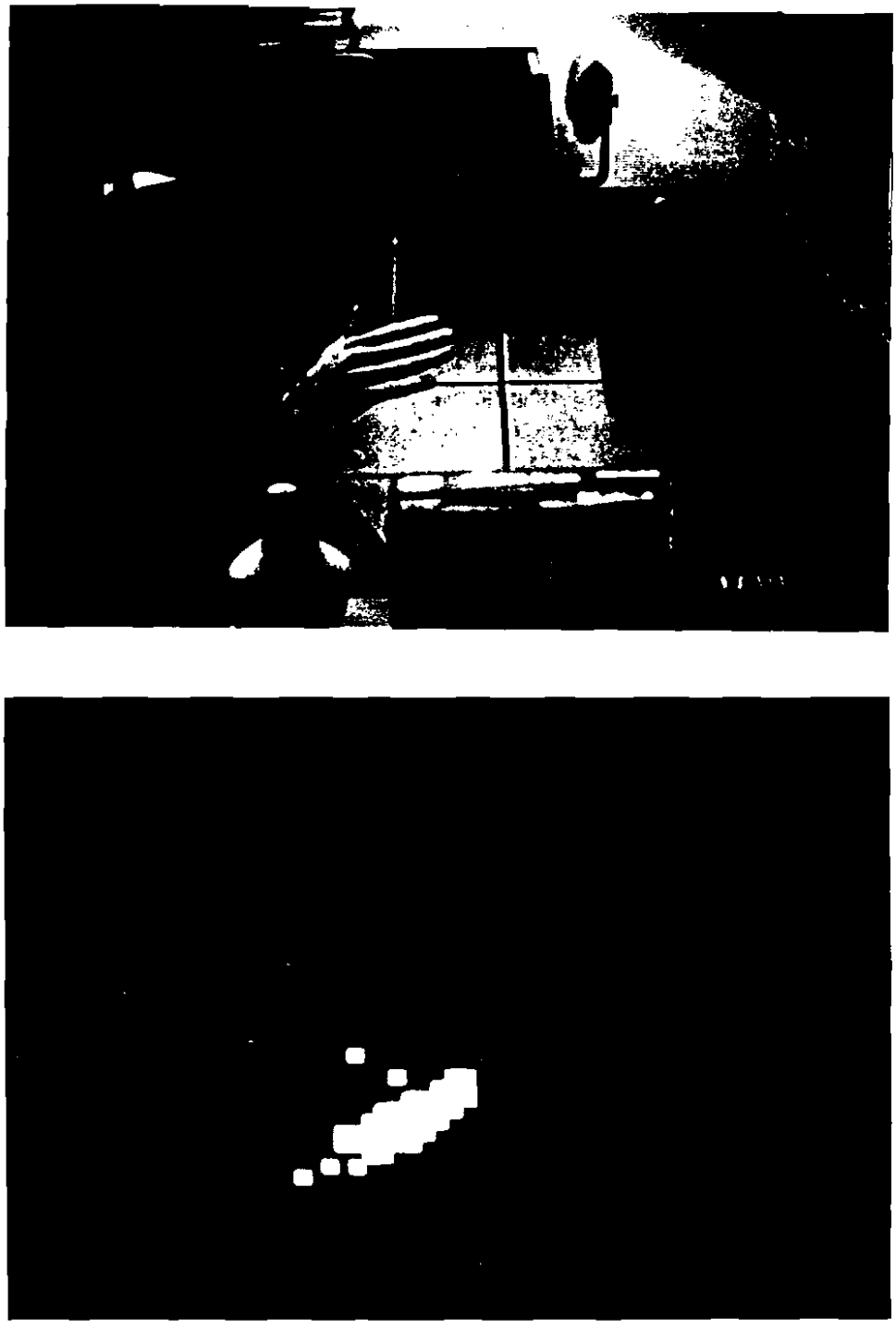

Figure 9: Detection of animate motion. The camera is translating to the left, which means that all the objects in the scene are apparently moving to the right with velocities that depend on their depth. The waving hand, however, can still be detected. The procedure runs in real time (10 frames/sec).

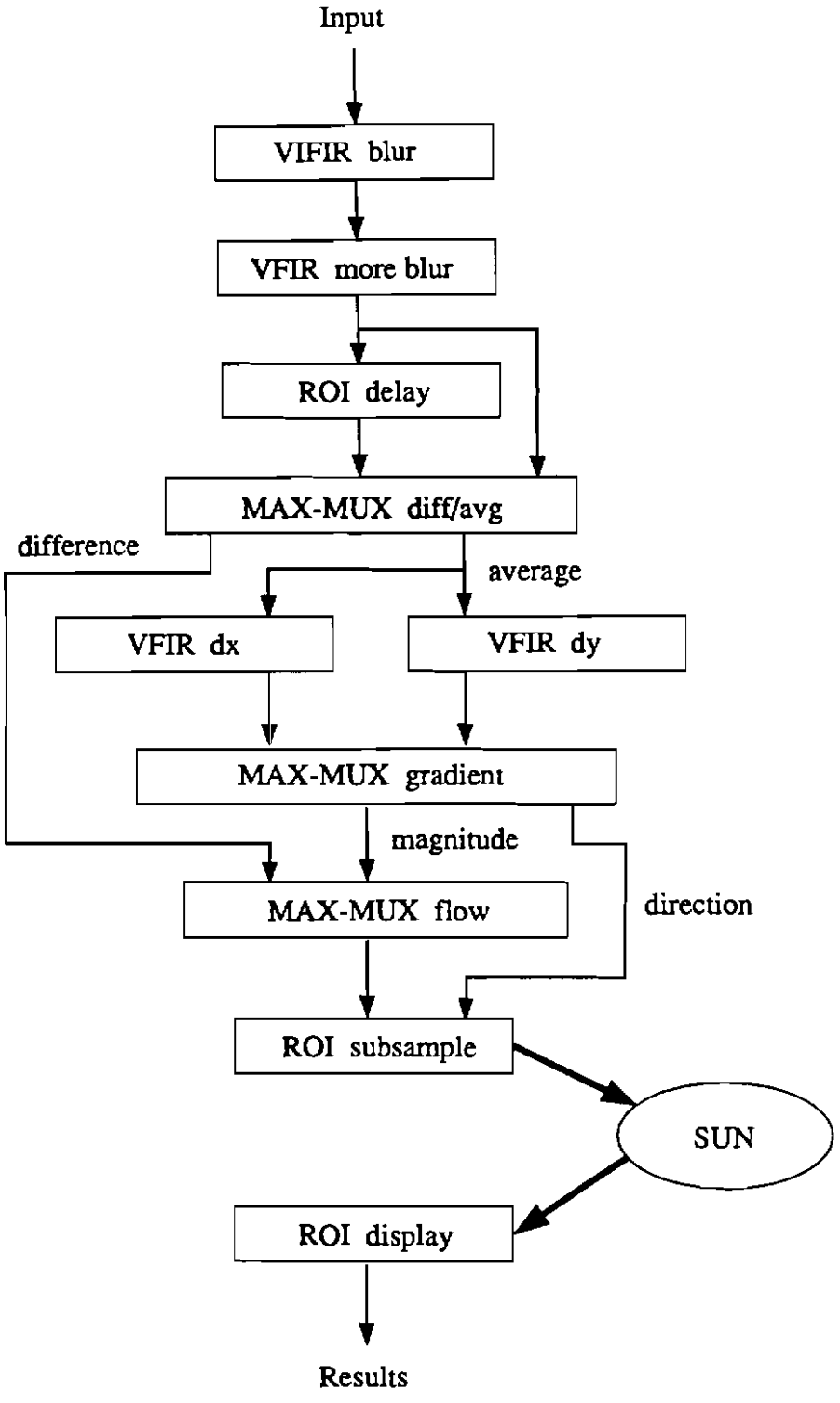

Figure 10: Dataflow diagram showing the computation of the gradient parallel flow component on the Maxvideo image processing system, transfer to the SUN for further processing, and retransfer to Maxvideo for real-time display of results.