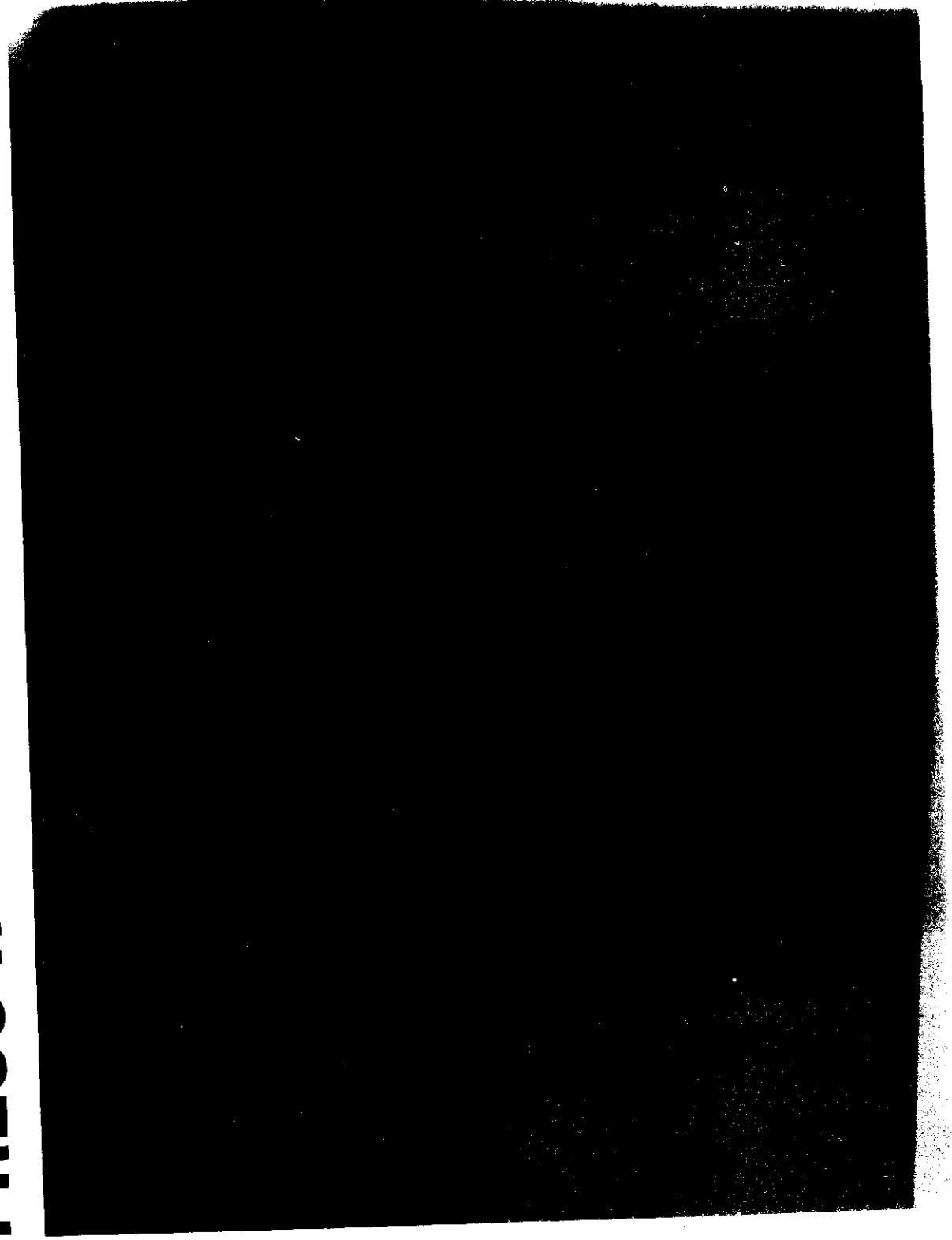


IEEE COMPUTER SOCIETY PRESS REPRINT

IEEE Computer Society
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1264

Washington, DC • Los Alamitos • Brussels • Tokyo

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE

IEEE COMPUTER SOCIETY

Motion, Depth, and Image Flow

by

J.S. Albus and T.H. Hong

Motion, Depth, and Image Flow

James S. Albus
Tsai Hong Hong

Robot Systems Division
National Institute of Standards and Technology
(formerly National Bureau of Standards)

Abstract

Motion relative to a surface is perhaps the most fundamental of visual perceptions. Image flow can be caused either by motion of objects in the world, or by motion of the eye through the world. If knowledge exists of eye motion, dense range maps can be computed locally for all stationary object pixels. Range from image flow can be computed either by gradient methods or by correlation methods. The former are simple but subject to quantization noise. The latter are complex but much more accurate. If eye rotation is known, image flow due to eye translation can be separated from flow due to rotation, and moving objects can be segmented from a stationary background.

Introduction

For creatures of nature, image flow is fundamental to the perception of depth, the detection of objects and events, and the determination of the 3-D shape of surfaces. Image flow is motion of an image on the retina of the eye, or on the photodetector array of a video camera. Image flow can be caused by two phenomena: 1) motion of objects in the world, and 2) motion of the eye through the world. Motion of the eye can be of two types: translation and rotation.

Background

The ability of organisms to see evolved in an environment full of moving objects and patterns, such as trees and grass blowing in the wind, sunlight filtering through moving leaves, running water, swimming fish, flying insects, scurrying rodents, flying birds, fleeing prey, and attacking predators. In the natural world, things that move are likely to be either threats to be avoided, or sources of food to be pursued. For creatures in such a world, the visual perception of motion, and the estimation of range and velocity of objects, is much simpler, and just as important for success in the struggle for survival, as the recognition of shape.

Visual segmentation of stationary objects in a natural environment is extremely difficult. It is confounded by shadows, surface texture, and markings that often produce brightness and color boundaries that are much more distinct than true object or surface boundaries. In order to take advantage of this, both hunter and hunted have evolved elaborate markings and coloration that make spatial analysis of stationary objects difficult. In such an environment, the most effective camouflage technique is to remain motionless.

On the other hand, segmentation of both stationary and moving objects based on differences in their image flow rates can be robust, especially in complex natural environments. For stationary objects and a moving eye, or for moving objects and a stationary eye, or both, shadows and surface markings simply provide texture that generates strong image flow signals; and image flow boundaries provide reliable and unambiguous surface and object segmentation boundaries.

For stationary surfaces, such as a leaf covered forest floor, differential image flow produced by translation of the eye can easily detect ground slope variations and surface discontinuities, such as drop-offs, that are virtually indistinguishable by spatial analysis alone. Shadows and surface markings improve the performance of image flow analysis for eye translation past stationary surfaces, as well as for objects moving relative to a background. This is true regardless of whether the relative motion is due to eye motion, object motion, or both.

It is well known that image flow due to translation is by itself a sufficient visual cue for safe movement through a natural environment filled with complex 3-D obstacles. Helmholtz [1] observed as early as 1925 that monocular image flow provides direct and immediate perception of 3-D space. The rate of apparent motion of a stationary object in the world induced by translation of the eye through the world is inversely proportional to the distance from the eye to the object. Close objects which are potential collision hazards produce large image flow. Distant objects produce little or no image flow.

The psychophysics of image flow was extensively studied by Gibson [2], who treated the perception of space as the perception of a collection of surfaces, and considered motion relative to a surface to be the most fundamental of visual perceptions.

Rogers and Graham [3] have showed that random-dot displays of monocular flow patterns produces the perception of solid oriented surfaces, and that image flow is adequate for shape and depth perception with no other depth information.

In many ways, depth from image flow is more primitive and simple than depth from stereo. The computation of depth from stereo requires registration of images from two entirely separate retinas, plus computation of relative displacement of corresponding features in the two images. Image flow, on the other hand, can be computed locally on a single retina from temporal and spatial derivatives, or from spatio-temporal cross correlation of adjacent pixels, provided eye motion (both translation and rotation) is known.

Most creatures of nature (and robot vehicles as well) possess non-visual vestibular and proprioceptive sensors, and other sensory-motor systems that enable them to estimate ground speed, air speed, or water speed independently of visual input. Of course, eye translation and rotation can also be computed visually from global data integrated over the entire visual field. The brain's final best estimate of eye translation and rotation is

CH2876-1/90/0000/1161$01.00 © 1990 IEEE

undoubtedly derived from the fusion of information from both visual and non-visual sources.

Yet, most researchers in image flow have not treated knowledge of eye translation and rotation as given. Instead, current image flow research tends to center on techniques for computing parameters such as the focus of expansion, and eye velocity, from data in the image itself [4,5,6]. In this paper, knowledge of eye motion, both translation and rotation, is assumed to be given.

When eye rotation is zero, the computation of image flow, and range from image flow, is quite simple. Thus, many creatures of nature have a vestibular reflex that can mechanically stabilize the eye against rotation. However, stabilization of the visual image against rotation can also be achieved neuronally. There is strong psychophysical evidence for neuronal mechanisms that scroll the retinal image so as to effectively cancel eye rotation. This causes the visual world to be subjectively perceived as stationary even though eye rotation causes the retinal image to move about.

Despite the fact that image flow provides a simple and effective measure of a number of visually important parameters such as distance, shape, surface orientation, and boundaries, until recently, most research in machine vision has focused on spatial analysis of static images. The field of image processing got most of its early funding support from aerial and satellite photo reconnaissance analysis, and the effect of this line of thinking is still apparent. The most common approach to flow analysis has been to treat it from the standpoint of photogrammetry. That is, first perform one or more levels of spatial analysis on each of a series of static images, to discover the position and orientation of a set of spatial features, such as edges, vertices, regions, objects, or landmarks. Then analyze the motion of the recognized features as they are tracked from one frame to another in a succession of still frame images.

This approach to image motion analysis works well in photogrammetry, and for tracking landmarks or moving objects. It is, however, a very limited for piloting by sighted creatures through a complex 3-dimensional world. In particular, it is ill suited for such tasks as driving an autonomous land vehicle, or flying at high speeds, like insects and birds routinely do, through a dense tree canopy filled with branches and leaves.

There are several reasons why this is true. First, shadows and surface markings in static images produce many false edges and boundaries which only confuse segmentation algorithms. Separation of object from background in static images of natural scenes is often virtually impossible.

Second, matching spatial features such as edges in one static image with corresponding features in another static image is a combinatorial problem which grows exponentially with the number of features in each image. The amount of texture in natural scenes typically produces so many features in every image that feature matching programs bog down under combinatorial explosion. The computing power required to match spatial features grows, and for any given system, visual performance degrades, with increasing image texture.

Yet, for biological creatures using vision for navigation, more texture improves performance rather than degrades it. It is well known that, for mobile systems piloted by biological vision, from insects to airplanes with human pilots, both accuracy and speed of response improve in environments with more texture, not less.

Third, treating shape analysis as a primitive operation, and image flow as a higher level process, is completely backwards from an evolutionary standpoint. It is well known that primitive species from insects to frogs discriminate shapes poorly. Yet, such creatures are able to perform rapid, precise, visually controlled maneuvers in a world filled with complex 3-dimensional shapes. Primitive natural vision systems apparently rely almost exclusively on image flow for visually guided behavior.

For example, a frog can recognize and track a flying insect with sufficient precision to catch it in mid-flight with a flick of the tongue. Yet a frog will starve to death while surrounded by dead insects because it cannot recognize stationary bugs.

For another example, a house fly can maneuver at speeds of more than one meter per second (about 100 body lengths per second) through a complex 3-D environment such as the branches of a bush or a patch of weeds. During such flights, a fly may encounter tens, or even hundreds, of randomly positioned obstacles (such as branches, stalks, or leaves) every second. The fly routinely accomplishes such flights without collision, despite widely varying viewing conditions such as deep shadows broken by sunlight filtering through a moving leaf canopy.

Flies can also see each other on the wing, and often engage in aerial combat maneuvers as complex as those of human fighter pilots. Analysis of films of such engagements show that the rate of turning of the chaser is proportional to the value of the error angle existing only 30 milliseconds previously [7]. This indicates that the neural image processing latency in the fly must be less than 30 milliseconds.

Yet, there is no evidence to suggest that a stationary fly can recognize stationary shapes at all.

There is also good evidence that many animals rely more on depth from image flow, than depth from stereo. Most prey animals, such as fish, birds, and rabbits have eyes on opposite sides of their heads. If they have stereo vision at all, it covers very little of the visual field. Yet such animals can swim, fly, or run rapidly through complex natural environments (such as reefs, forests, or thickets) filled with obstacles without collisions. They can also effectively flee predators and manuever relative to other members in flocks, herds, and schools.

Such performance strongly supports the idea that image flow analysis is a primitive, simple, and robust method for computing distance, segmenting images, recognizing surfaces, and estimating the 3-dimensional position and shape both of obstacles and targets. Even in human vision, motion is known to be one of the simplest and most important cues for directing attention.

Of course, the analysis of image motion is not confined to the lower levels of visual perception. For example, psychophysical experiments with points of light attached to the joints of humans in a dark room have shown that without movement, the stationary points of light are perceived as isolated random points. Yet, a few hundred milliseconds after the onset of movement, the same points of light can be perceived as 3-D human forms. These are easily recognized as performing complex movements, such as walking backward,

jumping, and sitting down. Complicated scenes with several persons dancing, individually and as couples, can be recognized. Individual persons can even be identified from the characteristics of their gait, and the sex of a subject can often be determined from characteristic movements. [8]

Image motion analysis thus permeates the entire range of visual perception from bottom to top. What is widely unappreciated is that image flow is a visual primitive that, combined with knowledge of eye or camera motion, is able to provide a simple and immediate measure of range and bearing at every pixel in the image.

The Equations of Image Motion on the Egosphere

There are six combinations of eye and object motion that can give rise to image flow.

Case I.

Eye translates without rotation through a stationary world.

Figure 1 is an illustration from Gibson [2] showing image flow on the egosphere of a bird as it flies over the ground. The egosphere is a spherical coordinate system with the self at the center. Note that the projection of the flow lines on the ground are straight lines, all parallel to the line of flight.

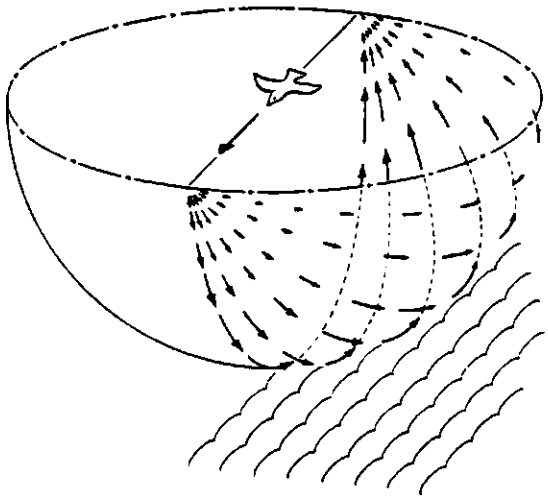


Figure 1. Gibson's example of flow induced by motion. The arrows represent image flow on the egosphere. [Reprinted from J.J. Gibson, *The Senses Considered as Perceptual Systems*, Houghton Mifflin, Boston, 1966, Fig. 9.3]

We may thus define a coordinate system for a velocity egosphere as shown in Figure 2, such that the center of the camera or eye is the origin, the camera velocity is along the y-axis, and the local vertical (anti-gravity) lies in the (yoz) plane. The point at which the y-axis intersects the egosphere is the focus of expansion (FOE). All image points on the egosphere move radially outward from the focus of expansion and travel along great circles (B = constant) on the velocity egosphere toward the negative y-axis, which is the focus of contraction (FOC). Here they converge. The velocity vector is coincident with the egosphere axis connecting the FOE and the FOC.

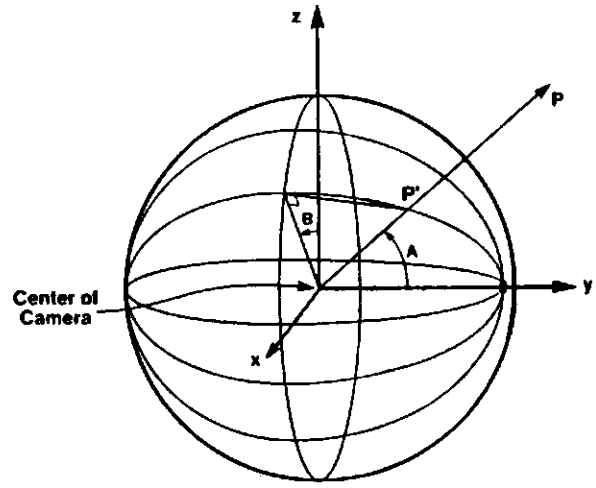


Figure 2. The velocity egosphere, a spherical coordinate system for image flow.

From the derivation in Figure 3, it can be seen that for stationary objects, the rate of motion of every pixel on the egosphere is given by

$$(1) \qquad dA/dt = \frac{v \sin A}{r}$$

where A is the angle between the velocity vector and the image pixel on the egosphere,
dA/dt is the image flow rate on the egosphere,
r is the range to the point covered by the pixel,
v is the eye velocity.

For stationary objects, all motion is along the great circle arcs joining the FOE and FOC. Thus

$$(2) \qquad dB/dt = 0$$

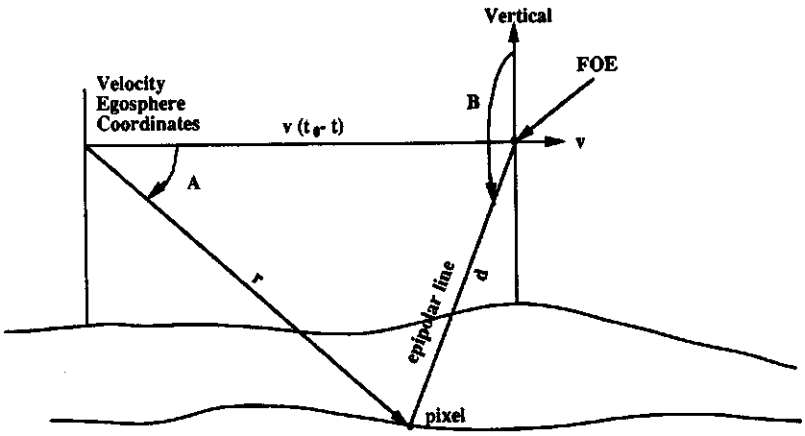
where B is the angle between the yoz plane and the plane of motion of the image flow line,
dB/dt is the time rate of change of B.

The Case I image flow problem is thus one dimensional in A. We can define a pixel array (i,j) such that i is the index along the A dimension, and j is the index on B. In most of the following equations, the j index will be constant, and will be omitted from the equations.

Solving (1) for r gives

$$(3) \qquad r = \frac{v \sin A}{dA/dt}$$

Thus, if the velocity v and the angle A between the pixel ray and the velocity axis is known, the range to an object can be computed directly from measurement of image flow dA/dt.



$$\cot A = \frac{v\,(t_\phi - t)}{d}$$

$$-\csc^2 A\, dA = \frac{-v\,dt}{d}$$

$$dA/dt = \frac{v}{d}\sin^2 A \qquad d = r \sin A$$

$$\boxed{dA/dt = \frac{v \sin A}{r} \qquad r = \frac{v \sin A}{dA/dt}}$$

where v = camera translation vector

A = pixel angular distance from FOE

dA/dt = image flow rate along epipolar line

t_ϕ = time when A = 90°

d = perpendicular distance of pixel from velocity vector

Figure 3. Derivation of image flow equations for eye translation only through a stationary environment.

There are at least two methods for computing dA/dt.

Horn and Schunck

The method developed by Horn and Schunck [9] expresses image flow in terms of a relationship between spatial brightness gradients and temporal derivatives. Transformed into the coordinate system of Figure 2, the Horn and Schunck equations yield

$$(4) \qquad dI/dt + \frac{\partial I}{\partial A}\, dA/dt + \frac{\partial I}{\partial B}\, dB/dt = 0$$

where dI/dt is the temporal derivative of brightness I,
$\frac{\partial I}{\partial A}$ and $\frac{\partial I}{\partial B}$ are the spatial gradients in the A and B directions,
dA/dt and dB/dt are the image flow rates in the A and B directions.

In general, the two unknowns dA/dt and dB/dt cannot be computed from the single equation (4). However, under the conditions of Case I, dB/dt = 0, and equation (4) becomes

$$(5) \qquad dA/dt = -\frac{dI/dt}{\frac{\partial I}{\partial A}}$$

The value of dA/dt computed from (5) can then be substituted in (3) to obtain range r.

The advantage of this method is its simplicity. In a sampled data system, dI/dt may be approximated by

$$(6) \qquad dI(i,t)/dt = I(i,t) - I(i,t-1)$$

where I(i,t) is the intensity of pixel i at time t

On the photodetector array defined above, $\frac{\partial I}{\partial A}$ can be approximated by

$$(7) \qquad \frac{\partial I(i,t)}{\partial A} = I(i,t) - I(i-1,t)$$

or from

$$(8) \qquad \frac{\partial I(i,t)}{\partial A} = 0.5[I(i+1,t) - I(i-1,t)]$$

The disadvantage of the Horn and Schunck method is its sensitivity to errors. There are four sources of errors:

First, the sampled data system approximations to dI/dt and $\frac{\partial I}{\partial A}$ are subject to Nyquist sampling frequency limitations. If I(i,t) contains either spatial or temporal frequency components that are higher than the sampling frequency, aliasing may occur.

Of course, the natural world is infinitely rich in detail. Thus, low-pass spatial filtering must be performed on natural images to smooth them before the Horn and Schunck method can be used. The resultant blurring of the images removes sharp features that contain the most accurate information about image flow.

Second, except in the vicinity of brightness edges, $\frac{\partial I}{\partial A}$ is small and may be zero. Division by small numbers magnifies errors.

Third, the sensitivity of photodetectors in any array is not uniform. The difference in signal from two adjacent photodetectors thus is not necessarily an accurate measure of the difference in illumination. In fact, in the biological eye, where accommodation occurs in the retinal photoreceptors, if the eye is fixated for an extended period, accommodation causes the the image to fade away completely. In this case,

difference in output of adjacent photodetectors provides no information at all about the spatial brightness gradient.

Fourth, thermal noise in photodetectors is not "white", but heavily biased toward low frequencies. The difference in signal from two adjacent photodetectors is thus subject to large amounts of differential low frequency thermal drift.

Thus, the Horn and Schunck method tends to be inaccurate for smooth images, and unreliable for highly textured images.

Cross Correlation

Cross correlation is another method for computing dA/dt. Consider for example, the cross correlation function between the temporal signal I(i,t) from pixel i and that from pixel i+1.

$$(9) \qquad \phi(\tau) = \int_{t_1}^{t_2} [\, I(i+1,t)\, I(i,t-\tau)]\, dt$$

where $\int_{t_1}^{t_2}$ is the integral over the time window $t_2 - t_1$

If τ_{max} is the value of τ where the cross correlation function $\phi(\tau)$ is a maximum, then

$$(10) \qquad dA/dt = \frac{\Delta A}{\tau_{max}}$$

where ΔA is defined as the pixel spacing A(i+1) - A(i).

The range to each point in the world can then be computed for each pixel directly from (3).

Figure 4 shows a data flow diagram by which the functions in (9) and (10) could be computed. A tapped delay line provides values of I(i,t-τ) for different values of τ. A series of multipliers and integrators provides a value of $\phi(\tau)$ for each value of τ. The duration of the time window $t_2 - t_1$ is set by the time constant of the integrators. A maximum circuit computes the τ_{max} as the value of τ that produces the maximum value of $\phi(\tau)$.

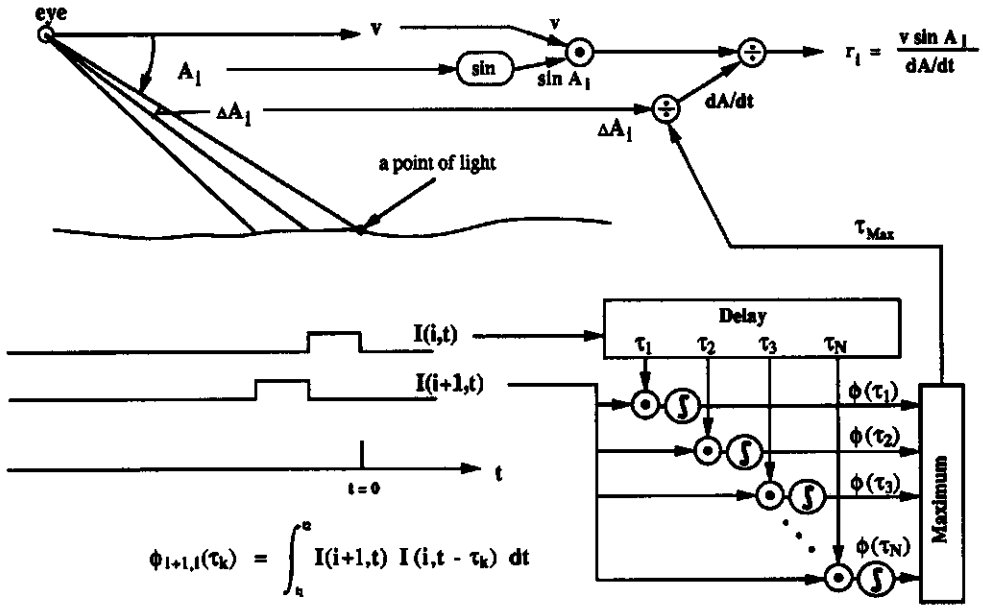


Figure 4. A data flow diagram for obtaining range from image flow by cross correlation between temporal signal I(i,t) from pixel i and I(i+1,t) for pixel i+1.

Figure 5 shows a digital electronic circuit that can compute the functions in (9) and (10). The temporal analog signal from pixels i and i+1 are transformed into digital samples and stored in a pair of shift registers. Between each pixel sample, the left shift register is shifted with respect to the right and the correlation is computed. Each shift operation corresponds to a delay of τ. For each value of τ, the values of I(i+1,t) over a selected window $(t_2 - t_1)$ are multiplied by the values of I(i,t-τ) and summed to give a value of $\phi(\tau)$. Each shift thus results in the computation of the cross correlation function for a single value of τ on the window $t_2 - t_1$. A series of shifts produces $\phi(\tau)$ for all τ. τ_{max} is the value of τ that produces the maximum of $\phi(\tau)$.

In either circuit (Figure 4 or 5), accuracy depends on the resolution in τ. This depends on two factors: the sampling rate, and the sharpness of the peak in $\phi(\tau)$.

The faster the sampling rate, the more accurately τ_{max} can be determined. For one percent accuracy, the sampling rate (in samples per second) should be $100/\tau_{max}$.

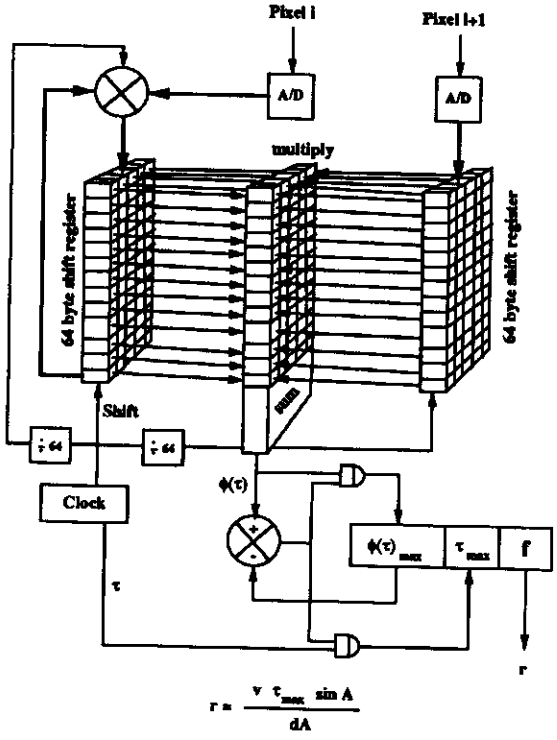


Figure 5. A microelectronic digital hardware that can compute range from image flow by cross correlation of temporal signals from pixels i and i+1.

The sharpness of the peak in $\phi(\tau)$ depends on the sharpness of edges in the image. These can be made sharper by taking the spatial or the temporal derivative before computing the cross correlation. For example, if $I'(i,t) = dI/dt$ we can compute

$$(11) \qquad \phi(\tau) = \int_{t_1}^{t_2} [\, I'(i+1,t)\, I'(i,t-\tau)] \, dt$$

The advantage of the cross correlation method is that it can be very accurate. The disadvantage is that it is quite complex. For each pixel along the flow line, there must be a delay line, a set of correlators that compute a correlation for each value of τ, and a circuit that detects which correlation is the maximum.

There is also an issue of how to select the length of the temporal integration window $t_2 - t_1$. A peak will appear in the $\phi(\tau)$ function for each range in the window. The amplitude of each peak will depend on the length of time the pixels dwell on surfaces at each range. In order to resolve rough surfaces with many sharp edges in depth, the temporal integration window should be short. In order to obtain precise depth measurements from large flat surfaces, the window should be large.

Case II

The eye rotates without translation in a stationary world.

In this case, all image flow is along small circles perpendicular to the axis of rotation as shown in Figure 6.

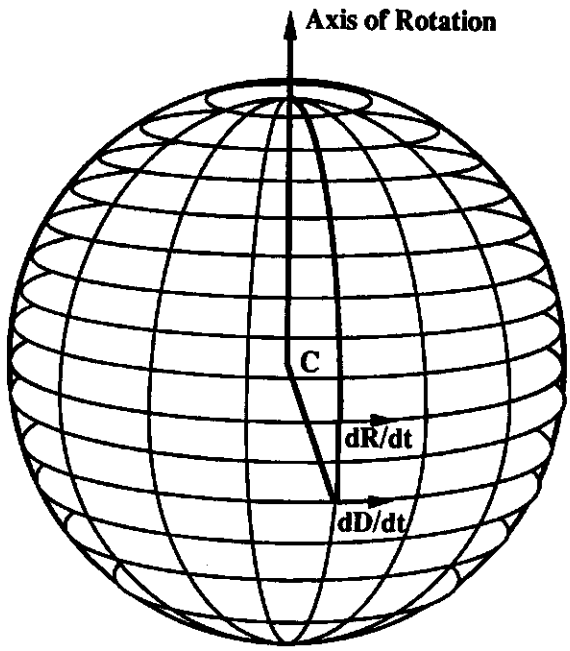


Figure 6. Eye rotation dR/dt without translation produces image flow dD/dt along small circle arcs around the axis of rotation.

The angular rate of motion of all image points along the small circles is the same and equal to the rate of rotation. However, the egosphere radii of the small circles varies as sin C, where C is the angle between the axis of rotation and the image point. From Figure 6, the image flow velocity due to rotation is given by

$$(12) \qquad dD/dt = (\sin C)\, dR/dt$$

where C is the angle between the image point and the axis of rotation,
dR/dt is the rate of eye rotation,
dD/dt is the image flow rate due to eye rotation.

In biological creatures, information about the rate of eye rotation dR/dt is provided by the vestibular system and stretch receptors in the ocular muscles. In artificial systems, it can be provided by rate gyros attached to the camera.

Rotation of the eye without translation yields no information about the distance to points in the world. Eye rotation is, nevertheless, important to perception.

Image Enhancement by Scanning

It is known that in biological creatures the eye is scanned repeatedly over objects of attention and fixation periods are short. The biological eye also exhibits high frequency tremor, which causes the pixels to scan repeatedly back and forth over edges. There is thus good reason to assume that the biological eye uses this motion to enhance the formation of spatial images.

As the eye is rotated, the measured rate of rotation is dR/dt, and the measured temporal derivative of the brightness from a given pixel is dI/dt. From (4) and (12) we can obtain

$$(13) \qquad dI/dt + \frac{\partial I}{\partial D} dD/dt = 0$$

Substituting (12) in (13) and solving for $\frac{\partial I}{\partial D}$ gives

$$(14) \qquad \frac{\partial I}{\partial D} = - \frac{\csc C \, (dI/dt)}{dR/dt}$$

$\frac{\partial I}{\partial D}$ can then be integrated to obtain a brightness image.

$$(15) \qquad I(D) = \csc C \int [(dI/dt) / (dR/dt)] \, dD + K$$

where K is an integration constant

All of the variables on the right hand side of equation (15) are known except for K. Thus, we can compute the spatial function I(D), which is brightness as a function of D (i.e. a small circle in an egosphere image). I(D) can then be computed by each single pixel as it sweeps over the scene. The set of images computed by all of the pixels can be registered and averaged to form a single spatial brightness image.

There are two advantages of generating an image by this method: First, it will contain more contrast and have sharper edges than a fixated image. This is because accommodation tends to reduce contrast in a fixated image. Second, the image will be more free of noise than a fixated image. By using only the temporal derivative of each pixel, the image contains none of the low frequency noise and differences in sensitivity which are characteristic of spatial images from photodetector arrays.

Case III.
The eye simultaneously translates and rotates in a stationary world.

In this case, the image flow at each pixel is given by the vector sum of dA/dt and dD/dt. The translational components of image flow can be separated from the rotational components by transforming each sampled image from the eye egosphere to the inertial egosphere. The inertial egosphere is fixed with respect to rotation. This translation is accomplished by rotating each eye egosphere image $I_e(i,j,t)$ through an angle $R(t) = R(t-1) - dR(t)$ before storing it on the inertial egosphere as $I_i(i,j,t)$. This produces a temporal series of images on the inertial egosphere with no rotational components. These images can then be analyzed by the methods in Case I to obtain range to each pixel.

Raviv [10] has recently derived an analytical solution to range from image flow for case III of the form

$$(16) \qquad r = F(x, z, V, W, dI/dt, \frac{\partial I}{\partial x}, \frac{\partial I}{\partial z})$$

where x and z are the image coordinates of an observed point on a flat image array,
V is the velocity vector in camera coordinates,
W is the rotation vector in camera coordinates,
$\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial z}$ are the x and z gradients of image brightness.

Case IV.
The eye remains fixed as objects in the world move.

In this case, image flow is due entirely to object motion. In general, the relationship between object motion and image flow is shown in Figure 7.

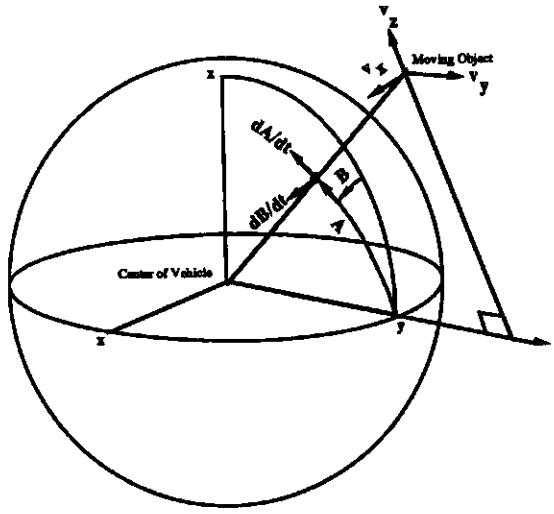


Figure 7. For an object moving with velocity components v_x, v_y, and v_z, the resulting dA/dt and dB/dt can be computed from equations (17) and (18), or (19) and (20).

For a stationary vehicle, image flow rates dA/dt and dB/dt are given by

$$(17) \qquad dA/dt = \frac{v_z \cos A - v_y \sin A}{r}$$

$$(18) \qquad dB/dt = \frac{v_x}{r \sin A}$$

where v_y is object motion parallel to the egosphere y axis,
v_z is object motion in the plane defined by B, and
v_x is object motion perpendicular to the B plane.

When the eye is fixed, only moving objects produce non-zero dI/dt. Non-zero dI/dt signals of course require brightness gradients which tend to exist at object boundaries. Clustering

of pixels with non-zero dI/dt thus provides a basis for depth segmentation. Region growing and blob analysis then yield centroid motion that can be tracked to determine dA/dt and dB/dt.

However, even if image flow components dA/dt and dB/dt are known, the distance r cannot be computed unless the components of object motion v_x, v_y, v_z are known or can be estimated. Alternatively, if the distance r is known or can be estimated, object motion can be computed. However, range and object motion cannot both be computed from image flow alone.

Case V.
The eye translates without rotation as objects in the world also move.

This case can be treated as a combination of Case I and Case IV. Combining equations (1) and (2) with equations (17) and (18) respectively, yields equations (19) and (20).

$$(19) \qquad dA/dt = \frac{v_z \cos A + (v - v_y) \sin A}{r}$$

$$(20) \qquad dB/dt = \frac{v_x}{r \sin A}$$

If the analysis of Case I is applied first, and the results subtracted from the incoming data stream, stationary objects will disappear, because once range is determined, the image flow of stationary points is completely determined.

The remaining pixels that exhibit motion inconsistent with the stationary hypothesis, must therefore arise from moving objects. These pixels can then be used to segment the remaining image into moving regions that can be tracked through a succession of images. The resulting centroid motion then gives rise to dA/dt and dB/dt to be substituted into equations (19) and (20). The components of object motion v_x, v_y, v_z can then be computed if range is known, or range can be computed if v_x, v_y, and v_z are known, as in Case IV.

Case VI.
The eye translates and rotates as objects in the world move.

In this case, eye rotation effects can be removed first by the method described in Case III. The problem then reduces to that of Case V which can be treated as described above.

Obstacle Avoidance Navigation

In addition to providing a convenient coordinate frame for analyzing image flow, the velocity egosphere is also an ideal representation for computing obstacle avoidance. For example, it is clear from Figures 2 and 3 that, if the FOE lies inside the image of an object on the egosphere, then a collision will occur unless the current direction of motion is modified. The minimum velocity deviation required is the angular distance between the FOE and the edge of the object image.

If the FOE lies outside of an object, the clearance between the edge of an object and the current path of the eye can be computed from Figure 3 as

$$(21) \qquad d = r \sin A$$

where d is the perpendicular distance between any point in space and the current velocity vector

Substitution of (3) in (21) gives

$$(22) \qquad d = \frac{v \sin^2 A}{dA/dt}$$

Thus, when eye velocity is known, not only the range, but the clearance between any point on a stationary object and the path of the eye can be computed directly from the observed image flow rate at the pixel covering that point.

Temporal Integration of Image Flow

The image flow discussion so far has been for pixels at a single point in time. It is thus subject to noise in the measurements. Integration of sensory information over an extended period can significantly reduce the single measurement effects of noise.

For example, let us define a linear sensor array such that the axes of the sensor elements lie in the B=180 plane. Let sensor element i lie between angle A(i) and A(i+1), as shown in Figure 8. Let I(i,t) = brightness of sensor i at time t. Let M be the number of elements in the sensor array, so that i = 1, . . . , M. Let dA(i)/dt = the image flow rate at the edge A(i)

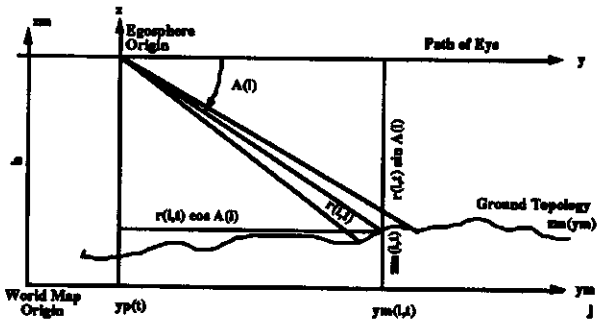


Figure 8. Temporal integration of image flow in world coordinates.

Let the origin of the world map lie in the B=180 plane, ym be parallel to y at a distance of h, and zm intersect the y axis. Then the curve zm(ym) is the intersection of the B=180 plane with the world map surface zm(xm,ym).

Let a grid of points on the curve zm(ym) be stored in a table zm(j). The world map thus is a data structure

j	zm(j)	ns(j)
1		
2		
.		
.		
N		

where N is the number of points on the curve zm(j)

and ns(j) is the number of times zm(j) has been updated

Now let the egosphere move along the y axis. For each time t, and each pixel i, i=1 to M, execute steps 1 through 8 of the following algorithm:

<Begin algorithm>

1. Compute the image flow dA(i)/dt by one of the methods described in Case I.

2. Compute range from the egosphere center to the intersection between the pixel edge at A(i) and the curve zm(ym) by formula (3), i.e. for each pixel i obtain

$$r(i,t) = v \sin A(i) / [dA(i)/dt]$$

The distance along the y axis of the intersection of A(i) with zm(ym) is then r(i,t) cos A(i).

3. Let yp(t) be the point on ym under the camera.

The world map ym coordinate of the intersection between the pixel edge at A(i) and the curve zm(ym) computed from the image flow is then

$$ym(i,t) = yp(t) + r(i,t) \cos A(i)$$

The distance from the y axis to the intersection of A(i) with zm(ym) is r(i,t) sin A(i).

The world map zm coordinate of the intersection between the pixel edge at A(i) and the curve zm(ym) computed from image flow is then

$$zm(i,t) = h - r(i,t) \sin A(i)$$

5. Find an entry j in the world map data structure that is closest to ym(i,t).

$$j = \text{round}[ym(i,t)]$$

6. Compute the difference diff(i,t) between the value of zm(i,t) computed from image flow and the value of zm(j) stored in the map data structure.

$$diff(i,t) = zm(i,t) - zm(j)$$

7. Update the map by the formula

$$zm(j,t+1) = zm(j,t) + kg(j)\ diff(i,t)$$

where kg(j) is the update gain factor for zm(j)

Set the update gain factor to be inversely proportional to the number of times the world map has been updated, i.e.,

$$kg(j) = 1 / (ns(j)+1)$$

8. Increment the value ns(j) = ns(j) + 1
<End of algorithm>

Now as the egosphere moves over the ground, a value of zm(j) will be computed every sample period for each edge i in the array A(i). The map update algorithm is thus a running average filter that produces a temporal integration over the computed terrain elevation points for an entire pass. This update algorithm also has the form of a neural net training algorithm, or a Kalman filter.

Figures 9 through 11 are experimental data from a pass of a linear array camera over a single flat textured surface. The Figures show plots of elevation = zm(j) as a function of translation along the y axis = (j). The plots are computed using

TERRAIN MAPS

Figure 9. Elevation for a single flat surface integrated over 10 frames.

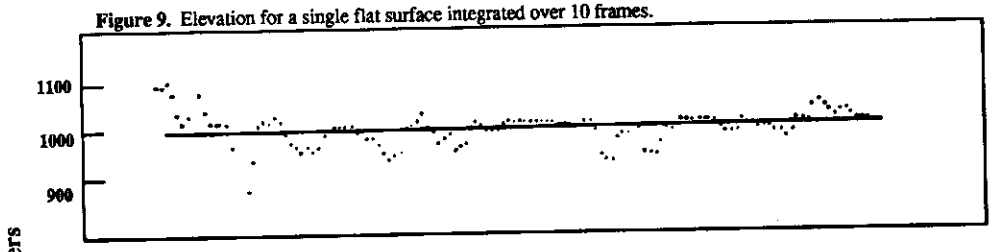


Figure 10. Elevation for a single flat surface integrated over 60 frames.

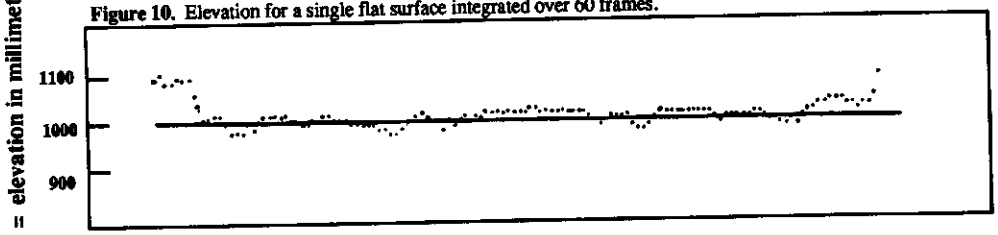
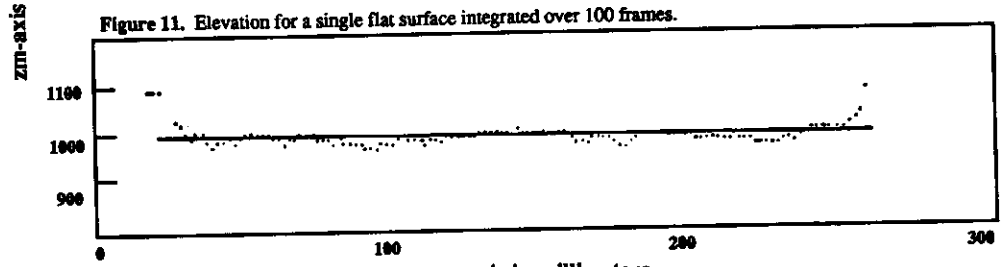


Figure 11. Elevation for a single flat surface integrated over 100 frames.



zm-axis = elevation in millimeters

Figure 12. Elevation for two flat surfaces with a 135 mm step integrated over 2 frames.

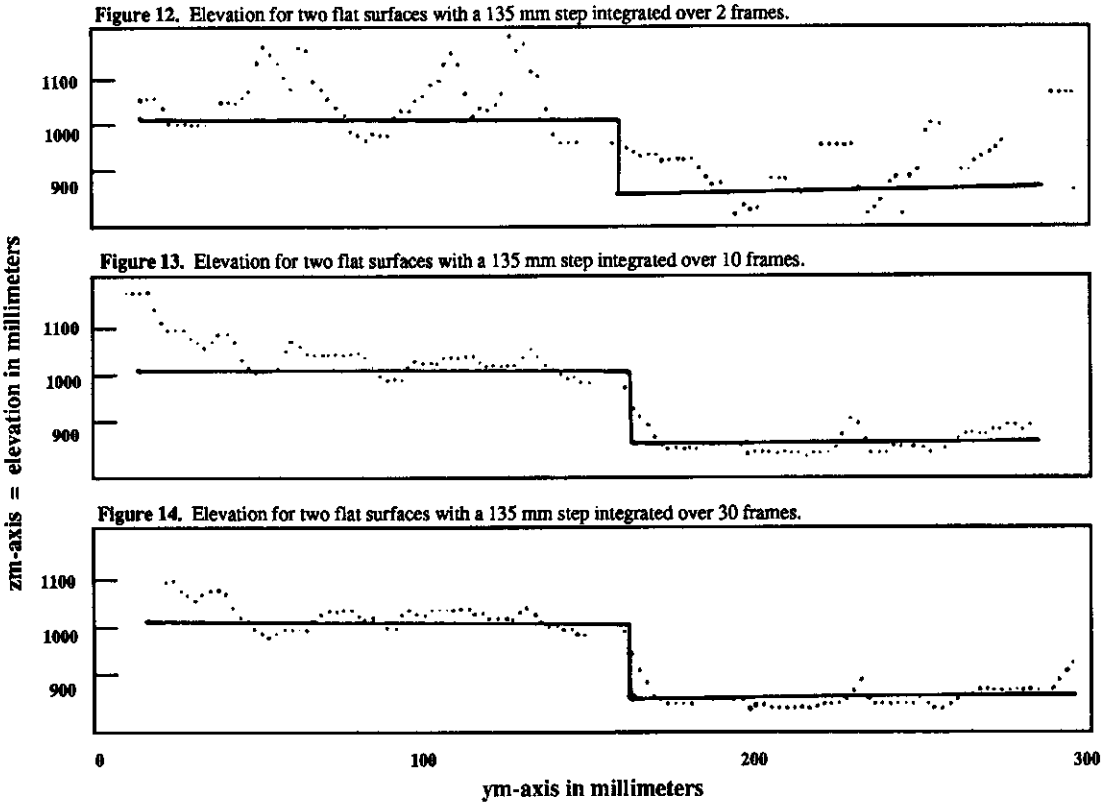


Figure 13. Elevation for two flat surfaces with a 135 mm step integrated over 10 frames.

Figure 14. Elevation for two flat surfaces with a 135 mm step integrated over 30 frames.

the Horn and Schunck method of formulae (4) through (8). Figure 9 shows that within ten frames, measured range from image flow is within about 10% of the correct value, except for the ends of the image where edge effects reduce the amount of temporal integration performed. Figure 10 shows that after sixty frames, accuracy is on the order of 3%. Figure 11 shows some further improvement after one hundred frames.

Figures 12 through 14 show data from a pass over a step discontinuity between two flat surfaces. In Figure 12, with only two frames of integration, accuracy is poor. Figures 13 and 14, with ten and thirty frames of integration respectively, show improvements similar to those in Figures 9 through 11.

Conclusion

In stationary environments, when eye velocity is known, image flow can be a simple, robust, and powerful method for generating dense range images. This may explain the ability of creatures of nature to manuever rapidly in a complex natural world. Accuracy of range from image flow is proportional to how accurately velocity is known. For many important robotic applications, such as driving unmanned ground vehicles, or flying unmanned air vehicles close to the ground, velocity is known well enough to compute useful range from image flow. Image flow may thus become a practical method for obstacle avoidance in unmanned vehicles.

For complex natural scenes, dense range images computed from image flow should prove to be much easier to segment than brightness images. The combination of image flow techniques with more traditional methods thus promises to vastly improve the performance of machine vision systems.

References

[1] Helmholtz, H. von, Treatise on Physiological Optics (translated by J.P.C. Southall) Dover Publications, New York, 1925

[2] Gibson, J.J. The Ecological Approach to Visual Perception Cornell University Press, Ithaca, N.Y. 1966

[3] Rogers, B. and M. Graham, "Motion parallax as an independent cue for depth." Perception 8, 125-134, 1979

[4] Waxman, A. M. and K. Wohn, "Image flow theory: A framework for 3-D inference form time-varying imagery", In Advances in Computer Vision, Volume 1 (C. Brown, Editor), Lawrence Erlbaum Associates, Publishers, Hillsdale, N.J. 1988

[5] Marr, D. and S. Ullman, "Directional selectivity and its use in early visual processing", Proceedings of the Royal Society of London, B211, 151-180, 1981

[6] Duncan, J.H. and T-C. Chou, "Temporal edges: the detection of motion and the computation of optical flow," CS-TR-2042, University of Maryland, 1988

[7] Horridge, G.A., "A different kind of vision: the compound eye", in Handbook of Perception, Vol. VIII, Academic Press, 1978

[8] Johansson, G. "Perception of motion and changing form." Scandinavian J. Psychology 5, 181-208, 1964

[9] Horn, B.K.P. and B.G. Schunck, "Determining optical flow", Artificial Intelligence 17, 1981

[10] Raviv, D. and J.S. Albus, "The computation of range from image flow given knowledge of eye translation and rotation", Journal of IEEE, SMC (to be published)