Computational Vision
U. Minn. Psy 5036
Daniel Kersten
Lecture 26: Spatial Layout, Structure from Motion

Initialize

Outline

Last time

Object recognition continued

Ways of matching stored information about 3D objects using 2D representations and transformations

Today

Computational theory for estimating relative depth, camera motion

Challenges to computational theories of depth and spatial layout

Spatial layout: Where are objects? Where is the viewer?

Recall distinctions: Between vs. within object geometry.

Lecture 15 on surface geometry.

Where are objects?

Absolute

Distance of objects or scene feature points from the observer.

"Physiological cues": Binocular convergence--information about the distance between the eyes and the angle converged by the eyes. Crude, but constraining. Errors might be expected to be proportional to reciprocal distance. Closely related to accommodative requirements.

"Pictorial cue"--familiar size

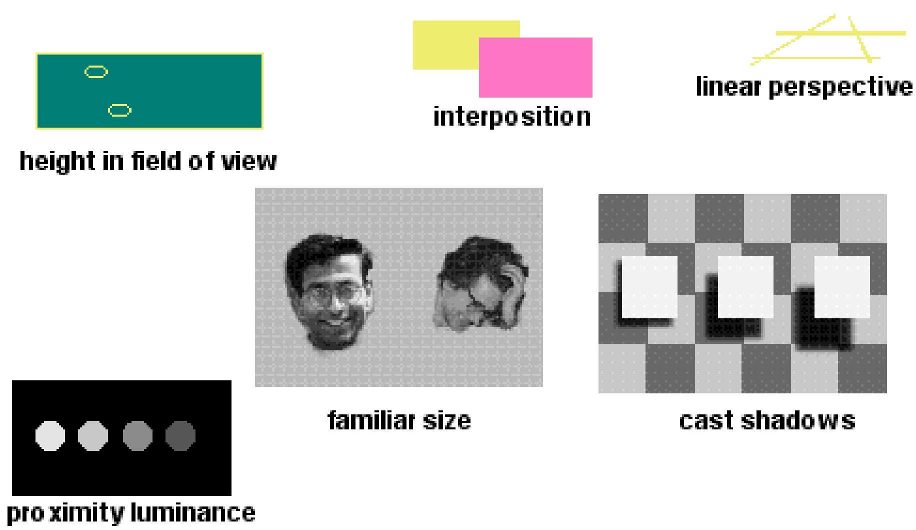
Relative

Distance between objects or object feature points. Important for scene layout.

Processes include: Stereopsis (binocular parallax) and motion parallax.

Also information having to do with the "pictorial" cues: occlusion, transparency, perspective, proximity luminance, focus blur, also familiar size & "assumed common physical size", "height in picture plane", cast shadows, texture & texture gradients for large-scale depth & depth gradients

Examples of pictorial information for depth



■ Cooperative computation & cue integration

...over a dozen cues to depth. Theories of integration (e.g. stereo + cast shadows). Theories of cooperativity (e.g. motion parallax <=> transparency).

Vision for spatial layout of objects, navigation, heading and for reach

Where is the viewer? And where is the viewer headed?

Computing scene structure from motion information provides information for vision. Can't say where the viewer is in absolute terms, but can say something about the relative depth relationships between objects, and can say something about heading direction, and time to contact.

Calculating structure from motion and heading from the motion field

Estimation of relative depth and eye (or camera) motion

Introduction

Earlier we saw:

1) how local motion measurements constrain estimates of optic flow, and thus the motion field.

2) how a priori slowness and smoothness contraints constrain dense and sparse estimates of the flow field (e.g. Weiss et al.).

How can we use an estimate of the motion field to estimate useful information for navigation--such as relative depth, observer motion, and time to collision??

Goals

Estimate relative depth, and eye's motion from motion field, estimates of time-to-contact

Ultimately we would like to gain some understanding of the environment from the moving images on our retinas. There are approaches to structure from motion that are not based directly on the motion field, but rather based on a sequence of images in which a discrete set of corresponding points have been identified (Ullman, S., 1979; Dickmanns).

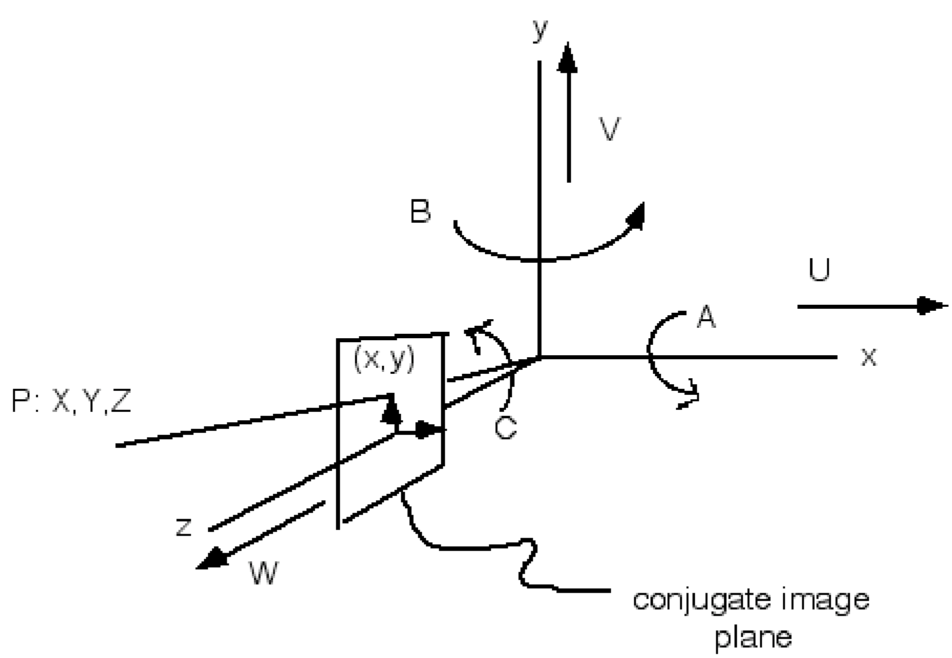
Alternatively, suppose we have estimated the optic flow, and assume it is a good estimate of the motion field--what can we do with it? Imagine the observer is flying through the environment. The flow field should be rich with information regarding direction of heading, time-to-contact, and relative depth (Gibson, 1957).

In this section we study the computational theory for the estimation of relative depth, and camera or eye-point heading from the optic flow pattern induced by general eye motion in a rigid environment. We follow a development described by Longuet-Higgins, H. C., & Prazdny, K. (1980). (See also Koenderink and van Doorn, 1976, Horn, Chapter 17, Perrone, 1992 for a biologically motivated model, and Heeger and Jepson, 1990).

Rather than following the derivation of Longuet-Higgins et al., we derive the relationship between the motion field and relative depth, and camera motion parameters using homogeneous coordinates.

Setting up the frame of reference and basic variables

Imagine a rigid coordinate system attached to the eye, with the origin at the nodal point. General motion of the eye can be described by the instantaneous translational (U,V,W) and rotational (A,B,C) components of the frame. Let P be a fixed point in the world at (X,Y,Z) that projects to point (x,y) in the conjugate image plane which is unit distance in the z direction from the origin:



Goal 1: Derive generative model of the motion field, where we express the motion field (u,v) in terms of Z, U,V,W,A,B,C.

■ Express velocity (X,Y,Z) of world point P in terms of motion of the frame of reference

Let **r**(t) represent the position of P in homogeneous coordinates:

$$\mathbf{r}(t) = (X, Y, Z, 1)$$

An instant later, the new coordinates are given by:

$$\mathbf{r}(t+\Delta t) = \mathbf{r} + \Delta\mathbf{r} = (X + \Delta X, Y + \Delta Y, Z + \Delta Z, 1) = \mathbf{r}\Delta R_{\theta_x} \Delta R_{\theta_y} \Delta R_{\theta_z} \Delta T$$

where infinitesimal rotations and translations are represented by their respective 4x4 matrices. (Note that matrix operations do not in general commute). Then,

$$\Delta T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\Delta x_0 & -\Delta y_0 & -\Delta z_0 & 1 \end{bmatrix}$$

and

$$\Delta R_{\theta_x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(\theta_x) & -sin(\theta_x) & 0 \\ 0 & sin(\theta_x) & cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -\Delta\theta_x & 0 \\ 0 & \Delta\theta_x & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + h$$

Using similar approximations for the other rotation matrices, and the relation

$$\Delta\mathbf{r} = \mathbf{r}\Delta R_{\theta_x} \Delta R_{\theta_y} \Delta R_{\theta_z} \Delta T - \mathbf{r}I$$

we have

$$(\Delta X, \Delta Y, \Delta Z, 0) = (X, Y, Z, 1)\begin{bmatrix} 0 & \Delta\theta_z & \Delta\theta_y & 0 \\ \Delta\theta_z & 0 & -\Delta\theta_x & 0 \\ -\Delta\theta_y & \Delta\theta_x & 0 & 0 \\ -\Delta x_0 & -\Delta y_0 & -\Delta z_0 & 0 \end{bmatrix} + \textit{higher order term}$$

By dividing by Δt, we can derive the following relations:

$$\frac{\Delta X}{\Delta t} = \frac{\Delta \theta_z}{\Delta t} Y - \frac{\Delta \theta_y}{\Delta t} - \frac{\Delta x}{\Delta t} = CY - BZ - U$$

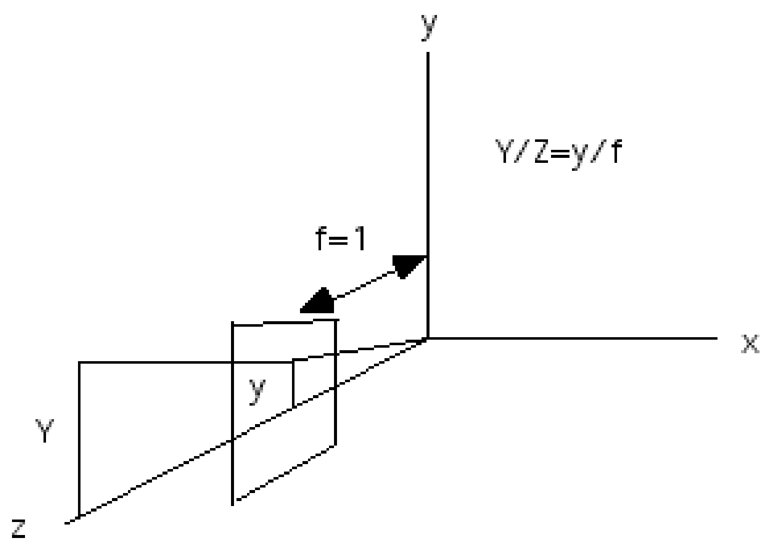
$$\frac{\Delta Y}{\Delta t} = -\frac{\Delta y}{\Delta t} - \frac{\Delta \theta_z}{\Delta t} X + \frac{\Delta \theta_x}{\Delta t} Z = -V - CX + AZ$$

$$\frac{\Delta Z}{\Delta t} = -\frac{\Delta z}{\Delta t} - \frac{\Delta \theta_x}{\Delta t} Y + \frac{\Delta \theta_z}{\Delta t} X = -W - AY + BX$$

(NOTE: Here the lower case Δx, Δy, Δz represent changes in the 3D world coordinates {X,Y,Z} due to the small translation, and should have 0 subscripts (as above) to distinguish them from the x and y used below. Below we use {x,y} to represent the *projection of* {X,Y,Z}. I hope to fix this notation in the future.)

So far so good. We have described the velocity of P in world coordinates in terms of the rotational and translational velocity components of the moving coordinate system. What is happening in the image--i.e. to the motion field or optic flow?

■ **Next step: Relate P velocity and depth Z to the motion field, {u,z}**



(Another note: light travels in straight lines, and so should the projected line in the figure above!)

For convenience, we assume the focal length to be one. The perspective projection is:

$$(x, y) = \left(\frac{X}{Z}, \frac{Y}{Z}\right)$$

and the motion field in terms of Z, and the rates of change of X,Y, and Z are:

$$u = \frac{dx}{dt} = \dot{x} = \frac{\dot{X}}{Z} - \frac{X\dot{Z}}{Z^2}$$

$$v = \frac{dy}{dt} = \dot{y} = \frac{\dot{Y}}{Z} - \frac{Y\dot{Z}}{Z^2}$$

To simplify notation, we've used the "dot" convention to indicate the temporal derivatives of X,Y, and Z.

■ Main result for goal 1, the generative model:

Substituting the expressions for the rate of change of X,Y, and Z, we have:

$$u = (\frac{-U + Wx}{Z}) + (-B + Cy + Axy - Bx^2)$$

$$v = (\frac{-V + Wy}{Z}) + (-Cx + A + Ay^2 - Bxy)$$

Note that we have organized the terms on the right of each equation so that the first parts do not depend on A,B, or C--that is

$$u_T = (\frac{-U + Wx}{Z})$$

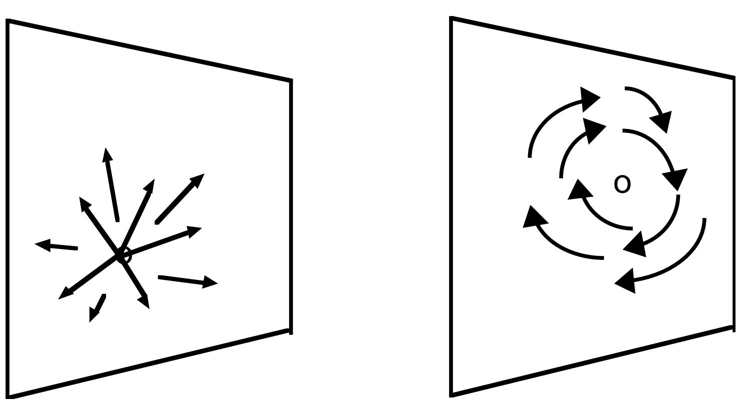
is a purely translational component, and the second term in brackets is purely rotational, and further does not depend on Z:

$$u_R = (-B + Cy + Axy - Bx^2)$$

So in general, we can write:

$$u = u_T + u_R \qquad v = v_T + v_R$$

The figure on the left below shows the flow field one would expect from a purely translational motion--there is a center of expansion (which could be off a finite image plane). The right panel shows the flow pattern of a rotational field.



Diversion: Using *Mathematica* to derive structure from motion and heading equations

Here is a start. I'll leave it to the reader to prove the rest of Longuet-Higgins and Pradzny's results using *Mathematica* to manipulate homogeneous coordinates.

```
XRotationMatrix[θ]//MatrixForm
```

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \text{Cos}[\theta] & -\text{Sin}[\theta] & 0 \\ 0 & \text{Sin}[\theta] & \text{Cos}[\theta] & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Recall the **Series[]** function:

```
??Series
```

```
Series[f, {x, x0, n}] generates a power series expansion for f about the
  point x = x0 to order (x - x0)^n. Series[f, {x, x0, nx}, {y, y0,
  ny}] successively finds series expansions with respect to y, then x.

Attributes[Series] = {Protected, ReadProtected}

Options[Series] = Analytic → True
```

Expand the rotation matrix into a Taylor series:

```
Series[XRotationMatrix[θ],{θ,0,1}]//MatrixForm
```

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1+O[\theta]^2 & -\theta+O[\theta]^2 & 0 \\ 0 & \theta+O[\theta]^2 & 1+O[\theta]^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Use Normal[] to chop off higher order terms:

```
Normal[Series[XRotationMatrix[θ],{θ,0,1}]]//MatrixForm
```

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -\theta & 0 \\ 0 & \theta & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Translational matrix is:

```
TranslateMatrix[-x0,-y0,-z0]//MatrixForm
```

$$\begin{pmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & y \\ 0 & 0 & 1 & z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Now let's put all the rotation and translational components together:

```
Normal[Series[TranslateMatrix[x,y,z].XRotationMatrix[θx],{θx,0,1}].Series[
YRotationMatrix[θy],{θy,0,1}].Series[ZRotationMatrix[θz],{θz,0,1}]]-Identi
tyMatrix[4]//MatrixForm
```

$$\begin{pmatrix} 0 & -\theta z & \theta y & x \\ \theta x\, \theta y + \theta z & -\theta x\, \theta y\, \theta z & -\theta x & y \\ -\theta y + \theta x\, \theta z & \theta x + \theta y\, \theta z & 0 & z \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{0} & \mathbf{-\theta z} & \mathbf{\theta y} & \mathbf{x} \\ \mathbf{\theta z} & \mathbf{0} & \mathbf{-\theta x} & \mathbf{y} \\ \mathbf{-\theta y} & \mathbf{\theta x} & \mathbf{0} & \mathbf{z} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Exercise: Use *Mathematica* symbol manipulation to derive

$$u = \left(\frac{-U + Wx}{Z}\right) + \left(-B + Cy + Axy - Bx^2\right)$$

Goal 2: Inference model: given (u,v), how can we obtain estimates of A,B,C,U,V,W,Z?

In general we can't obtain all seven unknowns (see Horn's book, chap. 17). One problem is that scaling Z by a constant, can be exactly compensated for by a reciprocal scaling of (U,V,W) yielding an unchanged motion field. Horn discusses least squares solutions for the direction of camera motion, and for its rotational component. See also Heeger and Jepson (1990).

Although one can imagine using Bayesian methods for the insufficiently constrained problem of estimating any or all of the seven unknowns, let's see how far one can get with simple algebra to get movement direction (not speed) and relative depth (not absolute depth). We follow the original work of Longuet-Higgins et al. (1980) for estimating the camera direction, and relative depth.

■ Known rotation, estimate translation.

First, suppose we know the rotational component. Then measurements of the motion field will give us the translational components. These components constrain U,V,W, and Z at each point (x,y) in the conjugate image plane.

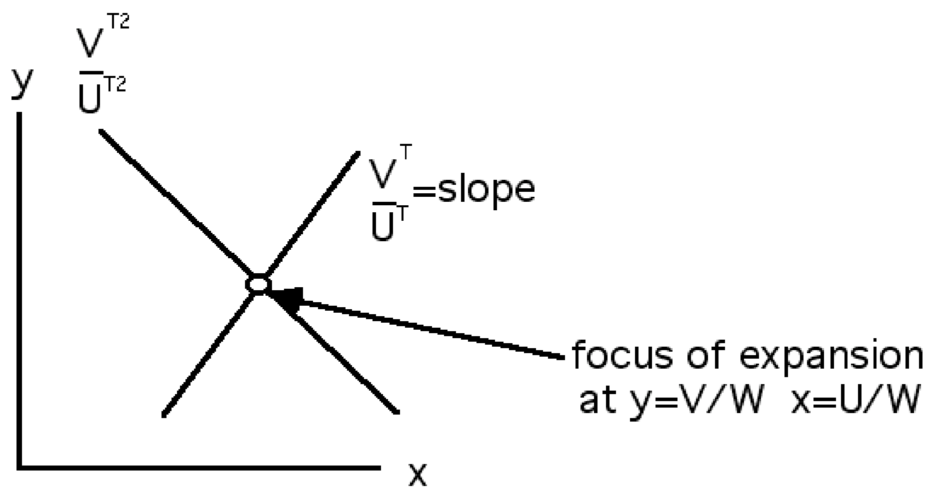
$$u^T = \frac{-U + xW}{Z}$$

$$v^T = \frac{-V + yW}{Z}$$

Combining these two equations to eliminate Z:

$$(y - \frac{V}{W}) = (x - \frac{U}{W})\frac{v^T}{u^T}$$

This equation is a straight line whose slope is determined by the ratio of the vertical and horizontal components of the flow field, and which passes through the point (V/W, U/W). This point depends only on the camera's translational velocity, so other motion flow field lines with different ratios of vertical and horizontal components of the flow also pass through this point. The point (V/W, U/W) is the *focus of expansion*.



Two motion field lines determine the focus of expansion, and thus the camera's translational direction, whose cosine is:

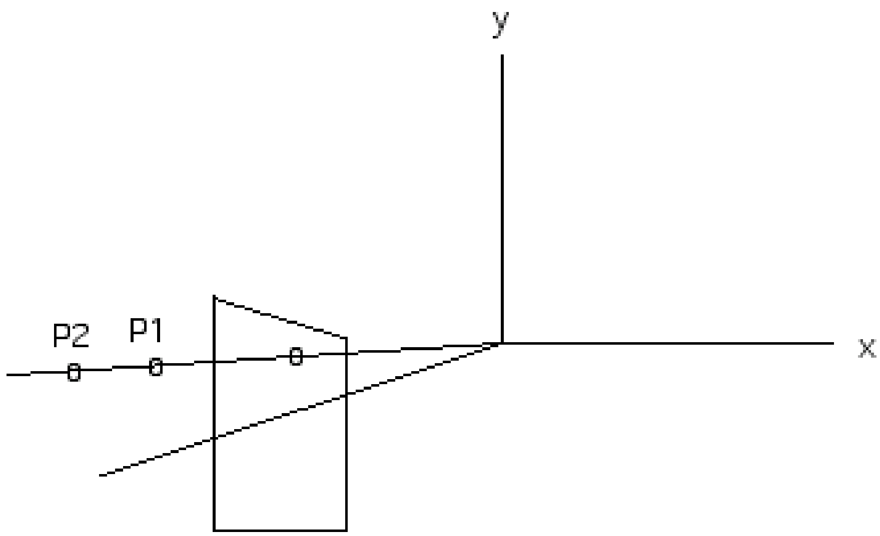
$$\frac{(U, V, W)}{\sqrt{U^2 + V^2 + W^2}} = \frac{(U / W, V / W, 1)}{\sqrt{U^2 / W^2 + V^2 / W^2 + 1}}$$

We can also obtain an estimate of the relative depth of points:

$$\frac{Z}{W} = \frac{x - U/W}{u^T} = \frac{y - V/W}{v^T}$$

■ Unknown rotaion: Estimate both rotation and translational components.

What if we don't know the rotational component? One solution suggested by Longuet-Higgins and Prazdny is to make use of *motion parallax,* where we have two 3D points that project to the same conjugate image point.



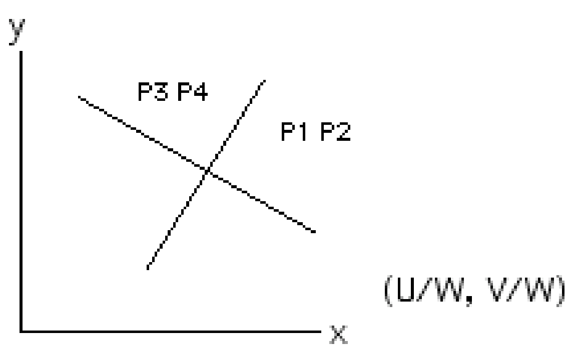
In general, these two points will have different motion field vectors at this image point. If we take the difference, we have:

$$u_1 - u_2 = (-U + xW)\left(\frac{1}{Z_1} - \frac{1}{Z_2}\right)$$

$$v_1 - v_2 = (-V + yW)\left(\frac{1}{Z_1} - \frac{1}{Z_2}\right)$$

$$\left(y - \frac{V}{W}\right) = \left(\frac{v_1 - v_2}{u_1 - u_2}\right)\left(x - \frac{U}{W}\right)$$

Again, finding the focus of expansion (xo,yo), which involves finding at least two motion parallax pairs,



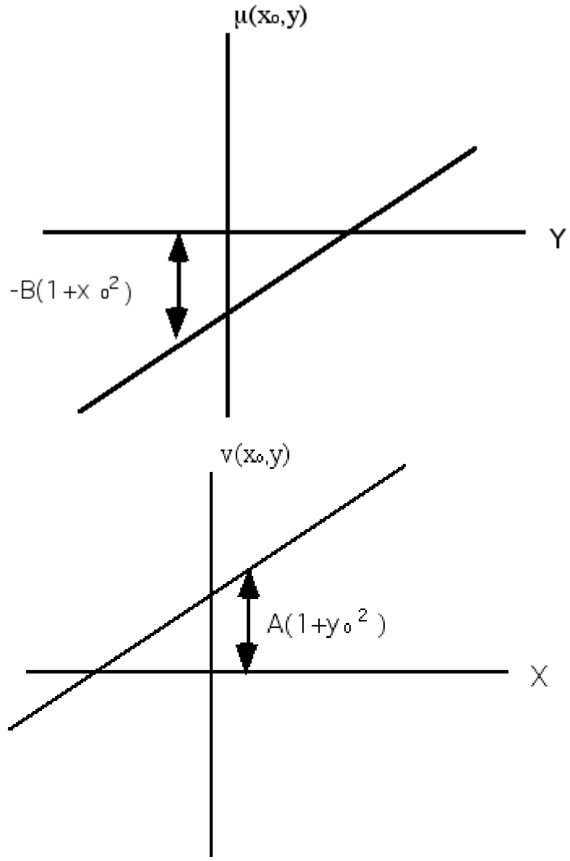
will give us the camera (or eye) direction

$$x_0 = \frac{U}{W}$$

$$y_0 = \frac{V}{W}$$

To find relative depth, we need to know A,B,C:

$$u(x_0, y) = (C + A x_0) y - B(1 + x_0^2)$$
$$v(x, y_0) = -(C + B y_0) x + A(1 + y_0^2)$$



These relations provide sufficient information to calculate A,B,C (from two or more points). A,B,C in turn determine u^R and v^R.

$$u = (x - x_0)\frac{W}{Z} + u^R;\ v = (y - y_0)\frac{W}{Z} + v^R$$

With some rearrangement, we can obtain a formula for relative depth:

$$u - u^R = (x - x_0)\frac{W}{Z}$$
$$v - v^R = (y - y_0)\frac{W}{Z}$$
$$\frac{Z}{W} = \frac{x - x_0}{u - u^R} = \frac{y - y_0}{v - v^R}$$

Although we won't take the time to go over the results, a potentially important form of information for relative depth, camera motion, and time-to-contact comes from an analysis of the flow patterns generated by textured surfaces (Koenderink, J. J., & van Doorn, A. J., 1976) and the above cited article by Longuet-Higgins and Prazdny. The idea is to compute estimates of the rotation, dilation, and shear of the motion field.

Exercise: Time-to-contact

Problem: Show that the reciprocal of the temporal rate of expansion of an object heading directly towards you is equal to the time to contact. (Lee and Reddish, 1981).

Heading experiments

Structure from motion: Psychophysics

Warren and Hannon (1988) provided the first compelling evidence that the human visual system could compensate for eye rotation purely from optical information. See also Royden, Banks & Crowell (1992) for the possible role of proprioceptive information in heading computation.

Structure from motion: Physiology

A possible neurophysiological basis for derivative measurements of flow (e.g. rotation, dilation, shear), see: (Saito, H.-A., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., & Iwai, E., 1986). For work relating to eye movement compensation in optic flow and heading, See Bradley et al. (1996). See Duffy (2000) for recent work.

Challenges to computational theories of structure from motion

Depth between objects

■ Depth from shadows

http://vision.psych.umn.edu/www/kersten-lab/demos/shadows.html

Depth from viewer

■ More on cue integration: Shadow displacement & size change for depth

Frame of reference issues in cue integration.

Modularity for cue integration: Shadow displacement & size change for depth

Frame of reference issues in cue integration (Schrater, & Kersten, 2000).

Earlier we looked at a simple graph for cue integration and showed how a optimal estimate (for the Gaussian case), say for depth, was a weighted combination of the estimates for the individual cues. The weights were determined from the relative reliabilities of the cues.

But a close examination of the generative models that result in multiple cues can show a more complex set of dependencies.

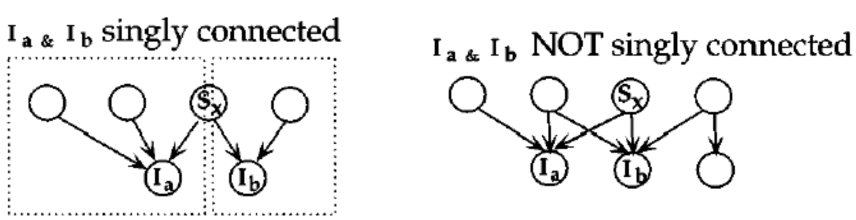
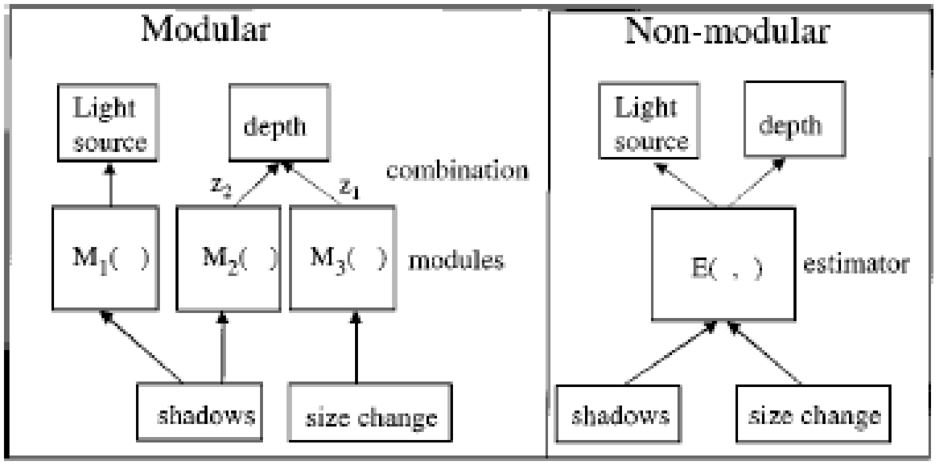
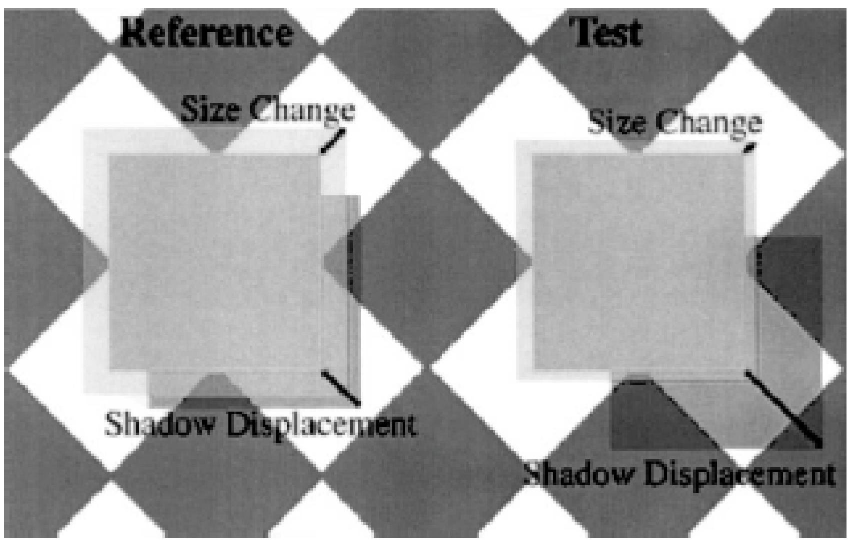


Figure 3. Whether independent data measures are singly connected to the estimated variable S_x determines whether or not estimation modules can be created for S_x. *Left* example of Bayesian modularity. Boxes show how the variables can be split to form two modules. *Right* example of a non-modular estimation.

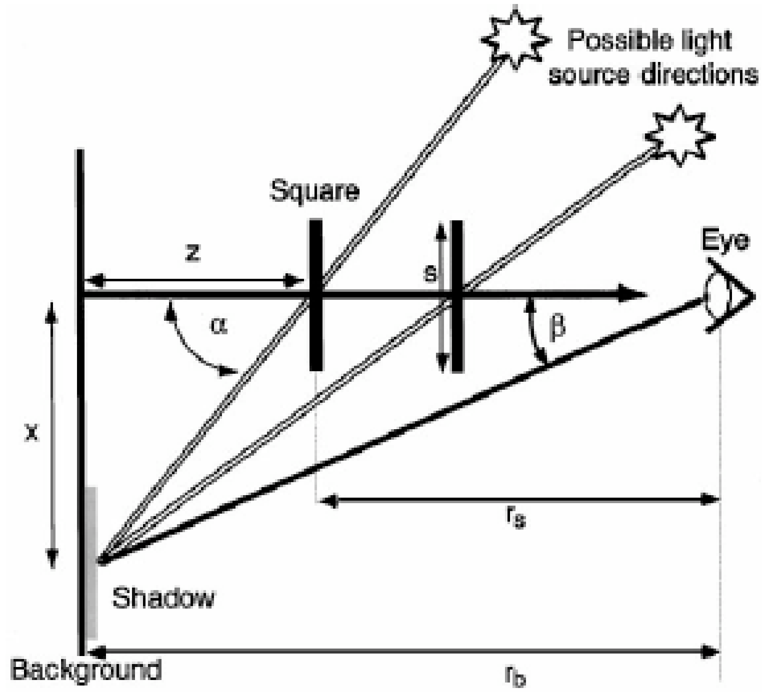
This has an impact on the architecture for optimal inverse inference--whether the algorithm can be broken into distinct modules or not. The non-modular case below is an example of what Clark and Yuille called "strong fusion". This is related to the notion of "cooperative computation" discussed earlier in Lecture 23 on perceptual integration.



Let's take a look at a specific case involving size and shadow position as cues for an object's 3D position in space.



The figure below shows some of the relationships between the data (shadow position β, size of the target square is **a**--not shown), and unknown parameters to be estimated (**z**, **rs**) of interest, (the unit-less parameter, **z/rb** is not shown), and unknowns to be integrated out (α, **s**, **rb**--depending on the task).



When we perceive a change in depth, *what variable does perceived depth correspond to?* Here are three possibilities: relative (unit-less) distance zr/rb, depth from the observer, rs, and distance from the background z.

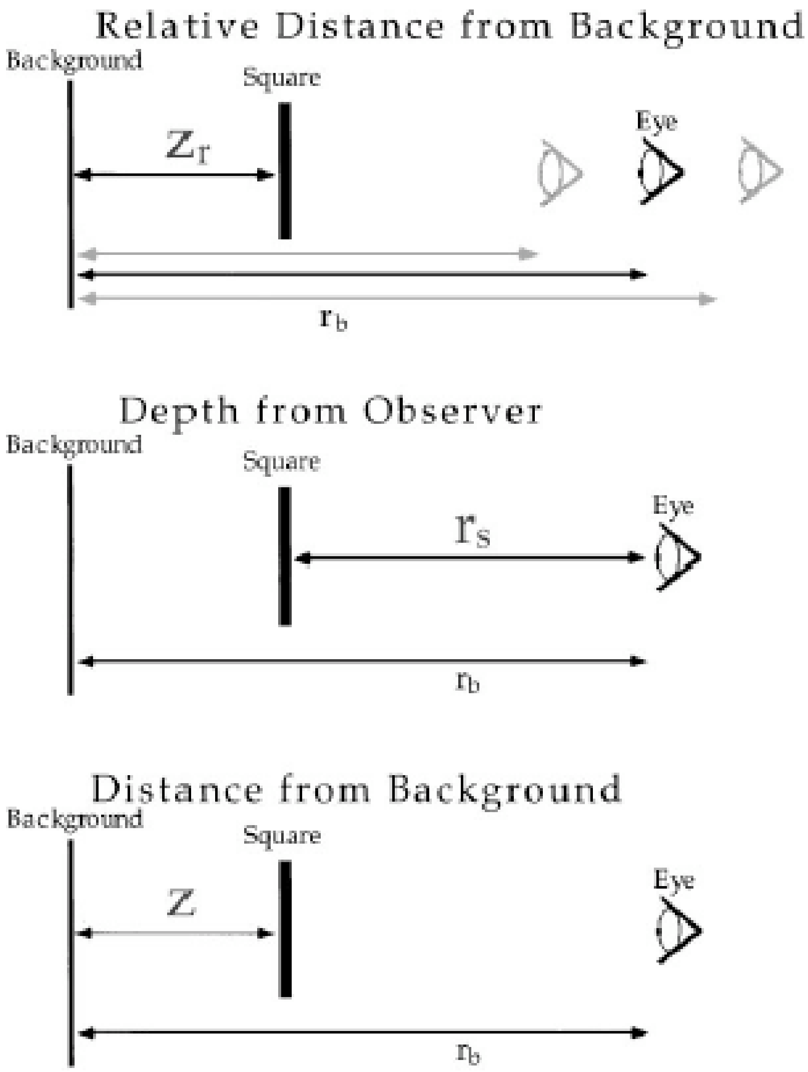


Figure 6. Diagram illustrating the depth variables to be estimated. The variable $z_r = z/r_b$ can't be shown directly, because it is an equivalence class of z and r_b distances.

Paul Schrater worked through the math and showed that these different assumptions about depth representation produced different generative models for producing the image size **a**, and shadow position, $\boldsymbol{\beta}$.

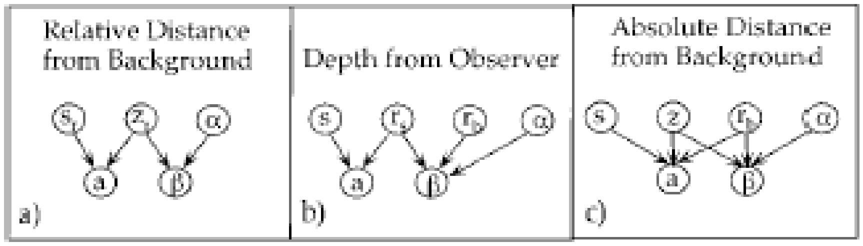


Figure 7. Bayes nets for the three depth representations. a) Bayes net for relative distance to the background. This task involves estimating object relations (world centered), and requires the least prior knowledge. b) Bayes net for distance to observer. Notice that the use of the shadow information requires integrating across two variables, hence the shadow cue should have more uncertainty for this task. c) Bayes net for metric depth from the background. Estimating the distance from the background, z, is complicated by the image size and shadow position measurements also being jointly dependent on the observer's distance to the background.

Fisher information is the asymptotic variance of the estimator, so can be used to calculate a weighted linear combination (an optimal estimator for the modular case).

Table 1. Table of MAP estimates and Fisher information values for the three depth estimate r tations. For the representations which admit modular estimates, the estimates are shown separ the shadow and image size cues.

Task	Est from shadow	Est from size	Shadow Fisher info	Size Fish
Relative z	$z_r = \tan(\hat{\beta})$	$z_r = 1 - \frac{\mu_{s_r}}{\hat{a}}$	$\frac{1}{\sqrt{2}\tan(\hat{\beta})^2}$	$\frac{2\hat{a}^2}{\mu_{s_r}^2(\sigma_{s_r}^2 +}$
Dist. from obs.	$r_s = \mu_{r_b}(1 - \tan(\beta))$	$r_s = \frac{\mu_s}{\hat{a}}$	$\frac{1}{\mu_{r_b}^2 \tan(\hat{\beta})^2}$	$\frac{2\hat{a}^2}{\mu_s^2(\sigma_s^2 +}$
Absolute z	$z = \frac{\mu_s \tan(\hat{\beta})}{\hat{a}(1 - \tan(\hat{\beta}))}$		$\frac{2\hat{a}^2(1 - \tan(\hat{\beta}))^4}{\mu_s^2 \tan(\hat{\beta})^2}$	

The shadow cue is most reliable when the target object is close to the background. But the size cue is most reliable when the target is close to the viewer.

There have been no systematic experimental studies of this general theoretical prediction.

■ Bottom line

Optimal estimators for depth depend critically on the representation of depth

Different representations result in different generative models, and thus different modular structures for optimal inference

Human judgments of closeness may be better predicted by a model that represents depth from the observer, rather than relative depth from the background, in either absolute (e.g. metric) units, or relative units. More experimental work is needed.

Appendix

References

Bradley, D. C., Maxwell, M., Andersen, R. A., Banks, M. S., & Shenoy, K. V. (1996). Mechanisms of heading perception in primate visual cortex [see comments]. Science, 273(5281), 1544-7.

Crowell, J. A., & Banks, M. S. (1996). Ideal observer for heading judgments. Vision Res, 36(3), 471-90.

d'Avossa, G., & Kersten, D. (1996). Evidence in human subjects for independent coding of azimuth and elevation for direction of heading from optic flow. Vision Research, 36(18), 2915-2924.

Duffy, C. J. (2000). Optic flow analysis for self-movement perception. *Int Rev Neurobiol, 44*, 199-218.

Gibson. (1957). Optical motions and transformations as stimuli for visual perception. Psychological Review, 64, 288-295.

Heeger, D. J., & Jepson, A. (1990). Visual perception of three-dimensional motion. Neural Computation, 2(2), 129-137.

Koenderink, J. J., & van Doorn, A. J. (1976). Local Structure of Movement Parallax of the Plane. J. Opt. Soc. Am., 66, 717-723.

Lappe, M., & Rauschecker, J. P. (1993). A neural network for the processing of optic flow from ego-motion in man and higher mammals. Neural Computation, 5, 374-391.

Lee, D. N., & Reddish, P. E. (1981). Plummeting gannets: a paradigm of

ecological optics. Nature, 293(5830), 293-294.

Longuet-Higgins, H. C., & Prazdny, K. (1980). The Interpretation of a Moving Retinal Image. Proceedings of the Royal Society of London B, 208, 385-397.

Martin, W. N., & Aggarwal, J. K. (1988). Motion understanding: robot and human vision. Boston: Kluwer Academic Publishers,

Perrone, J. A. (1992). Model for the computation of self-motion in biological systems. Journal of the Optical Society of America, A, 9(2), 177-194.

Royden, C. S., Banks, M. S., & Crowell, J. A. (1992). The perception of heading during eye movements [see comments]. Nature, 360(6404), 583-5.

Saito, H.-A., Yukie, M., Tanaka, K., Hikosaka, K., Fukada, Y., & Iwai, E. (1986). Integration of Direction Signals of Image Motion in the Superior Temporal Sulcus of the Macaque Monkey. The Journal of Neuroscience, 6(1), 145-157.

Schrater PR, Kersten D (2000) How optimal depth cue integration depends on the task. *International Journal of Computer Vision* 40:73-91.

Ullman, S. (1979). The Interpretation of Structure from Motion. Proceedings of the Royal Society London B, 203, 405-426.)

Warren, W. H., & Hannon, D. J. (1988). Direction of Self-Motion is Perceived from Optical Flow. Nature, 336((10)), 162-163.