Chapter 19

Three Dimensions and Two Eyes

Monocular and stereo depth cues; epipolar geometry; projective transformations; correlation; frontal-parallel plane;

19.1 Introduction

The discussion of edge detection involved only brightness changes, and we saw that these were quite amiguous. While brightness changes often occur at object boundaries, they don't ONLY occur there: brightness changes could be due to material changes (think sweater vs pants or a painted checkerboard), or they could be due to shadows or surface creases or ... whatever.

This last thought could provide a way forward: knowledge about surfaces requires knowledge about the third-dimension – depth – and stereo is a way to recover it. So perhaps if we understood stereo we would understand edge detection more deeply.

But there is much more to stereo: maybe it's completely about inferring structure in the third dimension; clearly this can be incredibly useful. For this reason we turn to stereo in this lecture, and begin with some interesting 'social' aspects.

To understand stereo we must re-visit the concept of projection; this time from 3 dimensions to 2 dimensions via projective geometry. Simply put, more distant objects appear smaller than closer ones. As we shall see, however, this *monocular depth cue* may be confusing for stereo. There's a complicated question here: While stereo is not necessary for inferring depth in certain situations, is stereo is sufficent to see depth in all situations?

This raises the question: why, in fact, do we have two eyes? How can we use them together and for what purposes? Our world is 3 dimensional; images are two dimensional. Thus maps from the world to the image are (relatively) easy to understand; these are dimension reducing. But how can we go in the other direction? The simplest guess: use more than one image. Vaguely this is like solving an equation when you're missing a variable; if you have a second (different) equation perhaps they can be solved simultaneously. Somehow we have to determine how to use the new

information provided by the second view. This is our main task in this and the next lecture.

In this Lecture we discuss the 3-dimensional map to the 2-dimensional image and the basics of projective geometry; the question of how to do stereo is begun, with the specification of the CORRESPONDENCE PROBLEM, and is elaborated into the next lecture as well.

This initial attempt lays out some of the basics, but, as we shall see, these simple approaches remain limited in their capability. Some applications are described. But for richer solutions we shall revisit the stereo problem in the advanced class.

19.1.1 Social Interaction and Hand-Eye Coordination

Everyone think that driving requires two eyes: we must judge how far cars are from us

We saw in the template lecture how difficult it is to design a proper 2-dimensional image template for faces. Here is what might seem a a curious – but will turn out to be important – fact: the proper social distance for interaction between individuals is about 1 - 2 m across basically all cultures (need ref. here). Why is this? One possible answer is that "reading faces"; i.e. not only identifying them but interpreting their emotional messages may well require a 3-dimensional description of them (Ekman).

Important for social interaction and reaching etc; See Fig. 19.1. Here's a simple experiment: hold your arms out straight to the sides, close your eyes, and now rotate your arms so that you touch finger tips directly in front of your body. Now, try it with your eyes open. What is the difference?

19.2 Projective Geometry and Camera Models

To understand stereo vision we must understand how the 3-dimensional world projects onto 2-D images; see Fig. 19.2. We are all familiar with the manner in which parallel lines converge to a point in perspective; if we're able to see where these points all fall, then we can see the HORIZON. The structure of such converging lines comprises PROJECTIVE GEOMETRY.

Projective geometry is based on the manner in which structure projects along straight lines – called RAYS. You can think of a ray as a line of infinite length that passes through the origin of a standard 3-d coordinate system; if we have a 3-D coordinate system (x, y, z) and we choose a point (x_0, y_0, z_0), then any collection of points (kx_0, ky_0, kz_0) for any $k \in \Re\backslash 0$, that is, any number k except 0, defines the ray from the origin through that point.

This construction is important because a point in space projects to a point on an image along the ray that connects the space point to the CAMERA CENTER at precisely the location where this ray intersects the image plane.

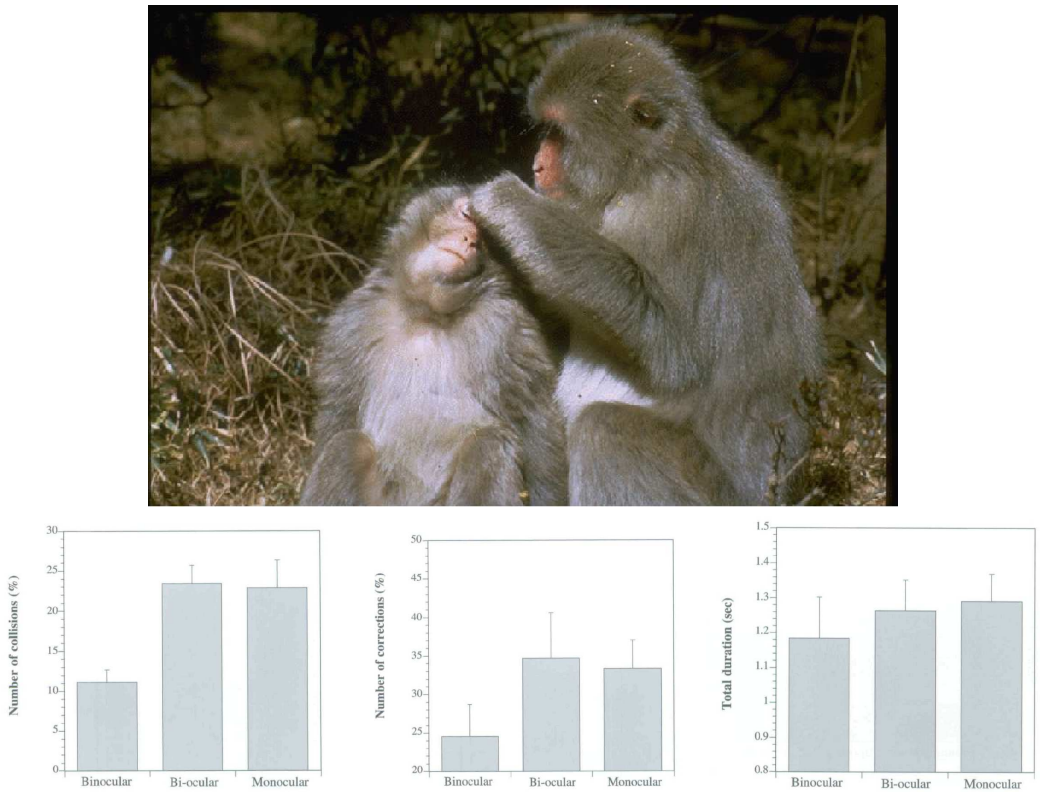


Figure 19.1: Two eyes are better than one (top) for grooming and (bottom) for grasping. Data from Bradshaw et al, 2004.

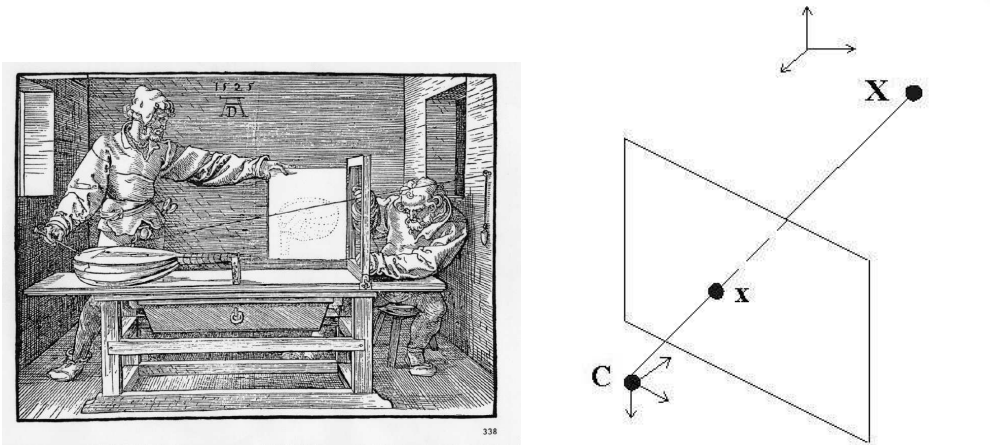


Figure 19.2: Projective geometry underlies stereo vision; it dictates how the 3-D world projects onto a 2-D image. LEFT) Classical etching by Durer (sixteenth century) illustrating the perspective transformation and how it was exploited by Renaissance artists. (RIGHT The geometry of projection for a point in the world.

The projective geometry specifies the ways in which we might form a 2-D image from a 3-D collection of points. To specify it, we shall need a model for an eye (or, equivalently, for a camera); recall Chap. **??**. Conceptually it is easiest to think of PINHOLE CAMERAS (Fig. 19.3.

Pinhole Camera

The pinhole camera is an imaging device in which all rays are forced to pass through a very tiny hole, thereby providing a means of focus.
There is an interesting physical tradeoff here – the smaller the hole, the sharper the image. This is because ideally only one ray of light should pass through. However this is only a mathematical abstraction – the larger the pinhole the more light gets through. Tradeoff for design: how large should the hole be to allow sufficent light for a bright enough image to see while being small enough so the image is still in focus. See Fig. 19.3.
Aristotle noticed pinhole camera effects while looking at an eclipse in the forest.
It is thought that artists during the renaissance may have used pinhole camera rooms to make images that were then traced into paintings; see David Hockney.

A consequence of projective geometry is that more distant objects appear smaller than closer ones, and lines on a flat plane converge to the horizon (consider white lines painted on an airport runway); see Fig. 19.3.

Quote from David Hilbert, Geometry and the Imagination, describing his figure (see Fig. 19.3, 3-rd image labeled (b)):

> If we draw the picture of a flat landscape on the blackboard, the landscape being a horizontal plane and the blackboard a vertical plane, then the image of the horizontal plane appears to be bounded by a straight line h, the horizon. Two parallel straight lines in the horizontal plane which are not parallel to the plane of the blackgoard appear in the picture as straight lines that meet on the HORIZON. In painting, the point of intersection of the two lines in the image is called the VANISHING POINT of the parallel lines.
>
> We see, then, that the images of parallel lines under center perspective are almost always not parallel. We see furthermore that the mapping effected is not one-to-one. The points of the horizon on the image plance do not represent any points of the original plane. Conversely, there are points

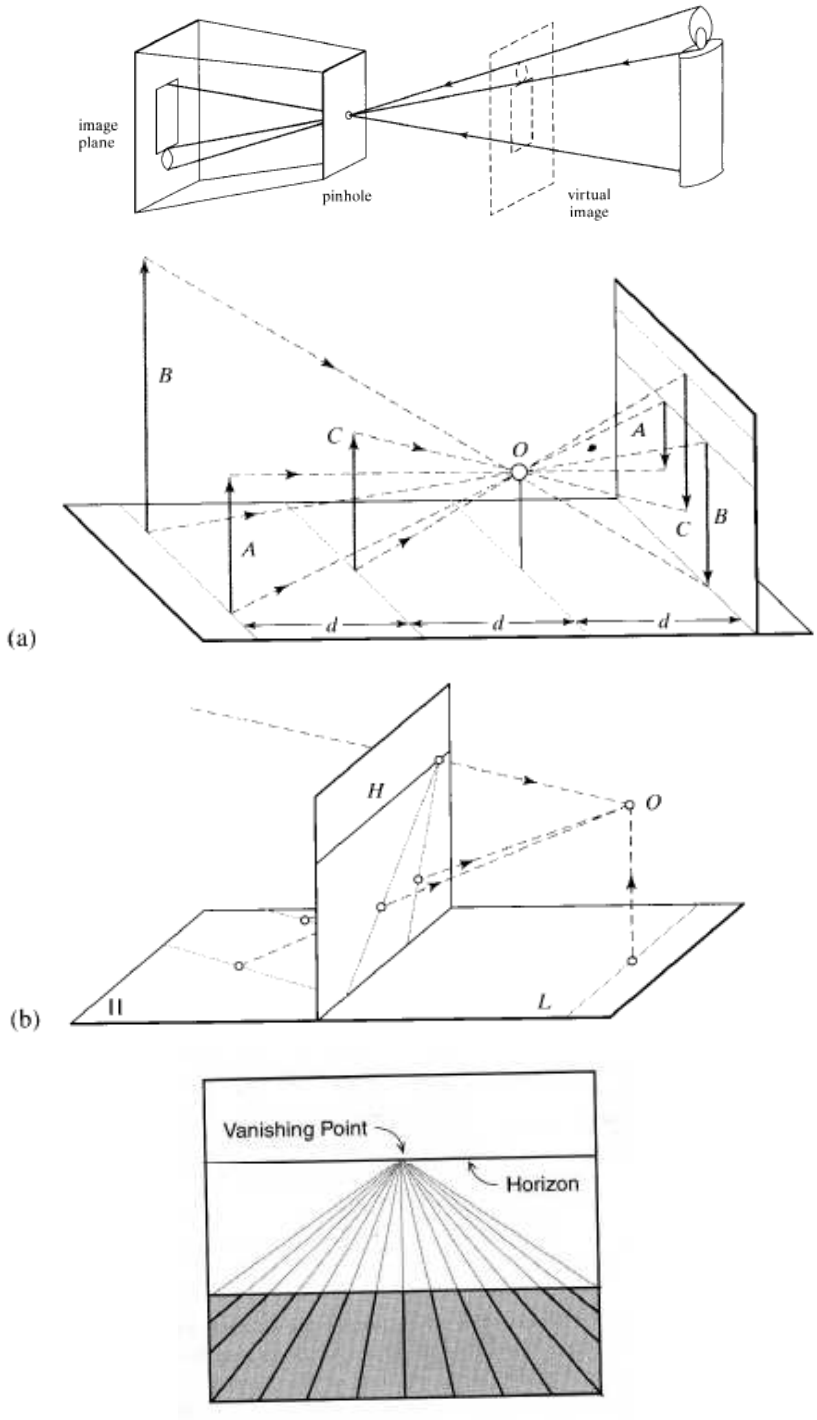


Figure 19.3: Converging light rays lead to perspective effects. (top) Simple camera model in which light rays pass through a pinhole; (this should go into the retina chapter as well.) Note the virtual image formed at a distance in front of the camera, which is sometimes useful for characterizing the geometry. (middle) A consequence of perspective is that further objects project to smaller images than closer ones. (bottom) Images of parallel lines in a (ground) plane converge to a point on the horizon. Images from Forsythe/Ponce. Last image: the vanishing point illustrates how, under projection, many point in the scene can map to a single point in the image (image: Palmer)

of the plane which do not have an image. These are the points of the straight line f that is vertically below the observer R and parallel to the image plane.

A basic model of projection can be derived from Fig. 19.4, top. Since the scene point $P = (x, y, z)$ projects to the image point $P' = (x', y', z')$ along a ray, collinearity implies $OP' = \alpha OP$ or

$$x' = \lambda x \tag{19.1}$$
$$y' = \lambda y \tag{19.2}$$
$$f' = \lambda z \tag{19.3}$$

$$x' = f'\frac{x}{z} \tag{19.4}$$
$$y' = f'\frac{y}{z}. \tag{19.5}$$

For the WEAK-PERSPECTIVE model, in which the points in the world are taken to be far from the camera, in the sense that depth variation (relief) in the world is small compared with distance to the camera, the FRONTO-PARALLEL PLANE becomes important (This is the plane orthogonal to the line-of-sight.) Then we can define a MAGNIFICATION FACTOR $m = -\frac{f'}{z}$ so

$$x' = -mx \tag{19.6}$$
$$y' = -my \tag{19.7}$$

(The minus signs are convention to make magnification a positive number, since the frontal-parallel plane is in front of the camera, z_0 is negative.)

When the change in depth in the scene is small relative to the distance from the camera, m is about constant. Normalizing image coordinates so that $m = -1$, we obtain ORTHOGRAPHIC PROJECTION:

$$x' = x \tag{19.8}$$
$$y' = y. \tag{19.9}$$

This is the simplest way to do projection: just think of the (x, y)-plane as the image and just drop the z-coordinate; this is also called PARALLEL PROJECTION. It is a useful approximation when objects are very very far from the image plane, or the scene is particularly flat.

19.3 Artist's Accurate Projections

Artists have been aware of the importance of perspective since Alberti and the renaissance. There's a lot to discuss here. First, the pinhole camera model may have

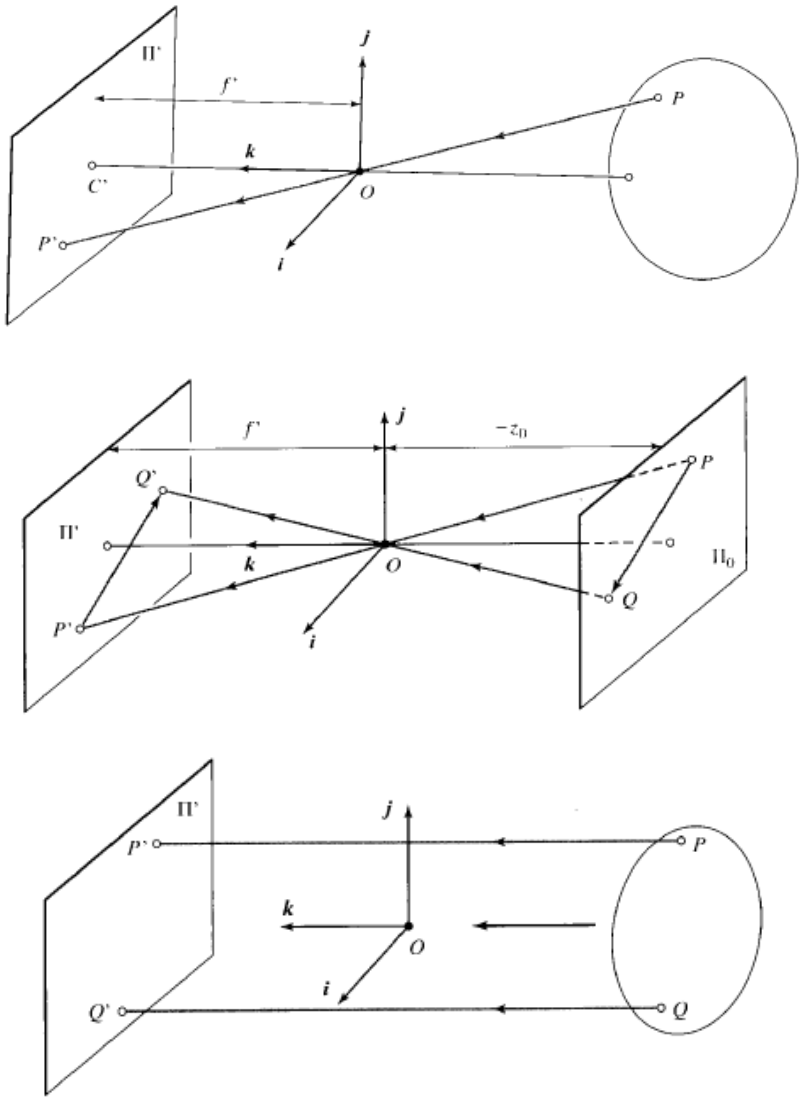


Figure 19.4: Geometry of projection. (top) Consider two points on the surface of an object in the world imaged through a pinhole by establishing a "right-hand" frame (i, j, k) directly at the pinhole (O). Direction k is the line-of-sight. (middle) The weak-projection model, in which the camera is at an approximately constant distance from the scene points. This introduces the fronto-parallel plane Π_0, on which scene structure lives. (bottom) The orthographic-projection model for objects very very far from the imaging device. Figures from Forsythe/Ponce.

been used by artists to get a very accurate model of perspective. This is now being widely investigated by David Hockney and others to determine whether/how they did it; See Fig. 19.5.

Such techniques are extremely important in architectural and machinist's drawings today.

19.4 The Reconstruction Problem

Given a single 2-D image it is of course impossible (in general) to reconstruct a 3-D scene accurately, unless additional information is given. This involves the study of MONOCULAR DEPTH CUES, to which we shall turn in later lectures. There are terrific examples of visual illusions that exploit this – see Ames Room demonstration (Fig. 19.6). However, there may be a solution to this reconstruction problem if a second image of the scene is available. This is a problem that has been studied at least since Kepler.

The simplest scene to consider is just a single point in 3-D. The positional difference in coordinates is the SPATIAL DISPARITY. If you view the tip of your finger about a foot in front of your eyes, spatial disparity can be illustrated by closing one eye and then the other: notice how your fingertip seems to "move" to a different position when viewed monocularly through the left or the right eye.

If the eyes are CALIBRATED; that is, if we know their separation and vergence angles, and if we have a proper model of the eyes as image-forming devices, then triangulation can be used to recover the 3-D coordinates of the point from its positions in the two images. This is essentially a problem in geometry that we shall not consider further.

19.4.1 Vergence and Disparity

Two important notions arise in understanding stereo from the geometry of the eyes: vergence and disparity; see Fig. 19.8.

Let capital letters denote a point in space, $(X, Y, Z) \in \Re^3$ and lower case letters (x, y) a point in the image. Suppose the focal length of the eye is f. Then, when expressed in radians, we know from before that

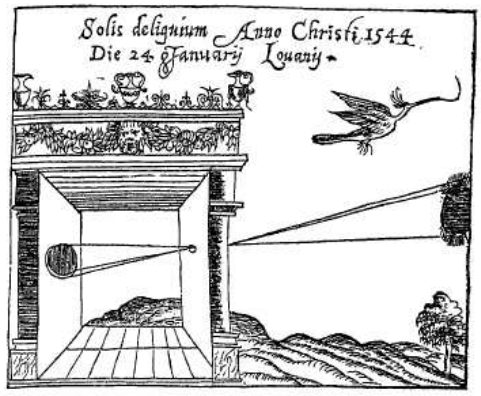
$$\frac{x_L}{f} = \frac{X}{Z} \tag{19.10}$$

and

$$\frac{x_R}{f} = \frac{X - D}{Z} \tag{19.11}$$

where D is the displacement between the eyes. Then

$$\text{disparity} = d = \frac{x_L}{f} - \frac{x_R}{f} = \frac{D}{Z} \tag{19.12}$$



One of the earliest known representations of the pinhole room camera obscura, from Gemma Frisius's *De radio astronomico et geometrico liber*, 1558

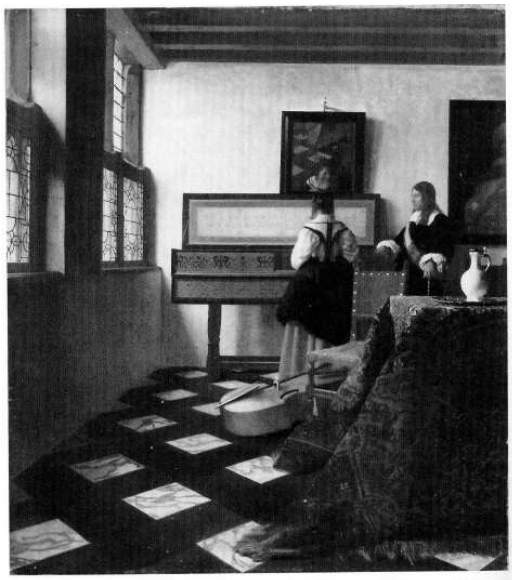


FIG. 2.19. Johannes Vermeer, *The Music Lesson (A Lady at the Virginals)*, c. 1665–1670. Oil on canvas, 73.7 × 64.0 cm. The Royal Collection, Her Majesty Queen Elizabeth II.

Figure 19.5: The camera obscura, or a room that acts like a pinhole camera, is now thought by many to be the foundation for the perspective effects captured by artists. See Kemp, Willats and especially Hockney.

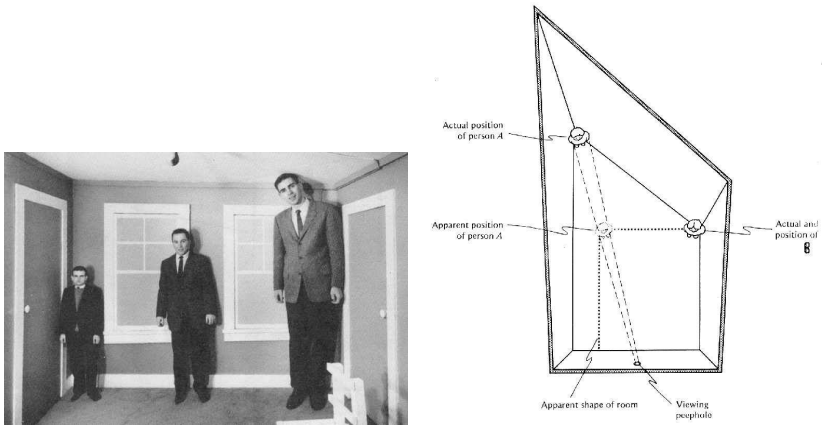


Figure 19.6: The Ames room demonstartion illustrates that we can be improperly fooled by looking at a specially constructed room from a specific vantage point.

or disparity is the offset on the retina of the images of a point in the world. It varies inversely with depth Z.

To characterize this offset, of course, we need to know in which direction the eyes are pointing. If we define the rotation of each eye from "straight ahead", then we compute the vergence angle θ_V.

Finally, the eye's angles can be related to disparity as well. If we imagine them looking straight ahead, then extending a ray from the point in space to the retina defines the angles $\theta_L = x_L/f$ and $\theta_R = x_R/f$. Measuring θ clockwise, we have

$$d = \theta_L - \theta_R. \tag{19.13}$$

The disparity equals the vergence angle when fixating on the point.

What do you think will happen if the eyes cannot verge; if one is pointed toward the point of 'fixation' while the other points outward. When the gaze angle is wrong – is exotropic – we notice it immediately. Stereo vision is impossible. It have been speculated that Rembrandt was exotropic – see his portrait in Fig. 19.9

19.5 Seeing in Stereo

It's useful to get a feeling for convergence by learning how to FREE FUSE a stereo pair of images. We start with an example from Wheatstone, that he used in his famous stereoscope in 1838; see Fig. 19.10. Although he developed an apparatus for presenting stereo pairs, it's helpful to us to learn how to free fuse them (if you're interested). Even if you're not, you might examine the structural differences between them.

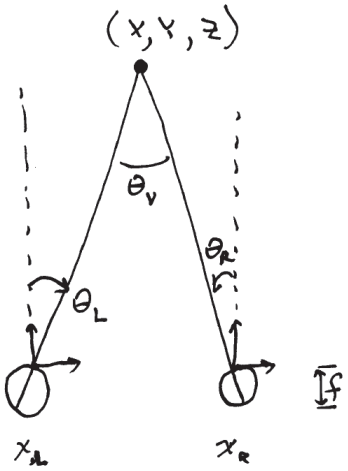


Figure 19.7: Vergence and the eye's geometry for stereo.

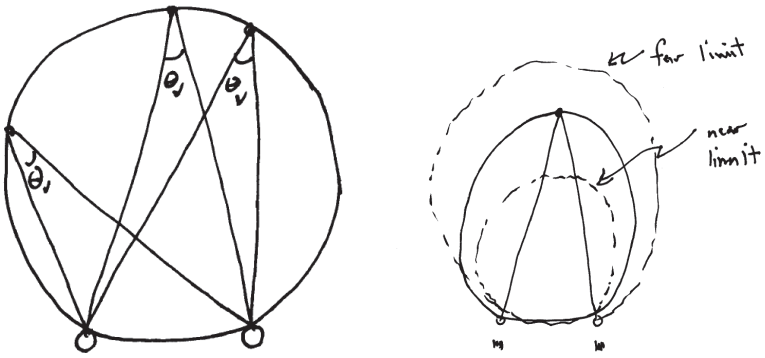


Figure 19.8: Vieth Muller circles are the positions in the world of constant vergence angle at different directions of gaze. Since vergence is related to disparity, there are limits to how far we can fuse images, or compute stereo correspondence; this is known as Panum's Fusional Area.

Figure 19.9: It's sometimes helpful, when learning to draw, to "flatten" the third dimension, e.g. by closing one eye. Livingstone and Conway speculate that artists, including Rembrandt, had the 'advantage' of being stereo blind. Do you think this might explain his incredible drawings? (Frankly, I'm skeptical .. but who knows.)

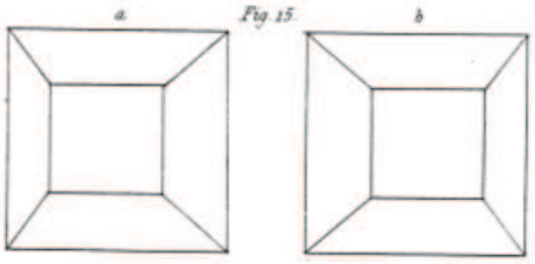


Figure 19.10: Examples from Wheatstone's original paper on the stereoscope. To fuse the images, try the procedure described in the text.

Hold the page/view the screen from about 24 inches. Now hold your finger below the pair of images about half the distance to your nose and fixate it. you should see your finger and, in the background, about four 'copies' of the line drawings. Now slowly move your finger toward the screen but maintain fixation on it – the drawings should start to converge so that, when you see three of them - magic! - the center one should start to look 3D. (This could take a little time and practice, but it's fun. Stick with it a little.)

This is called CROSS-FUSION, because in effect you're crossing your eyes so that the right image is 'seen' by the left eye, and vice versa.

Now, what must be going on in order to fuse these images? We turn to this next.

19.6 The Correspondence Problem

To triangulate a point, we must know where it is in the two images. This was trivial for a scene consisting of a single point. For general scenes the CORRESPONDENCE PROBLEM arises: determine which pairs of points (i.e., pixels in the image) correspond to the same point on the surface of the same object in 3D. See Fig. 19.11.

The correspondence problem amounts to a search problem: pick a point in the left image; search the right image until the match for the left point is found, and repeat until every point in the left image has a match.

Two issues arise. First, what should be matched – individual points, images patches, or features of the image patch? Second, complexity, or the number of computations. If the image is (N x N) pixels, then each point in the left image must be matched against every point in the right image: for a single point in the left image this is N^2 computations! This can be reduced by exploiting the structure of projective geometry.

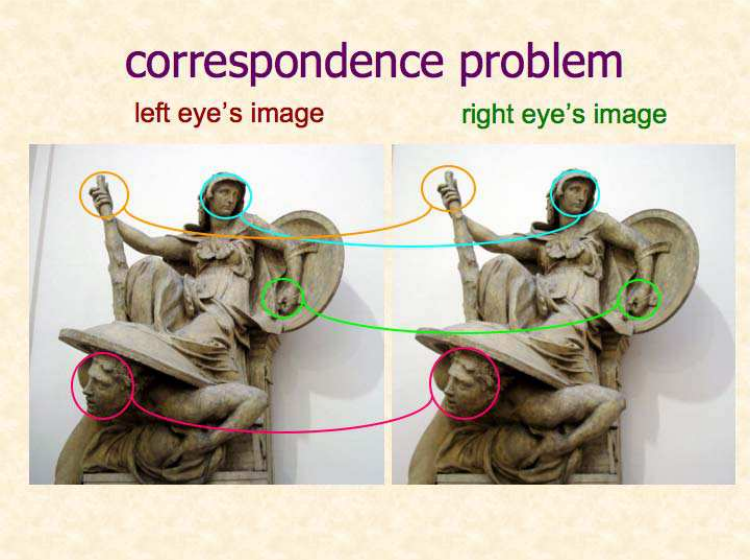


Figure 19.11: The correspondence problem: determine which patch in the left image corresponds to the patch in the right image projecting from the same neighborhood on the object. This must be solved for the patch centered around every point. Disparity then relates to the 'distance' between these patches on the retina.

19.6.1 The Epipolar Line

This full 2-D search can be facilitated enormously by exploiting a fundamental property of projective geometry. Observe that different points in the world may collapse onto the same image point, since any point along the ray in space projects to the same point in the image; see Fig. 19.12.

But also observe that this ray appears as a line in the other image. Again it is instructive to use a finger (or better: a pencil) to illustrate this. Hold the pencil about 4" from your left eye and point it right at the center of your eye; all points along the pencil other than the tip will then lie on the pencil-ray from the center of your eye to infinity. Now, close your left eye and open your right one – you should now be able to see the length of the pencil. This illustrates the EPIPOLAR LINE, or the image of the ray in one camera from another; see Fig. 19.13. The patching point in the right image must lie along the epipolar line, so search can now be done in one dimension.

To make these computations easier, it is often appropriate to rotate the image planes so that they are parallel to the line connecting the cameras; this is the process of RECTIFICATION; see Fig. 19.14.

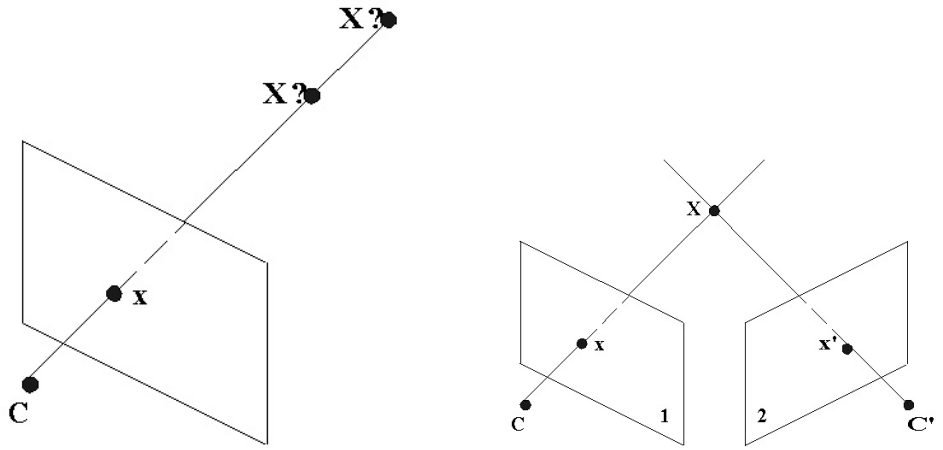


Figure 19.12: Ambiguity in perspective geometry. LEFT Any point along a ray through the camera center projects to the same image point.) (RIGHT) A second view can be used to disambiguate the depth of a point along the ray. The difference in relative coordinates is the spatial disparity of the point in one image relative to the other.

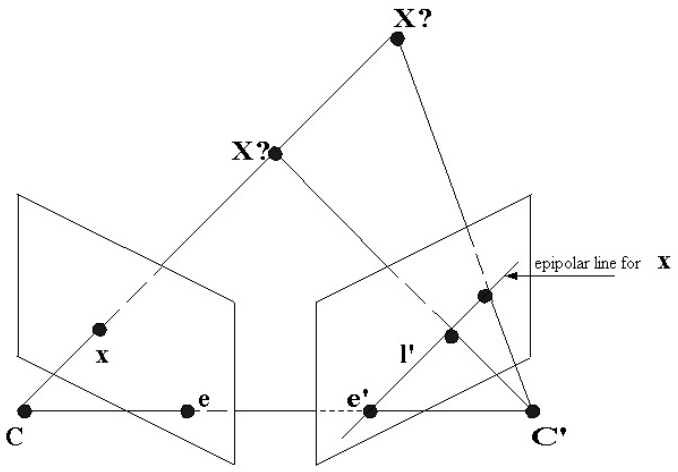


Figure 19.13: The epipolar constraint dictates that the search for a matching point in the left image can be along a 1-dimensional line in the right image, the epipolar line. Determining the epipolar line requires some degree of calibration.

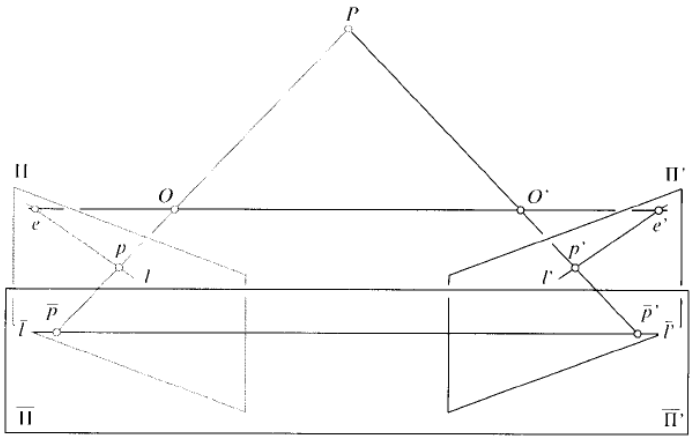


Figure 19.14: The process of image rectification projects the image planes until they are part of a common plane that passes through the camera centers; this makes epipolar lines horizontal and corresponds to placing the projective images at infinity.

19.6.2 Correlation Search Along the Epipolar Line

We are now in a position to state an algorithm for solving the correspondence problem: given a pixel in the left image, search along its epipolar line in the right image until a pixel with the same intensity is found. Clearly this algorithm is too local: many non-matching pixels will tend to have the same intensity (recall the histogram statistics for an earlier lecture).

A more structured search is to define the pattern around the pixel in the left image to include its context: take, e.g., a (5×5) window to define the pattern in the left image and search for the same pattern in the right image. Image correlation (as discussed previously) is the standard technique. See Fig. 19.15.

To put this in concrete terms we work with rectified images. Following Forsythe and Ponce, let $w_L(u, v)$ denote the left-image window centered at (u, v) of size, $p = (2m+1) \times (2n+1)$. One can think of this as a vector of length p, in which each entry of the vector is the intensity at individual pixels:

$$w = \begin{bmatrix} u_{(1,1)} \\ u_{(1,2)} \\ u_{(1,3)} \\ \vdots \\ u_{(2m+1,2n-1)} \\ u_{(2m+1,2n)} \\ u_{(2m+1,2n+1)} \end{bmatrix}. \tag{19.14}$$

We seek to match this against a window of the same size in the right image,

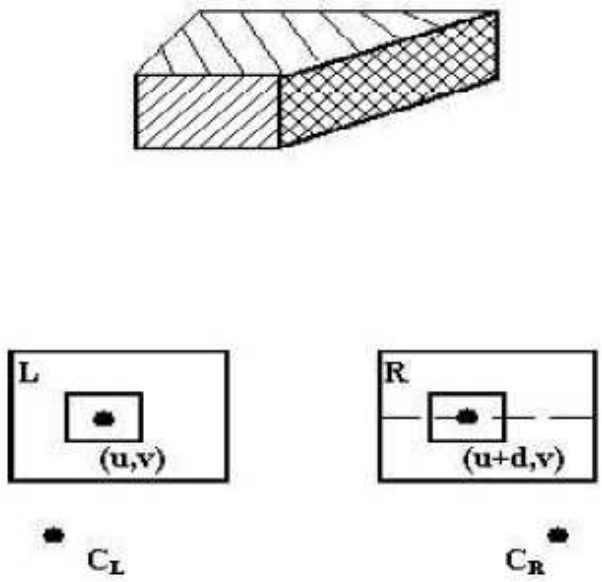


Figure 19.15: SSD matching along the epipolar line for solving the stereo correspondence problem. It amounts to an assumption that image structure lies in the frontal-parallel plane. d denotes spatial disparity.

$w_R(u+d, v)$ to determine whether d pixels is the proper disparity, or shift. (Remember, once the disparity for each matching pair of points in determined, we can get their depth through trigonometry since the cameras are assumed to be calibrated.) We calculate the normalized correlation:

$$C(d) = \frac{1}{w_L - \widehat{w_L}} \frac{1}{w_R - \widehat{w_R}} [(w_L - \widehat{w_L}) \cdot (w_R - \widehat{w_R})]. \tag{19.15}$$

Here $\widehat{w_i}$ denotes a vector of length p each entry of which is the average of the entries in w_i.

19.6.3 Analysis of Correlation-based Matching

A simple analysis of this expression shows that maximizing this normalized correlation function is equivalent to minimizing "SSD", the SUM OF SQUARED DIFFERENCES in pixel values:

$$\left| \frac{1}{w_L - \widehat{w_L}} (w_L - \widehat{w_L}) - \frac{1}{w_R - \widehat{w_R}} (w_R - \widehat{w_R L}) \right|^2 \tag{19.16}$$

This can be more efficent to implement and is widely used (often without the normalization).

More importantly, SSD matching works for a small class of images, but can be a problem for natural ones; see Fig. 19.16. While it is sometimes thought that being able to implement an algorithm is sufficent for testing it, a real advantage of characterizing the computation in abstract, mathematical terms is that we can analyze it to determine those classes of scene structure on which it works.

For maximizing correlation-based stereo or minimizing SSD, this happens when $I_R = \alpha I_L + \beta$ for $\alpha > 0$ and β constants. This can be shown by simply plugging the values in. Such an affine difference occurs mainly when the image is merely shifted in the frontal-parallel plane, and explains the results shown in Fig. 19.16. While this may be sufficent for certain applications, it hardly explains human stereo performance.

There are other heuristic constraints that have been employed to assist in the correspondence problem: prominent among them is the *ordering constraint*. Again, this can be illustrated by holding up your hand (in the plane orthogonal to the line of sight): notice that the order of fingers in the left eye is the same as the order seen in the right eye. However, this is just a heuristic and, although it has been widely used it does not hold in general: see Fig. 19.17.

19.6.4 Geometry of Correlation-based Matching

(Note: You saw this section before.) Continuing our geometric interpretation of various computations, we have that correlation is an inner product; See Fig. 19.18. What does this mean for thinking of "spaces of images" with an inner product, as analagous to spaces of "vectors" with a "dot" product defined between them?

19.7 Disparity in Binocular Neurons

With this background on the computational aspects of the stereo correspondence problem, we now turn to the neurobiology. Recall that the inputs from the two eyes are maintained separately in alternating layers of the LGN; these project monocularly to cortex, layer IVc; see Fig. 19.19.

We shall concentrate on those binocular neurons in the superficial layers that are driven by input from both eyes: that is, the maximal response of the neuron is given when there is a simulus in both the left eye and the right eye, and both are the correct orientation, contrast, and position. (There is a spectrum of binocularity, ranging from purely monocular to purely binocular, but we shall concentrate only on the binocular extreme in this chapter. You can consider the cells two lectures ago to be the monocular ones, although the distinction is only soft.)

The same issues as in computer vision arise in the neurobiology of stereo fusion. Incidently, Ptolemy was the first to draw diagrams (with false matches) like the one in Fig. 19.20.

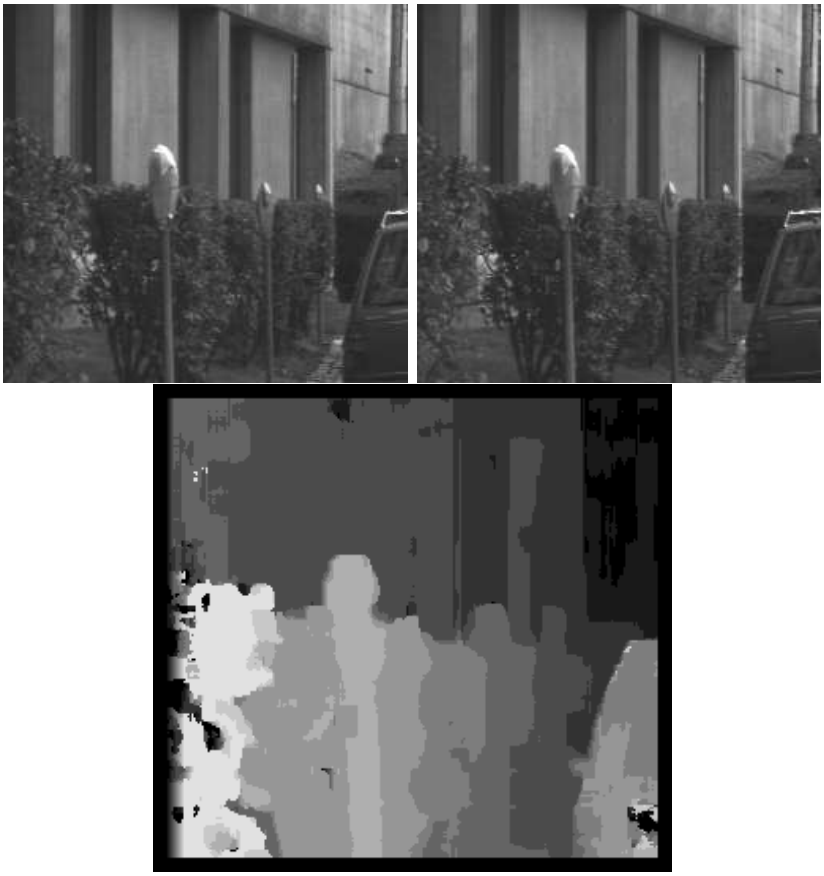


Figure 19.16: SSD matching fails for images that do not lie in the frontal parallel plane. (top) An stereo image pair of a street scene; note that the structure is not very complex but that it does not lie in the frontal-parallel plane. (bottom) The result of the standard solution to the stereo correspondence problem by searching along epipolar lines to find matching pixels. In this case SSD with a neighborhood of 11 pixels was used. Note the "scalloping" of the result into broken frontal-parallel approximations to the true scene. Gray levels correspond to depth in this result image.

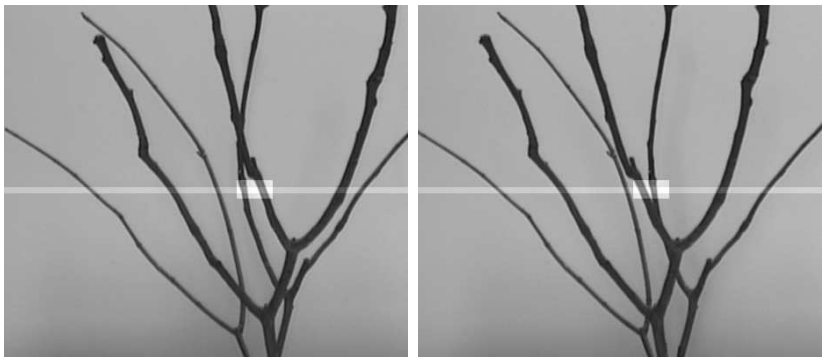


Figure 19.17: The ordering constraint fails for complex scenes. These images have been rectified so the epipolar lines are horizontal as shown. Consider the structure in the highlighted box.

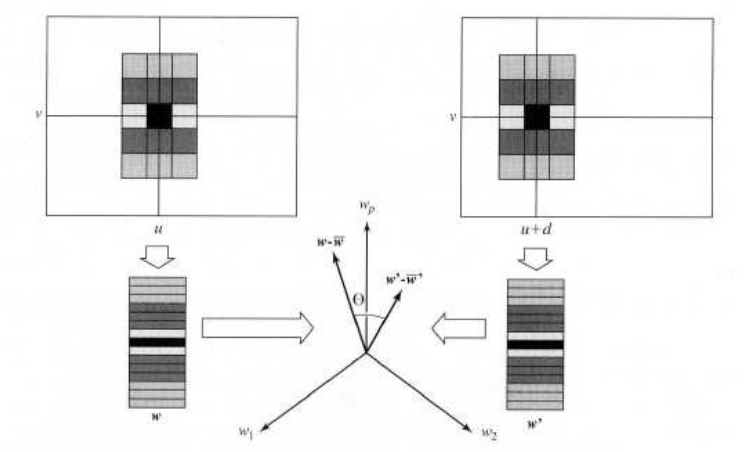


Figure 19.18: The geometry of correlation. If we think of the image patches as vectors, with some standard scan order and length equal to total number of pixels, then normalized correlation is the same as the "inner" product between these vectors. Figure from Ponce/Forsythe.

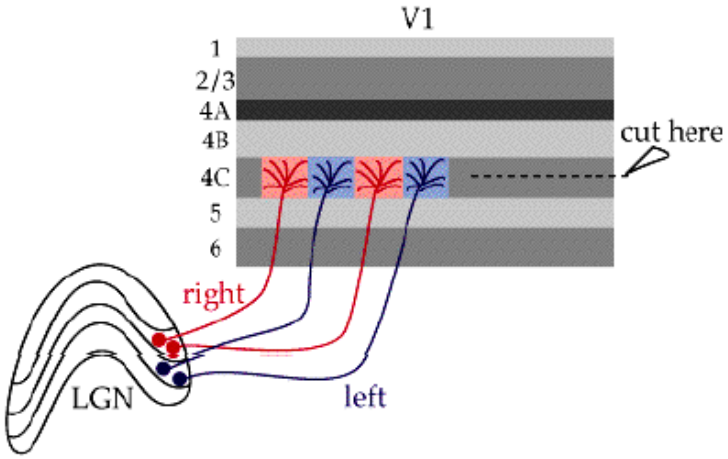


Figure 19.19: lgn-cortex preserves ocular dominance columns.

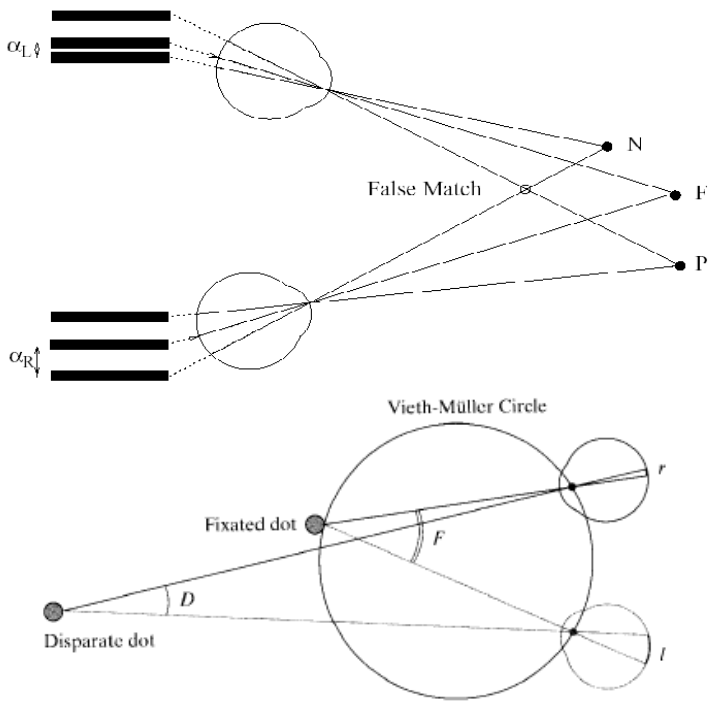


Figure 19.20: (top) Disparity is the angular displacement along the retina of the same point in space; the problem is determining which "point" in the left image corresponds to which point in the right; note both real and illusory matches give disparities. (bottom) The Vieth-Muller circle is that circle of constant disparity with the eyes verged.

19.8 Simple-Cell Combination Models

Remember the Gabor models for simple cells in Sec. 17.2.

If we suppose that stimulus in the left eye is a spot at some depth, then the ideal simulus in the right eye is basically the same stimulus displaced the correct amount from it; such displacements can be checked by wiring cells together with the correct offset so that, when they both fire, there is a proper stereo-disparity signal. (Remember the Reichardt model for motion that we introduced earlier – can you spot the equivalence between two slightly displaced images in space, taken at the same time, and two slightly displaced images in time, taken at the same place?)

More generally than spots, however, as we saw in the previous lecture, one might think of templates small enough to be general purpose while larger than single intensity values. Simple cells (putative "edge-like" and "line-like" signals) would seem to fit the task perfectly and are already available as building blocks.

Finally as we saw in the motion discussion, "coincident firing" of cells would indicate that they were both active. The result, of course, is to somehow interpret this coincident firing as a product of cells and, given the larger effective receptive field, to realize that if several were gathered together it would appear as a complex cell; see Fig. 20.14.

Given the structure of Gabor models for simple cells, there are two ways to accomplish the shift (for disparity); one could offset the receptive fields or one could shift the internal sine; see Fig. 19.21.

Both seem to be the case;

That is, we evaluate the sum (or sum squared!) of even- and odd-symmetric receptive fields at a position in the visual array; this give a response in proportion to match of the energy in the left and right image convolutions. That cell with the maximal firing rate would then signify the correct disparity locally.

> **Remark: we have simply derived a "biological implementation" of a correlation model!**

Such energy models are clearly implementing the frontal-parallel assumption very locally: they seek that image displacement which maximizes energy. But real structure is more complicated than this so we need another idea. If these local estimates are vaguely correct then perhaps we can refine them by looking at nearby estimates.

19.9 Artists' Perspective Revisited

We now return to the question of perspective and how it is used in art. Why are there so many, and how do they work?

A curious one is vertical oblique; See Fig. 19.23. Have you ever seen an example of this in the physical world?

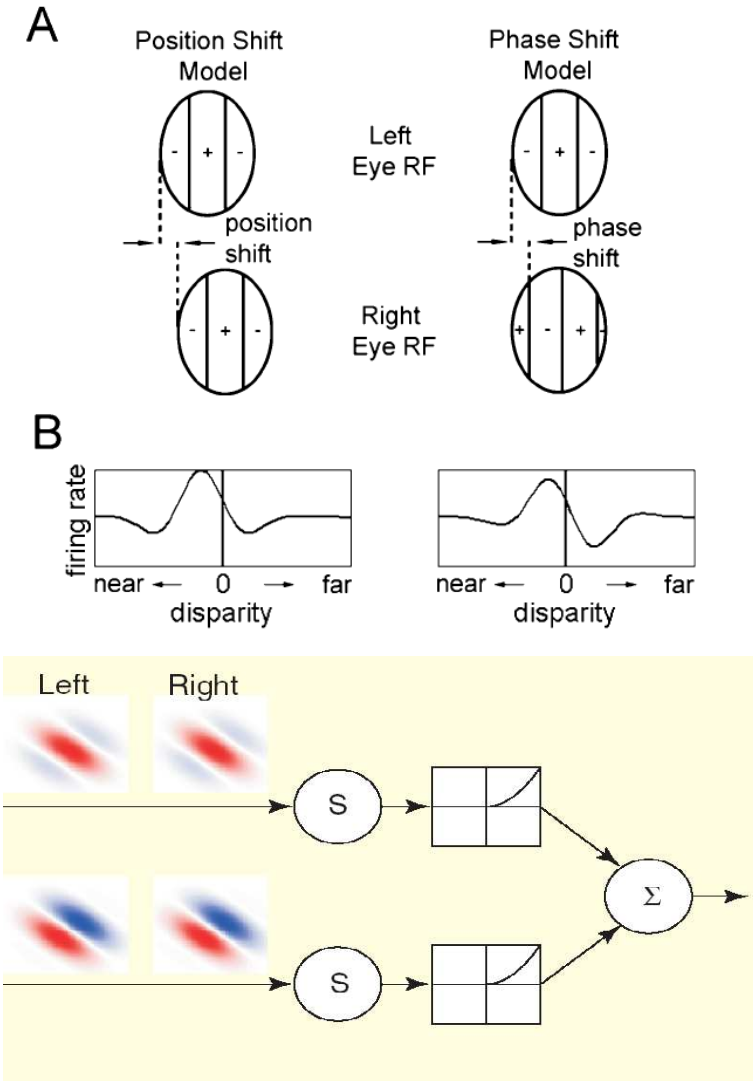


Figure 19.21: Energy models for stereo correspondence. (top) Determining whether a spatial shift or gabor phase shift is the right way to implement the disparity wiring. (bottom) Pooling over complex cells provides additional averaging. Top figure: Qian; Bottom figure: deAngeles

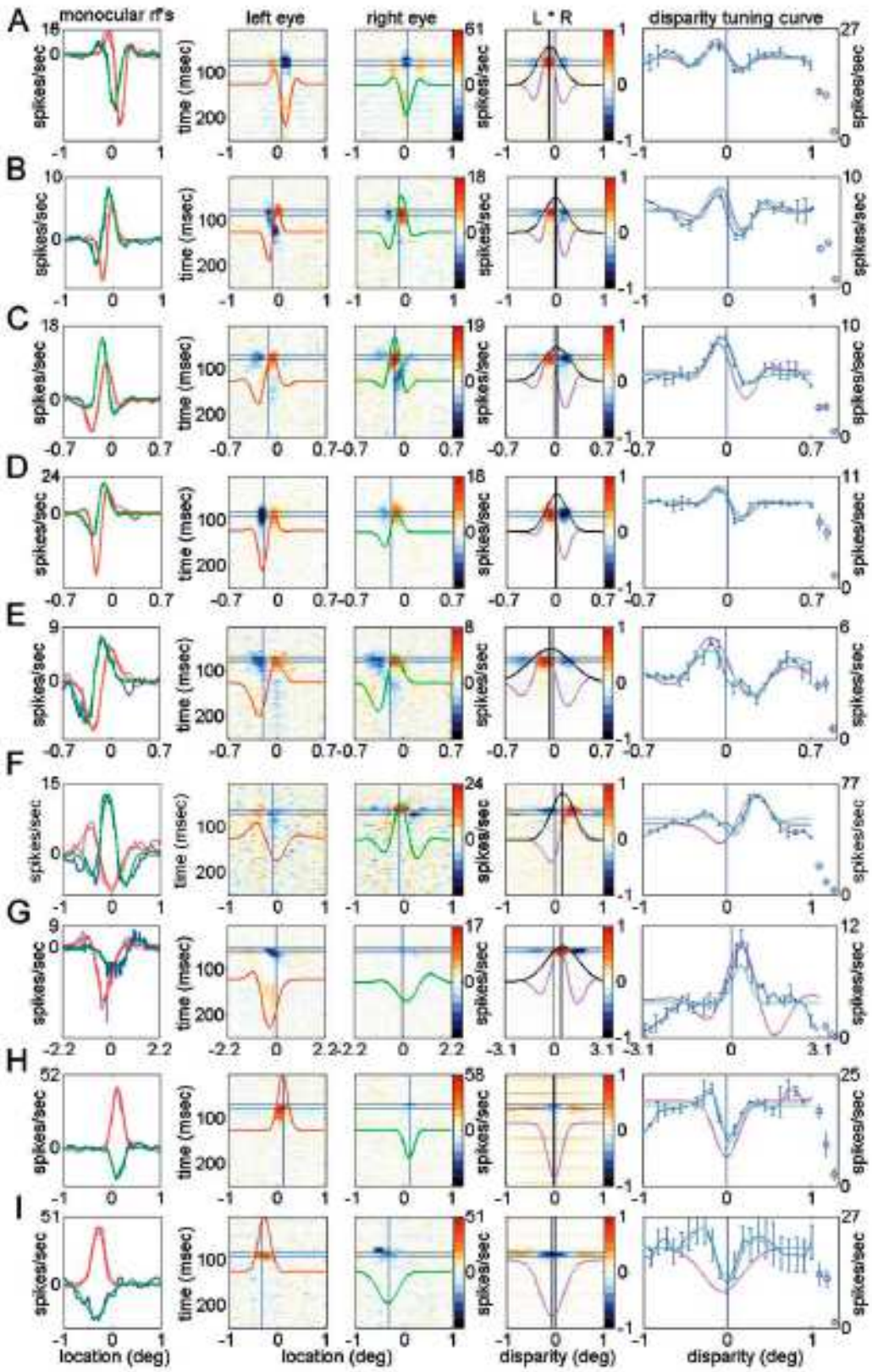


Figure 19.22: Evidence for stereo correspondence using simple stimuli. Data: Disparity-Tuned Simple Cells in Macaque V1 Neuron, Volume 38, Issue 1, Pages 103-114 D.Tsao, B.Conway, M.Livingstone

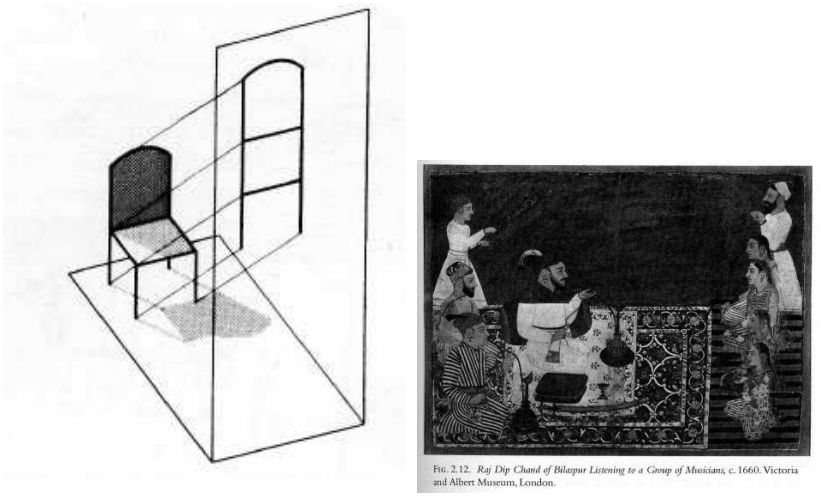


Figure 19.23: Vertical oblique perspective and its use in classical Indian art. Does the figure look correct mathematically? Three-dimensional? Is it interpretable? Figure from Willats.

Perhaps even more curious is the naive—and incredibly *qualitative*—perspective from Cennini in the 15-th century (quote from Willats):

> And put in the buildings by this uniform system: that the moldings which you make at the op of the building should slant downward form the edge next to the roof; the molding in the middle of the building, halfway up the face, must be quite level and even; the molding at the base of the building underneath must slope upward, in the opposite sense to the upper molding, which slants downward.

See Fig. 19.24. How can this possibly work?

Finally, we note that even for careful perspective it may not be consistent; there may be several different "pinholes" from which a single painting has been made, amounting to a kind of perspective collage: See Fig. 19.25. Why does this not bother us? In fact, does one even notice it? (Curiously, it has been centuries before art historians have analyzed these questions formally!)

19.10 Summary

How might we actually compute stereo in a brain? How accurate should it be and under which circumstances (reaching movements under eye control vs. artistic renderings).

Can we do better than this frontal-parallel solution?

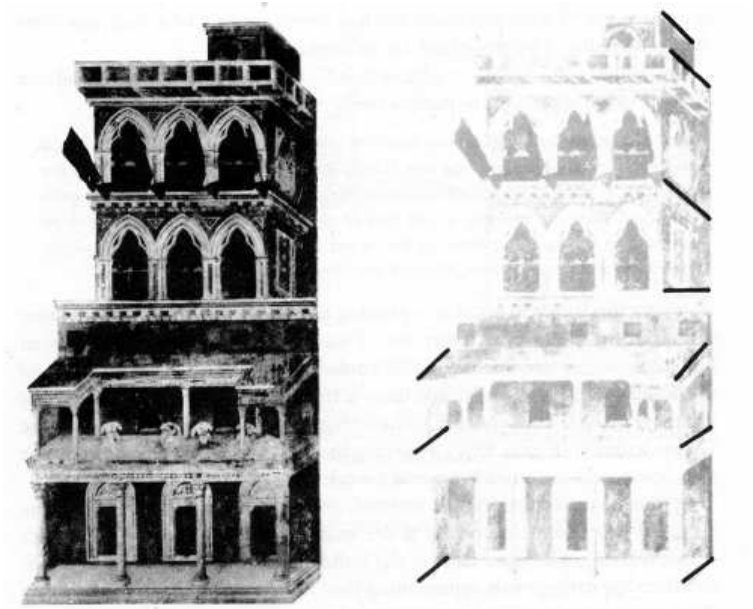


FIG. 2.22. (a) Giotto, *Painting of a Building* (detail of *The Dream of the Palace and Arms*), c. 1297–1300. Assisi, Upper Church of San Francesco. (b) The heavy lines show the directions of the orthogonals.

Figure 19.24: Naive perspective. Figure from Willats.

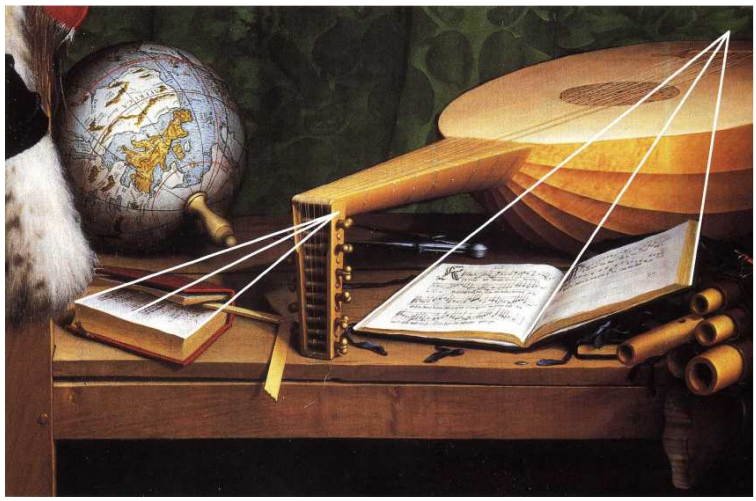


Figure 19.25: A collage of analytical perspectives in this detail of a painting by Holbein suggests that, if an imaging device were used to make the painting more accurate then it was viewed multiple times from multiple locations. Why don't we notice—or care—about this except when analyzing the history of art? Figure from Hockney.

Chapter 20

Correspondence Energy and Cooperation

introduction to neurophysiology of stereo; energy models; cooperative algorithms; local vs global stereo; coincidence firing; ocular column development

20.1 Introduction

In the previous chapter we 'solved' the correspondence problem by matching patches against patches individually. But this local approach was limited in its power. Moreover, there is far more structure to stereo than we are exploiting – the context in which each patch is placed. Looking around at objects in our world, they're often pretty smooth – at least the surfaces are – which implies that nearby disparities should not be too different. The idea behind a COOPERATIVE COMPUTATION is that somehow these are linked. In this chapter we describe an early way of doing this linking. Subsequent lectures will develop this use of context much more deeply.

In this chapter we also consider how the correspondence problem might be solved by networks of neurons. We consider two approaches: complex stereo cells, or "energy" based models, and richer interactions among neurons that implement a form of cooperation between them.

20.2 Energy Models Report Relative Disparity

Recall the Gabor filters from the last lecture. In the complex form this suggests two component filters: a sine harmonic inside a Gaussian enveloope and a cosine harmonic inside the envelope. Fig. 20.1.

By squaring the response and adding the outputs of such neurons together, we obtain the "energy" model of early visual computation. This was first developed for the study of motion (Adelson and Bergen) and has been applied widely. Fig. 20.2.

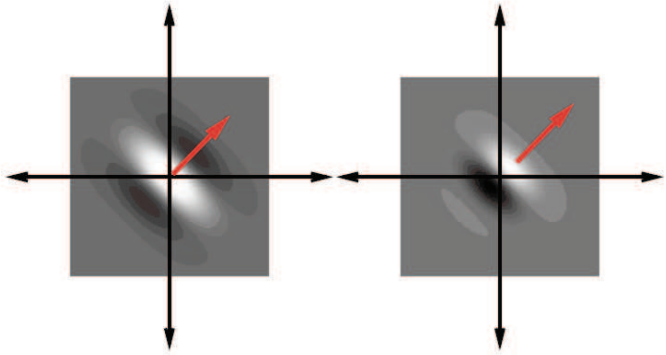


Figure 20.1: Illustration of real (left) and imaginary (right) components of a complex Gabor function. Think of these as receptive field pairs at the same position (note axes), and observe the complementary positions of the excitatory and inhibitory subzones of the receptive fields. Fig from Movellan

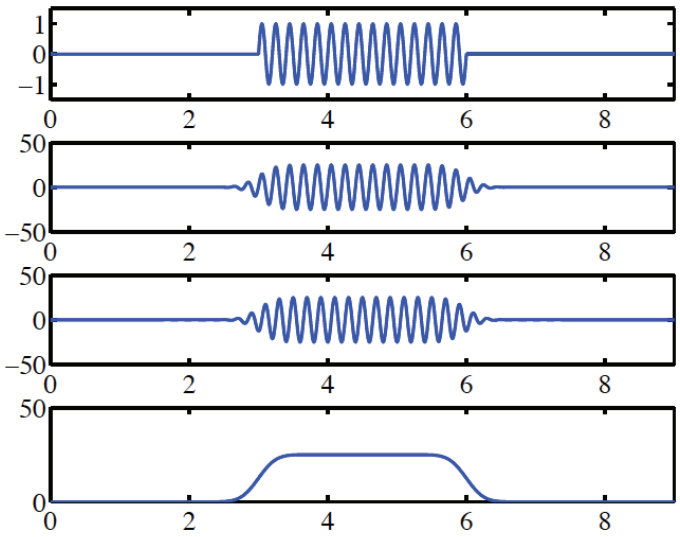


Figure 20.2: Illustration of the "energy" response obtained from squaring and summing (approximate) quadrature pairs of Gabors. (top) Input signal (think of a number of thin, parallel bars or a high spatial freq grating). (second) Cos Gabor convolved with input. (third) Sine Gabor convolved with input. (bottom) Energy output. Note this is "high" across a range of positions, like a complex cell's response. Fig from Movellan

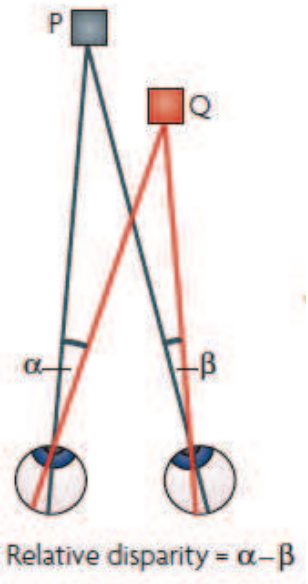


Figure 20.3: With the eyes verged on the point P, the nearby point Q has the relative disparity $\alpha - \beta$. Fig from Andrew Parker.

The advantage of energy models can be understood by thinking a little more deeply about the definition of disparity; we are sensitive to RELATIVE DISPARITY (Fig. 20.3) and not ABSOLUTE DISPARITY which requires a coordinate system directly on the retina (say, w.r.t. the center of the fovea). The evidence is that it's all done in relative coordinates.

20.2.1 Random Dot Stereograms

It is useful to consider matching binary images of random dots – so called RANDOM DOT STEREOGRAMS. These were developed by B. Julesz 50 yrs ago in an important demonstration that stereo fusion could happen *without* high-level object feedback.

Now, suppose we are matching a pixel in the left image with a pixel in the right – they better have the same brightness (black or white); it is unlikely that a black pixel will match a white one. However, energy models respond to both, as do neurons in V1! There is clearly more to stereo than this!! See Fig. 20.4.

20.3 Cooperative Computation

Neurons form networks, and thus far we have only been considering extremely simple feed-forward networks. Recalling the recurrent models introduced for the Limulus,

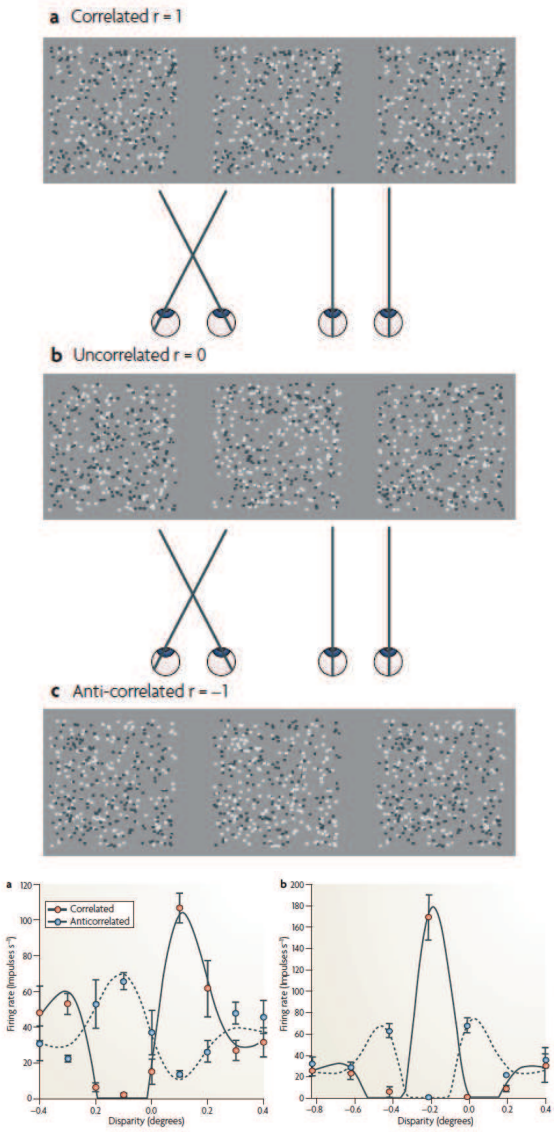


Figure 20.4: Random dot stereograms. (top) (a) By cross fusing the left/middle pair (or uncrossed fusing of the middle/right pair) a disk is seen in front of the background for correlated dot pairs. (b) when the dots are uncorrelated, only noise is seen; no depth effect. (c) when the paired dots are opposite contrast, so-called anti-correlated, correlation like "energy" models still establish correspondence, although we never see such patterns as giving rise to a depth effect. There must be more to stereo than energy. (bottom) Responses of the Gabor filters to correlated and anti-correlated random dot patterns. Notice that there is a strong response in both, as predicted by a correlation model; although there is a weaker response to the anti-correlated signal which would not be predicted by such a model. The sum of the squares of these responses is the energy model. Fig from Andrew Parker.

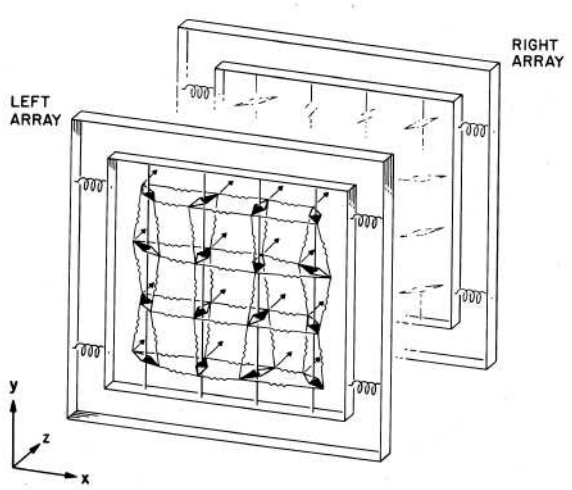


Figure 20.5: Julesz' "spring-loaded disparity model" based on a dipole model for matching pixels. Suppose there was a dipole representing the brightness of pixels in the left and right images; we would want their brightnesses to match under correspondence. But there is also a condition that matches should be near each other. Spring-loaded models provide a physical metaphor for this type of interaction. Example from Julesz, 1971.

however, perhaps we can get out of the frontal-parallel trap by allowing interactions between locally estimated disparities to modify them. This is the idea behind cooperative algorithms as a model for neural networks; see Fig. 20.5.

As a warm-up to start thinking about algorithms, remember the gradient ascent algorithm: start at a given position; take a step in the direction of the gradient and check for the new gradient; if 0 then done; otherwise take step in new gradient direction and repeat. If one simply looks at the springs in Fig. 20.5 then it might not be too huge a stretch to realize that something similar—or analagous—will be going on here.

To introduce such networks for stereo, it is useful to consider matching binary images of random dots – so called RANDOM DOT STEREOGRAMS. Now, suppose we are matching a pixel in the left image with a pixel in the right – they better have the same brightness (black or white); it is unlikely that a black pixel will match a white one.

In addition, if two pixel pairs derive from the projection of two nearby points in space on the same surface, then they should have (about) the same disparity. So there are two ingredients to the cooperation: local (pixel) matches and (global) consistency of nearby disparities. (If nearby disparities varied widely then they could not come from the same smooth surface.) The physical energy in the dipoles and the springs can represent each of these two ingredients – the orientation of the dipole represents how

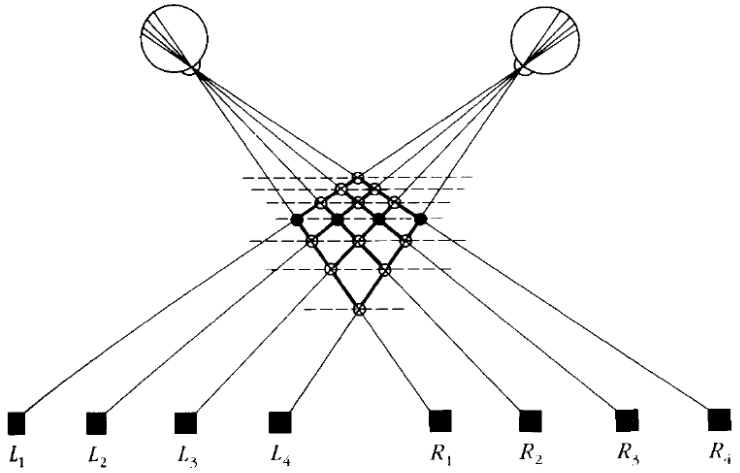


Figure 20.6: Kepler diagram for ambiguous matches in stereo. Image from Marr.

well the pixels match and the springs between dipoles represents consistency between nearby dipole orientations. For natural ferromagnetric materials, one would expect there to be domains in which the dipoles align (surfaces in stereo) and transitions between domains (edges between surfaces).

We develop this idea a little more fully by reviewing the Kepler diagram for ambiguous matches (Fig. 20.6. We first make sure that those matching image patches in the left and the right images have similar structure (what Marr calls the compatibility constraint); we need to remove those matches at the same location with many correspondes in the other image (what Marr calls the uniqueness constraint); we need to enforce those disparities that are spatially near one another to be similar (the smoothness constraint).

The diagram in Fig. 20.7 suggests how to transform these into a neural network. It represents a 1-dimensional stereo problem. The basic idea is to use two constraints:

- *continuity* Since surfaces are smooth, disparity values should vary very little over short distances.

- *uniqueness* Each point on a surface can have only one disparity (depth) value.

See figure caption for explanation.

For the basic representation, Marr starts with an array C(x,y,d) which takes on the value 1 when d is the correct disparity at position (x,y). Now, start with an initial, local estimate of $C^0(x, y, d)$ for an appropriate range of d's and iterate:

$$C^{n+1}(x,y,d) = \sigma\left[\sum_{x',y',d' \in S(x,y,d)} C^n(x',y',d') - \epsilon\left(\sum_{x',y',d' \in O(x,y,d)} C^n(x',y',d')\right) + C^0(x,y,d)\right] \tag{20.1}$$

until convergence. Here σ is the sigmoid function, ϵ balances the excitation with the inhibition, S represents the diagonal "neighborhood" for excitation (nearby values are similiar) and O represents the "vertical" neighborhood for inhibition (there is only one correct disparity value for each position). Results in Fig. 20.9.

Note that formula specifies a kind of balance between the initial match data and the consistency of nearby points; such a balance will be important for later consideration.

Remark: although we are now effectively minimizing a function over disparity, it still prefers the frontal-parallel solution!

While this is good news for the birthday cake example, it does not solve the problem for the natural world.

A key feature of such cooperative algorithms is that the solution tends to lie in domains of constant value. The good news here is the idea of introducing some constraints on the problem from domain knowledge; the bad news is that these are the wrong constraints.

20.4 Fusion, Binocular Summation, and Binocular Rivalry

How might stereo vision have evolved? While it is clearly useful in providing information about the third dimension, what is the 'evolutionary pressure' that drove it? The first requirement is two eyes facing in about the same direction (for us this if forward) with a large region of BINOCULAR OVERLAP. That is, think about laterally-placed eyes and ask: how do they rotate forward? Clearly this must preceed any development in the brain of circuitry for stereo.

It's difficult to make a case that evolutionary pressure derives from the third dimension, because how would evolution 'know' to drive toward this? It turns out that there's a simpler issue to think about that could have preceeded stereo.

We begin with an experiment, which is a variation on the one in XXX. Consider looking at an image of a sine wave grating, but in which the contrast is low. Our task is to state whether a sine grating or a constant image was present. (This is an example of a two-alternative, forced choice task. It is an excellent way to do psychophysics, because the analysis is well understood.) Should the result – the minimum contrast necessary to detect the presence of the sine grating –

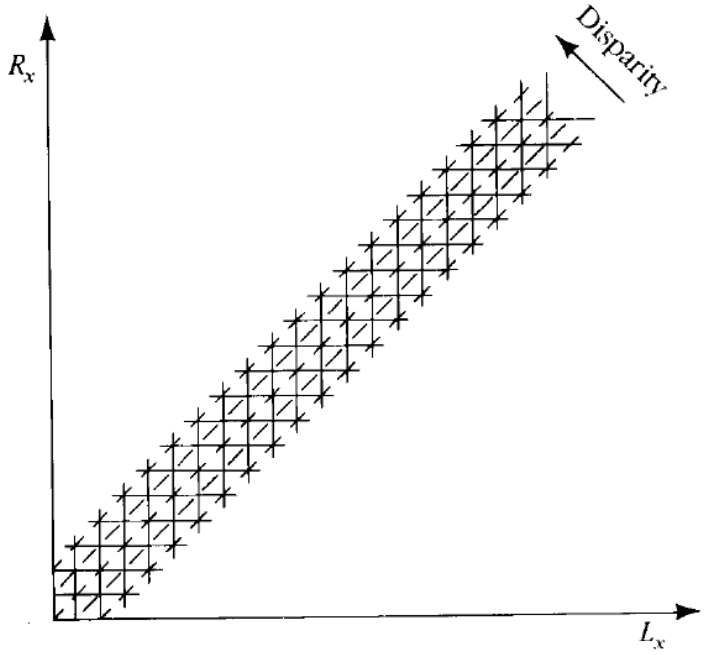


Figure 20.7: Compatibility structure for Marr's cooperative stereo algorithm. Let L_x represent position in the left image, and R_x position in the right. Now, a short vertical line indicates that, for some position in the left image there are several matches in the right one; this should not be allowed so there should be inhibition between disparities corresponding to these lines. Now, nearby positions should have about the same disparity, so there should be excitation along the diagonal, dashed lines. This defines an excitatory/inhibitory network among cells representing disparity. Image from Marr. reference: Cooperative computation of stereo disparity (1976) by D Marr, T Poggio Science

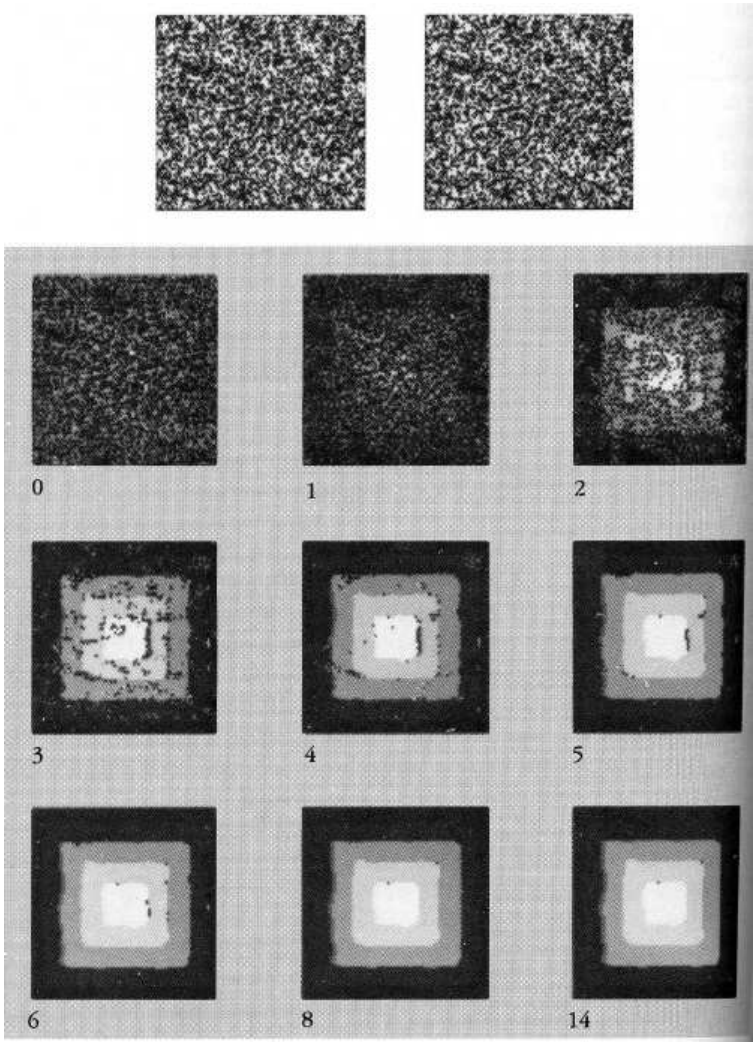


Figure 20.8: Marr's result on a random dot stereogram of a "birthday cake." Note that it consists of a series of nested frontal-parallel planes! Image from Marr.

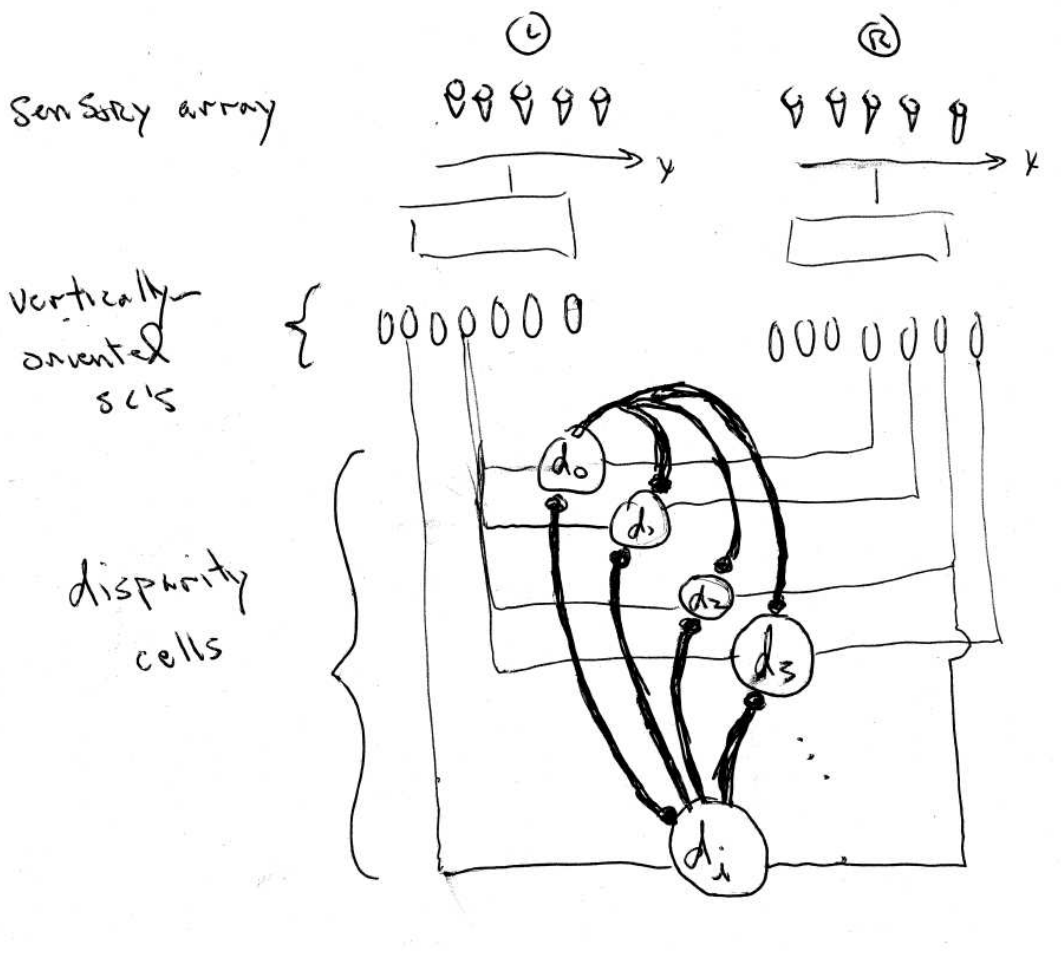


Figure 20.9: Layered networks that compute features at different levels of abstraction. We start with the L and R sensory arrays distributed along the single spatial dimension x. the empty boxes denote the projection from retina to LGN to V1, where we show a number of simple-cell receptive fields (oriented vertically) along x. The local disparity computation then yields some initial estimates of disparity, represented by cells d_i via the thin lines. Finally, there is a network (thick lines) running recurrently among the disparity neurons that implement e.g. smoothness constraints.

Stereo may well have evolved from the binocular overlap and summation of images in both eyes. This provides a $\sqrt{2}$ improvement in signal/noise ratio, which could really matter in low-light (dawn; dusk; rainforest) situations.

To do this correctly for a system in which early receptive fields are orientationally selective would require some care in computing the summation. Clearly, orientation will matter. This is also true for computing stereo correspondence. We shall return to this in a few lectures.

Suppose two different images are placed in front of our eyes; this causes a complex rivalry pattern that seems hardly normal. Somehow this says the eyes' inputs should be coordinated. It also raises the issue of when it is possible to "fuse" the information from the two eyes. We're going to have to do a little work before we can determine what these limits are.

20.5 Ocular Dominance Columns and their Development

(This section is only for those interested in starting to think about learning and development; it is outside the context of the current class lectures.) To maintain an overall retinotopic mapping from retina to cortex we are confronted with two competing notions: each eye seeks its own consistent map; how can the two be integrated into a single solution?

Cortical development has achieved a solution of ocular dominance bands in which nearby points in the left image are "nearby" in the cortex while nearby points in the right image are also "nearby" in the cortex; see Fig. 20.10. The solution is part genetic and part developmental. We here concentrate on the developmental part and illustrate a class of models that can achieve qualitative solutions.

20.5.1 Hebbian learning

Basic idea: "cells that fire together wire together." In other words, the change in synapses over time:

$$\frac{\Delta S}{\Delta t} = (\text{post} - \text{synaptic})(\text{pre} - \text{synaptic}) - (\text{decay})$$

coincidence firing and the product computation

20.5.2 Models for Activity-Dependent Development

Basic problem: given geniculo-cortical projections that are overlapped and uniform, how can these be organized? Layout from Miller paper in Fig. 20.11.

Swindale model:

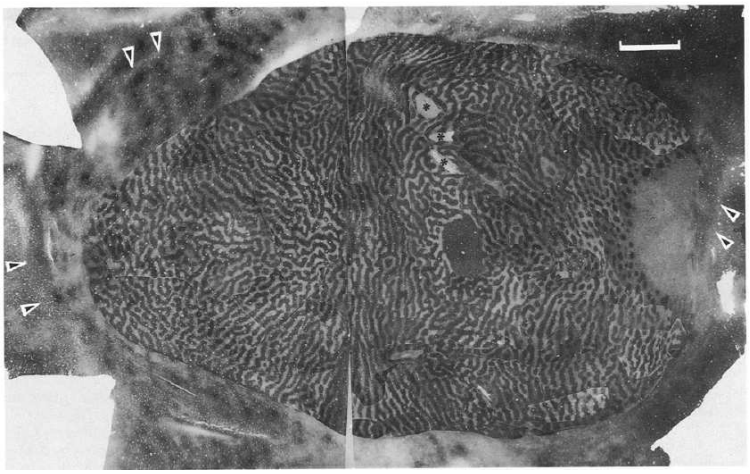


Figure 20.10: The ocular dominance bands in Cebus monkey. Black indicates cells enervated by one eye; white by the other. Note that this solution maintains some semi-open neighborhood around each point. Image from Rosa, Gattas, Fiorani.

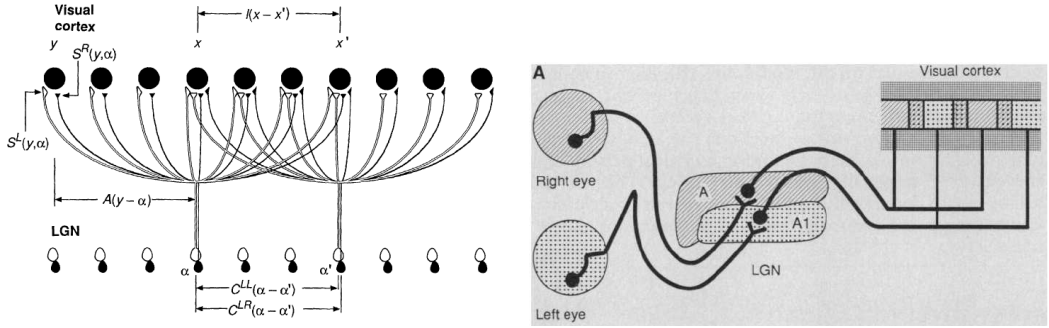


Figure 20.11: Setup for the activity-dependent organization of ocular dominance columns from Miller paper. Starting with a general projection from the LGN to cortex, with a uniform spanout; how can the structured connections pattern (right) be learned?

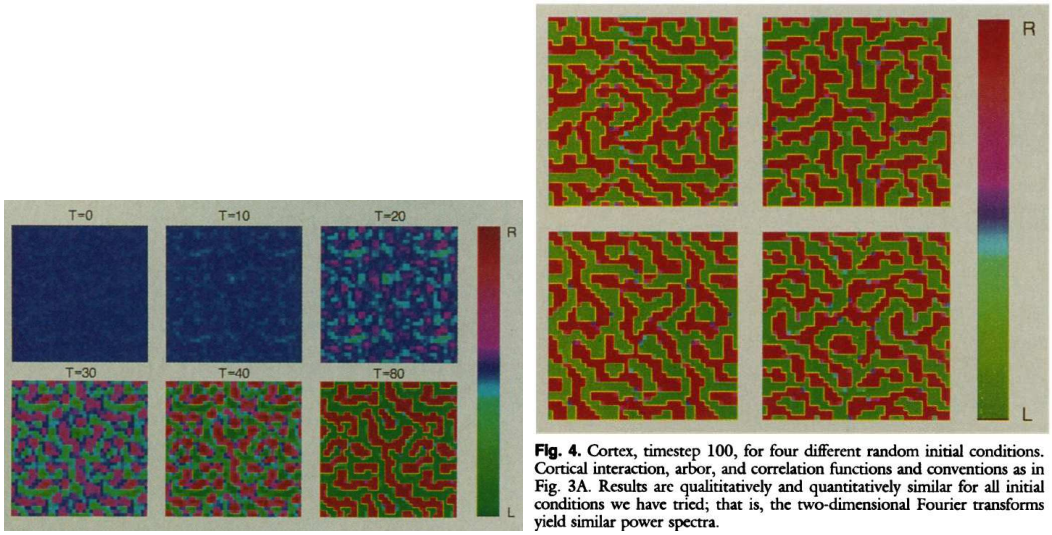


Fig. 4. Cortex, timestep 100, for four different random initial conditions. Cortical interaction, arbor, and correlation functions and conventions as in Fig. 3A. Results are qualititatively and quantitatively similar for all initial conditions we have tried; that is, the two-dimensional Fourier transforms yield similar power spectra.

Figure 20.12: Results of running the Miller evolution equations in time; notice how the OD bands derive from small "islands" of non-uniformities. The reaction

$$\frac{\partial n^R}{\partial t} = f(n^R)[w_a * n^R + w_i * n^L] \frac{\partial n^L}{\partial t} = f(n^L)[w_a * n^L + w_i * n^R] \tag{20.2}$$

where n^i is the synapse density for the left/right eye.

Miller, Keller, Stryker Model results in Fig. 20.12.

20.5.3 Deprivation and Ocular Dominance Development

For animals derived in the dark, or with one eye sutured closed, one would expect the cortex to develop in a non-standard manner, with the deprived eye commanding less cortical representation than the unsutured eye. This is precisely what happens, and confirms the activity-dependence component of cortical development.

20.6 The Pulfrich Effect

There is a fascinating demonstration that links both retinal processing and stereo, to illustrate how earlier processing links influence later ones.

20.7 Summary

what have we done in this lecture: Gabor models of receptive-field structure led to different implementations of the disparity offset; but reduce to correlation in compu-

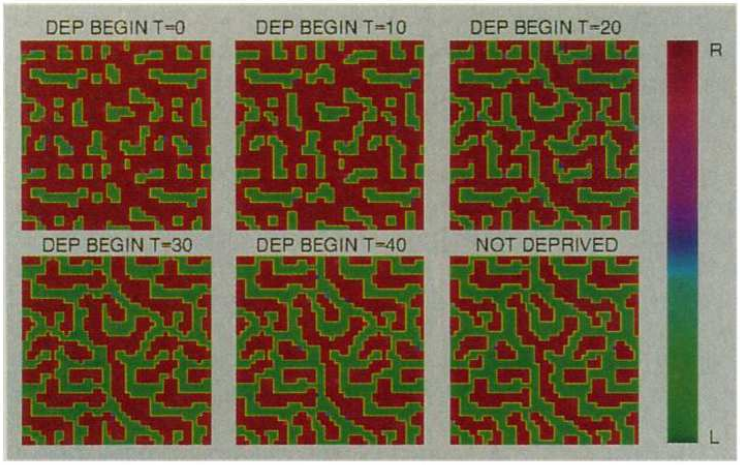


Fig. 7. Results of monocular deprivation. Results at timestep 200 are shown for initiation of monocular deprivation at five different times (timestep 0, 10, 20, 30, and 40). The sixth panel shows, for comparison, timestep 200 in an identical run but without deprivation. Arbor, correlation, and cortical interaction functions, initial conditions, and conventions as in Fig. 3A except

Figure 20.13: Results of running the Miller model with one eye deprived of input from different points during development.

Figure 20.14: Illustration of the Pulfrich effect. A pendulum swinging in the frontal-parallel plane is viewed in stereo (left). When a neutral density filter is placed over one eye, the pendulul appears to swing in an ellipsoidal arc, with the depth information induced by a delay caused by neural integration factors. Figure from: `http://www.siu.edu/~pulfrich/Pulfrich_Pages/explains/expl_txt/explaint.html`

tational competence.

Networks at a next level of abstraction start to incorporate properties of images that we expect to find in scenes. These become quite general purpose. We used the same type of continuity argument that arose with logical/linear operators. But this is clearly only the beginning.

While we have understood something about the geometry of stereo and the computational requirements of the corresondence problem, we still are stuck in the frontal-parallel plane assumption about the world. This is really a key part of computational reasoning: what can we do and what are the fundamental abstract limitations of an approach. It can only be obtained by an abstract mathematical analysis; see Fig. 20.15. Somehow we're going to have to get more general surfaces into all of this and the way they interact with light.

How to move on from here – diagram illustrating piecewise flat approximation and the need for a "higher-order" constraint from the surface world. Mr. Taylor rises again.

20.8 Notes

The classic papers on binocular summation are:

Gordon E. Legge Binocular contrast summation I. Detection and discrimination Vision Research, Volume 24, Issue 4, (1984) Pages 373-383

Gordon E. Legge Binocular contrast summation II. Quadratic summation Vision Research, Volume 24, Issue 4, (1984)Pages 385-394

On the Pulfrich effect:

Pulfrich, C. (1922) Die Stereoskopie im Dienste der isochromen und heterochromen Photometrie. Die Naturwissenschaften, 1922, 10, 553 - 564; 569 - 574; 596 - 601; 714 - 722; 735 - 743; 751 - 761.

Burr, D. C. & Ross, J. (1979) How does binocular delay give information about depth? Vision Research, 1979, 19, 523 - 532.

Important to connection to optical flow: see Qian and Anderson.

Nice description of the cooperatative stereo algorithm in Marr's book **Vision**. For students interested in the subject, this book is essential reading.

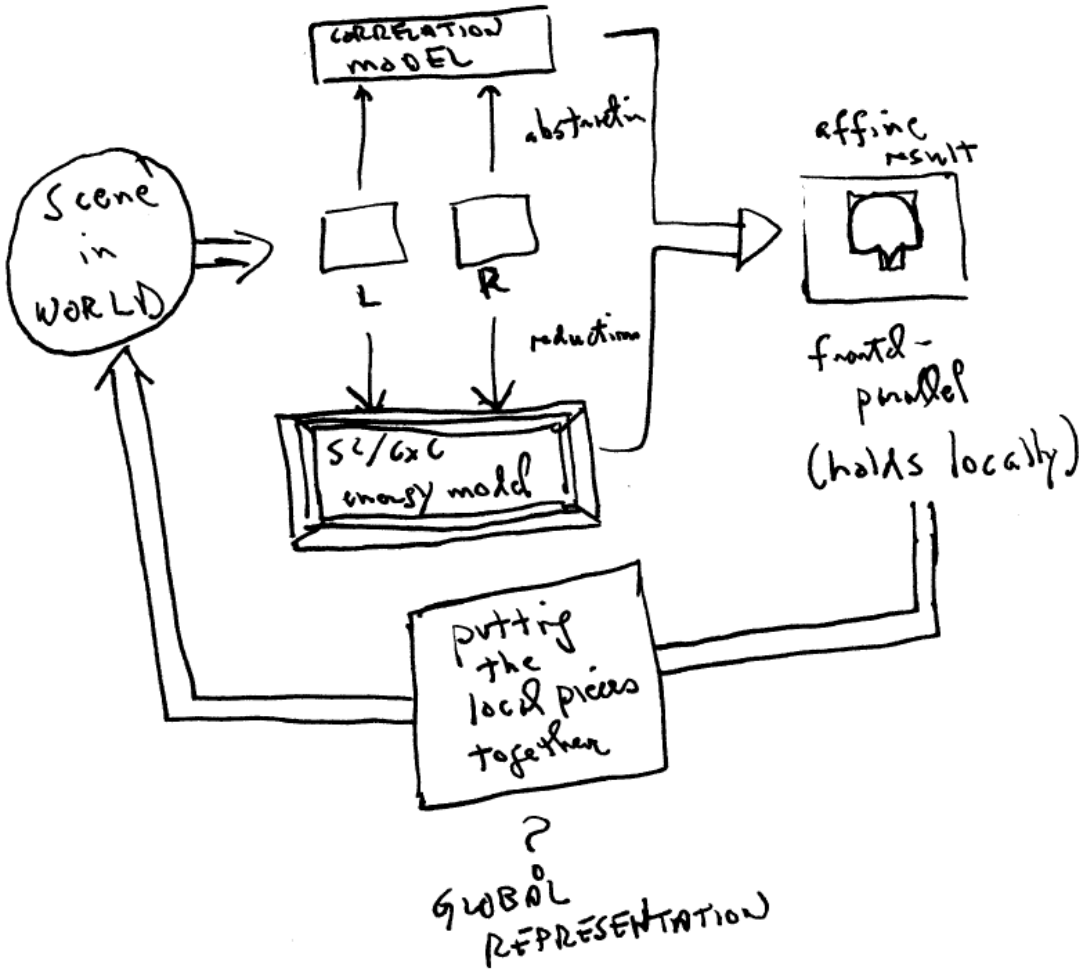


Figure 20.15: Computational analysis reveals the fundamental capabilities of an approach at an abstract level.