13. On the Computation of Image Motion and Heading in a 3-D Cluttered Scene

Michael S. Langer[1] and Richard Mann[2]

[1] School of Computer Science
    McGill University, Montreal, Quebec, Canada

[2] School of Computer Science
    University of Waterloo, Waterloo, Ontario, Canada

1   INTRODUCTION

When an observer moves through a static 3-D scene, the retinal image deforms in a way that depends both on the scene geometry and on the observer's motion. These deformations thus provide information about the scene geometry and the observer's motion (Gibson, 1950). Human observers can in many cases perceive the direction of 3-D heading from retinal image motion, even under passive viewing such as watching TV. Performance depends on several factors, such as whether there is sufficient depth variation in the scene and whether the image motion is due to real or simulated eye movements. A variety of scene geometries have been used ranging from a single ground plane to a "3-D cloud of dots". A recent review of human heading perception can be found in (Warren, 1998).

Several computational models of heading perception have also been proposed. These models typically assume that accurate estimates of retinal image velocities can be pre-computed and they take these velocities as their starting point. Such an assumption is often reasonable. Many computer vision algorithms have been developed for computing image velocities, and these algorithms often perform well enough to be used reliably as input to a heading computation. A review of computational models of heading which assume pre-computed image velocities can be found in (Hildreth & Royden, 1998).

Unfortunately, there are many types of scenes for which one cannot simply assume that image velocities are computable. In particular, densely cluttered 3-D scenes such as a forest or grassland. When an observer moves through a densely cluttered 3-D scene, the occlusion relationships between the

objects present are so complex that it is unlikely that the visual system can recover an accurate velocity field. Think of the many branches and leaves visible in a bush. When an observer moves past the bush, motion parallax occurs and the image projection of the various leaves and branches slide past each other. To our knowledge, there is no evidence – either computational or psychophysical – that an accurate velocity field can be computed from such complex motions.

Despite the difficulty of recovering an accurate velocity field in a cluttered 3-D scene, animals such as cats, monkeys, and birds can easily navigate through such scenes. Humans appear to have this ability as well, as cited above. These observations raise the question of how this remarkable performance is achieved. How can an observer compute heading in a 3-D cluttered scene while by-passing the step of computing an accurate image velocity field? Our goal in this paper is to suggest an answer to this question.

We argue that the goal of estimating accurate image velocities in a cluttered 3-D scene is not merely unattainable. The goal is not even well-defined. To estimate the image velocity at a point, the visual system must analyze the time varying intensities in a local neighborhood of the point. But in a cluttered 3-D scene, multiple surfaces are present in most local image neighborhoods because of dense occlusions. Since each surface present typically lies at a different depth, each surface defines a different velocity because of motion parallax. Thus any single velocity vector that is computed really represents a weighted average velocity for the neighborhood, where the weighting depends in a complex manner on the luminance and relative areas of surfaces visible in the neighborhood. It is quite unclear what information this average velocity vector carries about scene depth.

This argument leads us to abandon the traditional goal of pre-computing an image velocity field prior to the heading computation. Instead we introduce a different goal, which is to estimate a family of velocities that is present in a local image region. We show that for an observer moving through a 3-D cluttered scene, the family of velocities in a local image region has a linear structure in image velocity space. This linear structure represents the local motion parallax. We then briefly review our recent algorithm for recovering this family of velocities in a local image region (Langer & Mann, 2001, 2002; Mann & Langer, 2002). Finally, we suggest how these local velocities could in principle be used to compute heading.

2   THE MOTION FIELD SEEN BY A MOVING OBSERVER

The general equations of retinal image velocities field seen by an observer moving through a static scene were derived in (Longuet-Higgins & Prazdny, 1980). Let the observer's translation velocity be (T_x, T_y, T_z) and rotational velocity be $(\Omega_x, \Omega_y, \Omega_z)$. Let $Z(x, y)$ be the depth of the surface point visible at image position (x, y). Assuming the image plane is at $Z = 1$, the velocity field can be written as the sum of a translation field and a rotation field as follows

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \frac{T_z}{Z(x,y)} \begin{bmatrix} x - x_T \\ y - y_T \end{bmatrix} + \begin{bmatrix} xy & -(1 + x^2) & y \\ 1 + y^2 & -xy & -x \end{bmatrix} \begin{bmatrix} \Omega_x \\ \Omega_y \\ \Omega_z \end{bmatrix} \quad (1)$$

The first term is the translation field. It depends on inverse depth. The second term is the rotation component. It does not depend on depth. The special image position

$$(x_T, y_T) = \frac{1}{T_z}(T_x, T_y) \quad (2)$$

is called *the axis of translation* (AOT). When no camera rotation is present, this position is called the *focus of expansion* (FOE). Because we are allowing for camera rotation in this paper, we will use the term AOT rather than FOE. We also define the special image position

$$(x_\Omega, y_\Omega) = \frac{1}{\Omega_z}(\Omega_x, \Omega_y) \quad (3)$$

to be the *axis of rotation* (AOR). In general, the AOR can point in any direction. Specific examples are when the camera rotates about the optical axis (roll), the vertical (pitch), or the horizontal axis (tilt).

There are three standard observations about the translation and rotation fields. First, each vector in the translation field points away from the AOT. Second, the speed of the translation vectors depends on depth. The translation field is smooth over image regions in which depth is smooth. At a depth discontinuity, there is a discontinuity in the magnitude (speed) of the field but not the direction. Third, the rotation field is smooth across the image and, in particular, does not depend on depth.

These observations are the basis for several algorithms for recovering the T and Ω vectors, including the algorithm we will introduce later in the chapter. The algorithms stem from the theory of (Longuet-Higgins & Prazdny, 1980). The first implemented algorithm was described in (Rieger & Lawton, 1985). The idea of their algorithm is to subtract the velocities of neighboring image points that straddle a depth discontinuity. In taking the difference of the two nearby velocity vectors, the rotation components of the two vectors approximately cancel because the rotation field is smooth. Thus, the difference of the two velocities is approximately equal to the difference of the translation components only. Since translation components point toward the AOT, the difference of two neighboring vectors should also point toward the AOT. As analyzed in (Rieger & Lawton, 1985) this last result assumes that the depth difference is large, the points are nearby in the image, and that the velocities are computed accurately.

Rieger-Lawton's experiments address the case of two continuous surfaces separated by a single occlusion boundary. Provided that the depth variation is sufficiently large and accurate flow vectors can be computed near the boundary, heading can be accurately computed. Royden extended this analysis to smooth depth variations, such as an inclined plane (Royden, 1997). Others methods, such as that proposed by Heeger and Jepson (1992), avoid taking local differences of the velocity field and instead minimize a global least squares function.

All these methods assume that the image velocity vector is well-defined in a local image region and that this vector can be computed prior to applying the method. We argued earlier in the chapter, however, that such an assumption is not suitable for a 3-D cluttered scene since many depth discontinuities occur even in local patches. In the next section we develop a model of motion extraction that is suitable for 3-D cluttered scenes and we argue that it captures the information necessary to determine heading direction.

3   MODEL OF LOCAL MOTION PARALLAX

Although a cluttered 3-D scene such as a forest produces a dense set of depth continuities, the motion field is still rich in information. Two specific observations can be made from Eq. (1). First, *the direction of the translation field is smooth and hence is locally constant*. The only exception is within an image region that is near the AOT since the direction of the translation field has a singularity at the AOT. Note that the speed of the translation field is not locally constant since it depends on depth, which may have many discontinuities. Rather, the direction of the translation field is locally constant.

a) Translation Fields (Normalized)

T = (0, 0, 3) T = (-1, 0, 0) T = (-1, -2, 3)

b) Rotation Fields

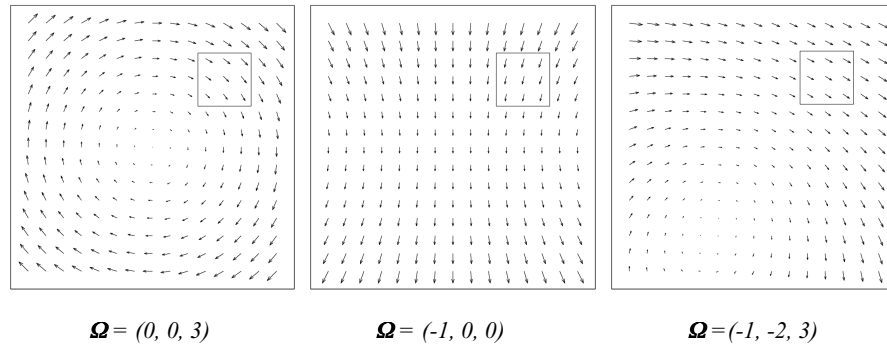Ω = (0, 0, 3) Ω = (-1, 0, 0) Ω = (-1, -2, 3)
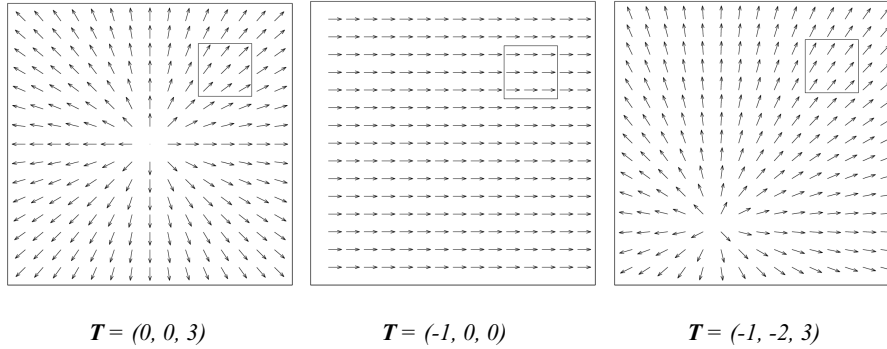
Figure 1. Examples of image motion for a 90° field of view centered on optical axis. a) Normalized translation fields, i.e. direction vectors only. b) Rotational fields. Each vector is approximately 6° from its neighbor. A large square region containing 3-by-3 vectors is shown that is approximately 30° from the optical axis. For each example, the fields are approximately constant within this square region.

Second, *the rotation field is smooth and hence locally constant*. Note that even though the rotation field vanishes near the axis of rotation (AOR), the field is still smooth.

   Examples of the translation and rotation fields are given in Figure 1. These examples support the above two observations. Each vector field shows a 90 degree field of view centered at the optical axis. A 15 × 15 sampling of the vectors fields at roughly 6 degree intervals is shown. Figure 1a shows examples of the translation fields, with velocities normalized to unit length so as to indicate direction only. Except for points near the AOT, the direction of the fields is locally constant. Figure 1b shows examples of the rotation fields. These fields are locally constant everywhere as well.
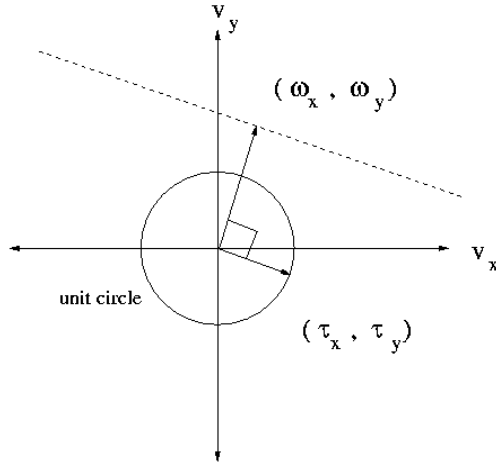
Figure 2. Model of local motion parallax. The set of velocities in a local image region lie on the dotted line of Eq. (4). The τ vector is of unit length and the ω vector is perpendicular to the dotted line.

The two observations above imply a linear model of the image velocities in a local region that does not contain the AOT: *the set of velocities in this region lies on a line in velocity space.* The direction of the line is the direction of the translation component of the velocities, which points to the AOT. The line may be parameterized as follows (see Figure 2)

$$(v_x, v_y) = \alpha(x, y)(\tau_x, \tau_y) + (\omega_x, \omega_y) \tag{4}$$

The velocity at (x, y) depends on the position x, y and on the depth $Z(x, y)$. This dependence is captured by the function $\alpha(x, y)$. The vectors (ω_x, ω_y) and (τ_x, τ_y) are constant for each local region.

When depth is continuous in an image region, the $\alpha(x, y)$ function is also continuous and thus the velocity vectors form a continuous field. An example is the shear field produced by lateral motion relative to a slanted plane. When the 3-D scene is densely cluttered, however, a dense set of depth discontinuities is present and so the velocity field is also discontinuous. When an observer moves past a tree or bush, many leaves are present, even in a 6° field such as the distance between vectors in Figure 1. Since the leaves lie at different depths the velocity field has many discontinuities, yet the model of Eq. (4) still holds.

The vector (τ_x, τ_y) is defined to be the unit vector parallel to the line of Eq. (4). This vector is uniquely defined up to a sign change, provided that $\alpha(x, y)$ indeed takes more than one value over the image region so that the line itself

is well-defined. The vector (ω_x, ω_y) is defined to be the perpendicular vector from the origin to the line of Eq. (4). This vector is typically not of unit length. For our model of local motion parallax, (τ_x, τ_y) is the unit vector pointing in the direction of translation. (ω_x, ω_y) is the component of the rotation vector that is perpendicular to (τ_x, τ_y) .

The model of Eq. (4) is similar to the model of (Longuet-Higgins & Prazdny, 1980) and (Rieger & Lawton, 1985). The image velocities within a small region of the image are assumed to be the sum of a rotation component, which is approximately constant, and a translation component, which depends on depth. By treating the constant and depth-dependent vectors abstractly in our model, we may assume without loss of generality that these vectors are perpendicular.

Given local motion parallax, how do we compute heading direction? We suggest an idea similar to (Longuet-Higgins & Prazdny, 1980) and (Rieger & Lawton, 1985). For several image regions, we estimate the line of Eq. (4), in particular, the (τ_x, τ_y) vector. This vector points toward the axis of translation (AOT) since the difference of any two vectors on this line is due to the difference in translation field only, and hence points toward the AOT. Once we compute the τ vector for several regions of the image, one can recover the AOT by taking the intersection of these lines using standard methods.

The problem that remains, of course, is how to estimate the τ vector from the retinal image of a moving observer. This estimation problem is non-trivial because depth discontinuities occur in abundance even in local regions and so classical methods for computing image velocities cannot be relied on. In the next section, we present a recent method we have developed for solving this problem.

4   MOTION PARALLAX IN THE FREQUENCY DOMAIN

Our idea for estimating the τ vector of the translation field in a local image region uses a frequency-based motion estimation technique, which we have introduced previously (Langer & Mann, 2001), that extends previous frequency-based techniques for measuring image motion (Heeger, 1987). The frequency-based techniques use the following *motion plane* property (Watson & Ahumada, 1985). A 2-D image region that is moving with uniform velocity (v_x, v_y) produces a plane of energy in the 3-D spatiotemporal frequency domain. This plane is given by
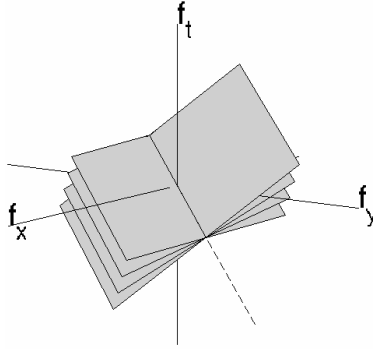
$$v_x f_x + v_y f_y + f_t = 0 \tag{5}$$

Figure 3. The motion described by Eq. (6) results in a family of planes in the frequency domain. The planes intersect at a line that passes through the origin. We refer to this as a "bowtie" pattern.

where f_x and f_y are the two spatial frequencies and f_t is the temporal frequency. In Langer & Mann (2001, 2002), we extended this motion plane property to motions that satisfy Eq. (4) by substituting Eq. (4) into Eq. (5). This substitution defined a one-parameter family of motion planes

$$(\omega_x + \alpha\tau_x)f_x + (\omega_y + \alpha\tau_y)f_y + f_t = 0 \; . \tag{6}$$

As in Eq. (4), the ω and τ constants here are fixed for the local image region and the parameter α takes on a range of values that depends on image position and depth.

The motion planes in Eq. (6) all intersect at a common line that passes through the origin, as illustrated in Figure 3. For this reason, we say that the family of planes has a *bowtie pattern*. We refer to the line of intersection of the planes as the *axis of the bowtie*. We show in Langer and Mann (2002) that the direction of the axis of the bowtie in the 3-D frequency domain is

$$(-\tau_x, \tau_y, \sqrt{\omega_x^2 + \omega_y^2}) \tag{7}$$

4.1   Parallel Motion Parallax

An interesting special case occurs when the line of Eq. (4) passes through the origin so that the ω vector is (0,0). An example is an observer moving through a static 3-D scene and tracking a particular visible surface point in the scene. To track the point, the observer rotates the eye so that the image velocity of the tracked point is reduced to zero (Lappe & Rauschecker, 1993).
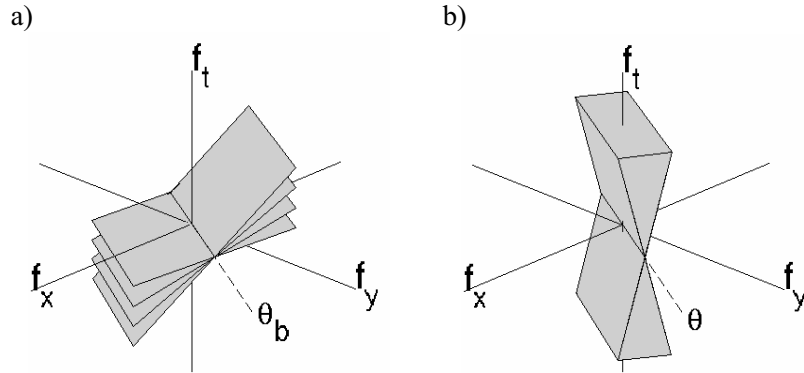
a)   b)

Figure 4. Parallel motion parallax (optical snow). a) Bowtie pattern in the frequency domain. The axis of the bowtie is in the f_x-f_y plane. The angle θ_b is measured from the f_x axis. b) Wedge used to estimate the orientation of the bowtie in Section 4.2. The power within the wedge reaches a minimum when the wedge is aligned with the bowtie in (a), that is, when $\theta = \theta_b$.

Since the tracked point has zero velocity, the line of Eq. (4) that represents the image velocities near that tracked point must itself pass through the origin. The reason is that this line includes the velocity of the tracked point, which is zero. If the line of velocities passes through the origin, then the ω vector is zero. The bowtie model of Eq. (6) reduces to

$$\alpha\tau_x f_x + \alpha\tau_y f_y + f_t = 0 \qquad\qquad (8)$$

We refer to local motions for which the ω vector is zero as *parallel motion parallax.* The reason is that all local motions are in the direction of the τ vector, and thus are parallel to each other. In earlier papers (Langer & Mann, 2001; Mann & Langer, 2002) we called such motion *optical snow*. The reason for this term is that the motion is similar to what a static observer sees during snowfall. Snowflakes that are near the observer fall with a greater image velocity than snowflakes that are far from the observer because of the inverse depth effect (parallax). Yet each local image region may contain snowflakes at a wide range of depths. (Note that this example is slightly different since we have in mind for this example a static observer and a translating 3-D scene, namely a rigid cloud of snowflakes falling vertically.)

For the case of optical snow, the ω vector is zero and Eq. (8) holds such that the axis of the bowtie lies in the (f_x, f_y) plane. We parameterize the direction of this axis with an angle θ_b such that $(\cos\theta_b, \sin\theta_b)$ is the direction of the axis of the bowtie (Figure 4a). Since the axis of the bowtie is in direction $(-\tau_y, \tau_x, 0)$, it follows that
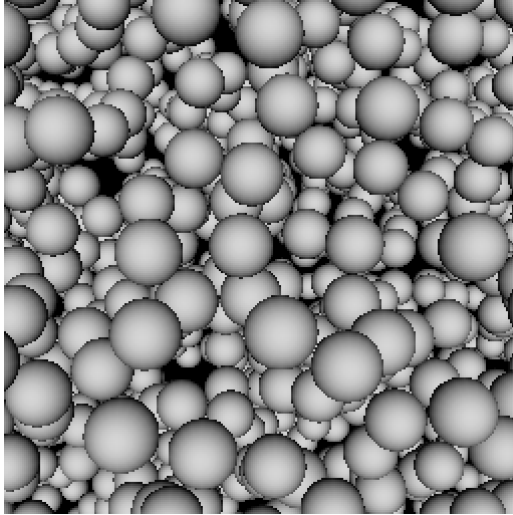
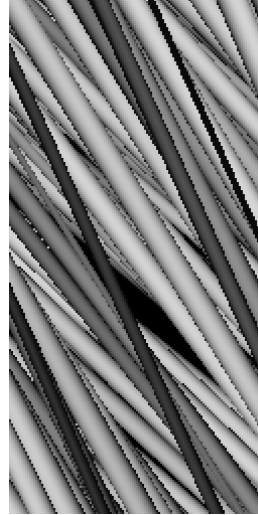a) XY slice   b) YT slice

Figure 5. Synthetic "falling spheres" sequence. a) XY slice of the sequence at the first frame. b) YT slice of the sequence taken at rightmost pixel column of (a). The data consists of 128 frames, each of which is a 256×256 image.

$$(\tau_x, \tau_y) = (-\sin\theta_b, \cos\theta_b) \tag{9}$$

As an example, consider a computer generated image sequence of falling snow as discussed above. The scene is a set of static spheres seen by an observer moving in the vertical direction and was generated using OpenGL (Figure 5). The spheres have constant 3-D size and are placed at random positions within a view volume prior to rendering.

Figure 5a shows one frame of the sequence (XY slice) and Figure 5b shows a YT slice (Adelson & Bergen, 1985). Because different spheres lie at different depths, different spheres have different image speeds. In Figure 5b, each sphere appears as a space-time bar. Different bar orientations indicate different speeds (Bolles et al., 1987).

To visualize the bowtie of power in the 3-D frequency domain for this sequence, we compute the 3-D power spectrum and project it orthographically onto a set of planes,

$$\cos\theta \ f_x + \sin\theta \ f_y = 0 \tag{10}$$

where $\theta \in [0,\pi)$. For any θ, we compute the projection by summing the power along lines parallel to $(\cos\theta, \sin\theta)$. Figure 6 shows the projected power for

several angles θ. A bowtie indeed appears at $\theta = \theta_b = 0$ as expected. Since the image velocities are in the vertical direction, these velocities are of the form $(0, \alpha)$ and so the axis of the bowtie has $\theta_b = 0$. That is, all the motion planes pass through the f_x axis. Moreover, as θ deviates from $\theta_b = 0$ the bowtie gradually diminishes (Langer & Mann, 2001).

One aside worth noting is the aliasing effects in Figure 6 at $\theta = 0$, where the bowtie wraps around the boundaries of the plot. Aliasing is due to the edges of the spheres, which are represented with floating point accuracy by OpenGL. For image sequences taken by a real moving camera, aliasing effects are reduced because of optical blurring at the sensor level prior to spatial sampling. An example is given in (Langer & Mann, 2001).

In the example of Figure 6, the slopes of the planes making up the bowtie were all positive. The reason is that the camera is not rotating and so all the y speeds are negative. A second example, outlined below, has both positive and negative speeds.

Take the same 3-D scene of spheres and the same moving observer but now rotate the observer about the x-axis direction as the camera translates in the y direction. The rotation is such that the optical axis passes through a 3-D point near the center of the view volume in all frames, that is, the observer tracks this 3-D point, reducing its image velocity to zero. Objects on the near side of the tracked 3-D point have downward image velocity and objects on the far side of the tracked point have upward image velocity. Figure 7 shows the XY and YT plots as well as the projected spectrum where the projection is along the bowtie axis. In the YT plot, the slopes are positive for the background objects and negative for the foreground objects. In the projected bowtie, the speeds are both positive and negative.

4.2   Estimating the Translation Direction

The visualization of the bowtie pattern presented in the last section suggests a method for estimating the direction of the motion, that is, the τ vector. The method assumes the ω vector of Eq. (4) is zero such that the axis of the bowtie lies in the (f_x, f_y) plane (Langer & Mann, 2001). At the end of this section, we discuss how to generalize the method to the case where the ω vector is non-zero. Results for this case will be presented in a future paper.
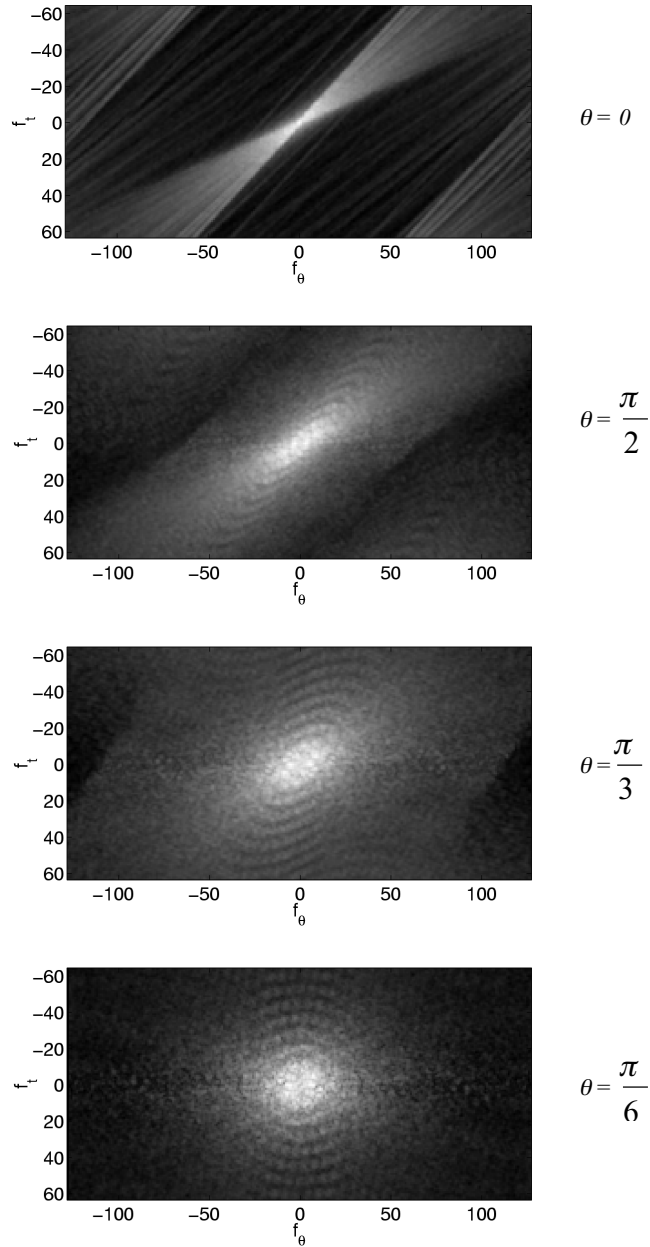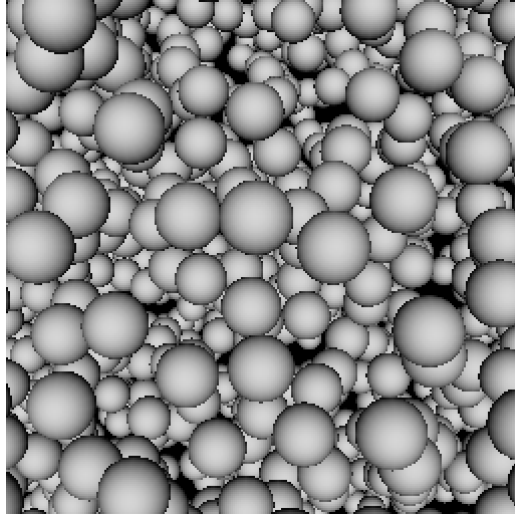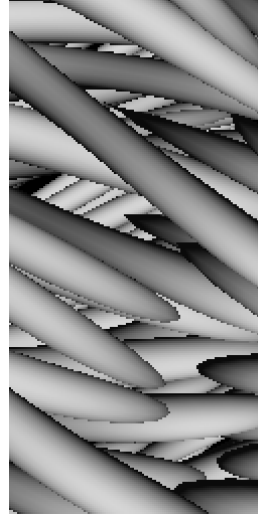
Figure 6. Projected power spectrum for the falling sphere sequence in Figure 5. The bowtie is evident at θ = 0 where the direction of projection is identical to the axis of the bowtie.

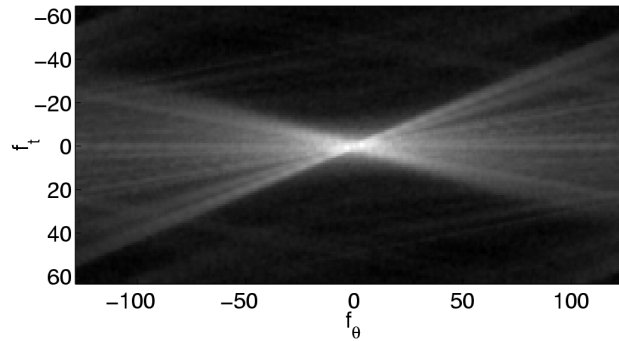a) XY slice   b) YT slice

c)

Figure 7. A sequence similar to that shown in Figure 5, except that as the camera moves upwards it rotates about the x-axis to "track" a point near the center of the view volume. a) XY slice. b) YT slice. c) Bowtie pattern ($\theta = \theta_b = 0$). Positive and negative speeds are present.

For the case where ω is zero we need to estimate the angle θ_b, which is equivalent to estimating the τ vector. We consider all possible directions θ for the axis of the bowtie. For each direction, we sum the power in a wedge of frequencies where the wedge is oriented at an angle θ, as in Figure 4b. The wedge is defined by a slope v_{max}, which is chosen by the user. This slope
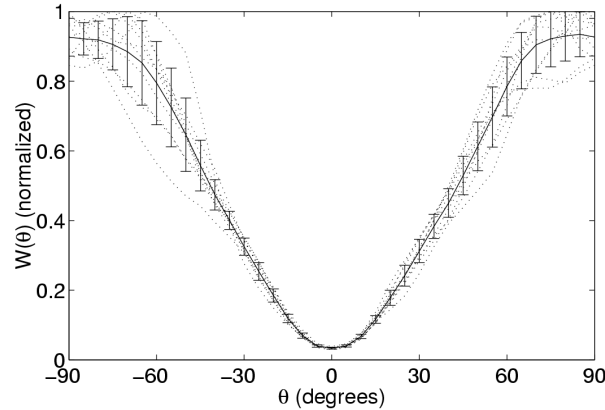
Figure 8. W(θ) for the synthetic sequence in Example 1. The minimum occurs at $\theta = 0°$ corresponding to vertical motion. The results of ten runs are shown (dotted lines), along with the mean response (solid lines) and error bars (one standard deviation).

corresponds to an upper bound on speeds that the user assumes is present in the image sequence.

The direction θ_b is found by looking for a minimum in power within the wedge over all θ. The minimum in power should occur at θ_b because at this angle the bowtie and the wedge should have little overlap. Formally, let $W(\theta)$ be the power in the wedge when the wedge is oriented in direction θ. We are looking for the minimum of $W(\theta)$. Analogously to the visualization of Section 4.1, we compute $W(\theta)$ for a given image sequence by rotating through angles θ at fixed angular increments.

Figure 8 shows $W(\theta)$ for the sequence in Example 1. In this experiment, we chose $v_{max} = 8$ pixels/frame. We show $W(\theta)$ for ten different sequences, as well as the mean curve and the error bars at one standard deviation. Note that while there is some variability in the response, $W(\theta)$ has a well-defined minima at $\theta = 0°$ which corresponds to vertical motion.

How could this algorithm be generalized to the case where ω is non-zero? Our idea is to define a more general wedge. The two planes of the wedge again have high slope but now the planes intersect at the axis of the bowtie, which in the general case does not lie in the (f_x, f_y) plane; recall Figure 3. The same principle applies as in the parallel case, namely that the wedge should contain a minimum of power when the axis of the wedge is identical to the true axis of the bowtie. The idea of the method is thus to search for the general bowtie axis. The search would be over both the azimuth angle θ as before, and also over an elevation angle as well. The elevation angle would determine the length of the ω vector. It remains to be seen how well such a

method can recover the ω and τ parameters for motion in a small image region. This is a topic of our present research.

5 SUMMARY

In this paper we have considered the problem of computing image motion and heading in a static 3-D cluttered scene. The starting point for our analysis was our observation that the motion fields in such scenes contain a rich set of depth discontinuities, and as a result it is unlikely that accurate image velocities can be computed. Thus, models of heading perception that are based on pre-computed image velocities may not be feasible in 3-D cluttered scenes. We have proposed that instead of trying to recover accurate velocities, an alternative goal may be to recover parameterized families of velocities that are present in local image regions. We have argued that this family of velocities has a simple linear form, which gives rise to a bowtie pattern in the 3-D power spectrum. We then outlined a method that could be used to estimate the family of image velocities in a patch, and subsequently to compute direction of heading.

ACKNOWLEDGEMENTS

REFERENCES

Adelson, E. & Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A, 2(2)*, 284-299.

Bolles, R. C., Baker, H. H., & Marimont, D. H. (1987). Epipolar-plane image analysis: An approach to determining structure from motion. *Int. J. Comput. Vision, 1,* 7-55.

Gibson, J. J. (1950). *The Perception of the Visual World.* Houghton Mifflin, Boston.

Heeger, D. J. (1987). Optical flow from spatiotemporal filters, *First International Conference on Computer Vision,* p. 181-190.

Heeger, D. J. & Jepson, A. D. (1992). Subspace methods for recovering rigid motion. I: algorithm and implementation. *Int. J. Comput. Vision, 7,* 95-117.

Hildreth, E. C. & Royden, C. S. (1998). Computing Observer Motion from Optical Flow. In Watanabe, T. (Ed.) *High-Level Motion Processing – Computational, Neurobiological, and Psychophysical Perspectives* (p. 269-294), Massachussets: MIT Press.

Langer, M. S. & Mann, R. (2001). Dimensional analysis of image motion, *IEEE International Conference on Computer Vision,* p. 155-162.

Langer, M. S. & Mann, R. (2002). Tracking through optical snow, *2nd Workshop on Biologically Motivated Computer Vision (Lecture Notes in Computer Science), (in press).*

Lappe, M. & Rauschecker, J. P. (1993). A neural network for processing of optical flow from egomotion in man and higher mammals. *Neural Comput., 5,* 374-391.

Longuet-Higgins, H. & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proc. R. Soc. Lond. B, 208,* 385-397.

Mann, R. & Langer, M. S. (2002). Optical snow and the aperture problem, *International Conference on Pattern Recognition,* Quebec City, Canada.

Rieger, J. H. & Lawton, D. T. (1985). Processing differential image motion. *J. Opt. Soc. Am. A, 2,* 254-260

Royden, C. S. (1997). Mathematical analysis of motion-opponent mechanisms used in the determination of heading and depth. *J. Opt. Soc. Am. A, 14(9),* 2128-2143.

Warren, W. H. (1998). The State of Flow. In Watanabe, T. (Ed.) *High-Level Motion Processing – Computational, Neurobiological, and Psychophysical Perspectives* (p. 315-358), Massachussets: MIT Press.

Watson, A. & Ahumada, A. (1985). Model of human visual-motion sensing. *J. Opt. Soc. Am. A, 2(2),* 322-342.