Lecture 8. Stereoscopic Vision

Reading Assignments:

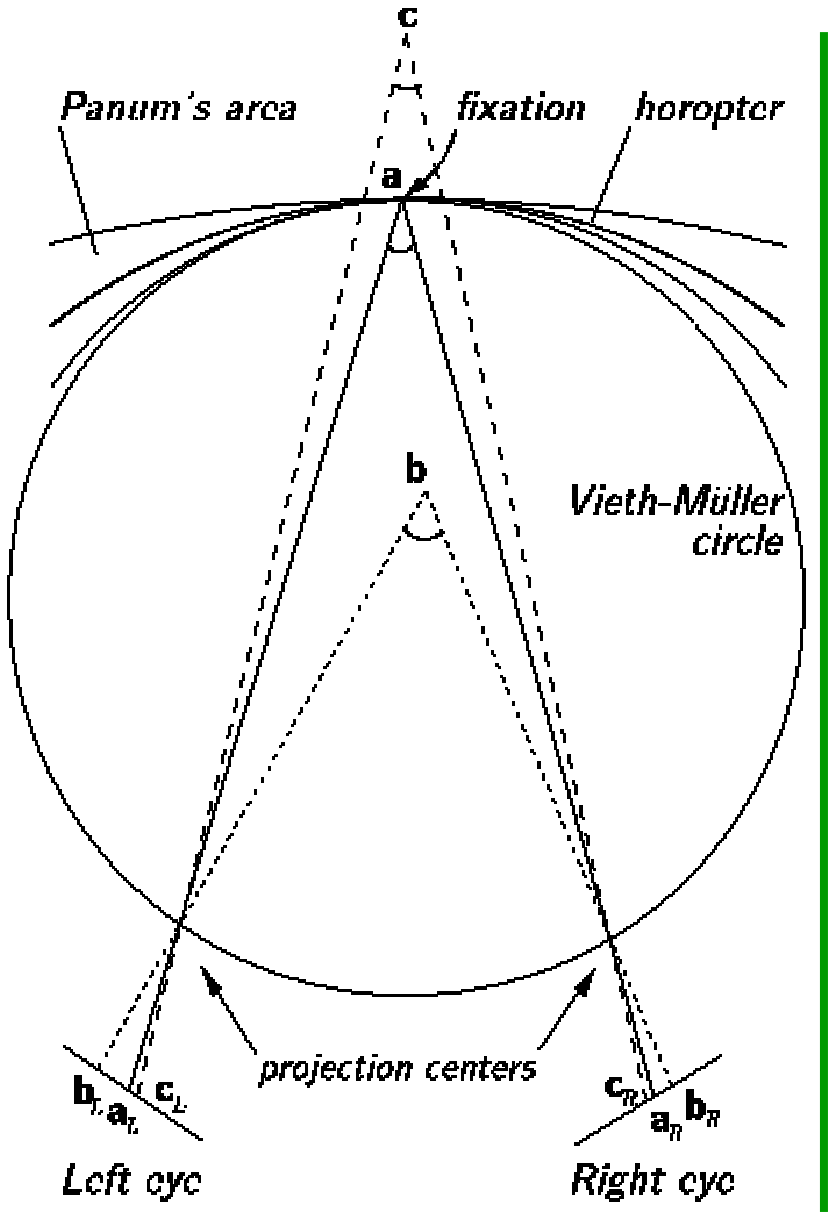
Second part of Chapter 10.

Seeing With Multiple Eyes

From a single eye: can analyze color, luminance, orientation, etc of objects.

But to locate objects in depth we need multiple projective views.

Objects at different depths/distances yield different projections onto the retinas/cameras.



Depth Perception

Several cues allow us to locate objects in depth:

Stereopsis: based on correlating cues from two spatially separated eyes.

Optic flow: based on cues provided to one eye at moments separated in time.

Accommodation: determines what focal length will best bring an object into focus.

Size constancy: our knowledge of the real size of an object allows to estimate its distance from its perceived size.

Direct measurements: for machine vision systems; e.g., range-finders, sonars, etc.

Stereoscopic Vision

1) Extract features from each image, that can be matched between both images.

2) Establish the correspondence between features in one image and those in the other image. Difficulty: partial occlusions!

3) Compute disparity, i.e., difference in image position between matching features. From that and known optical geometry of the setup, recover distance to objects.

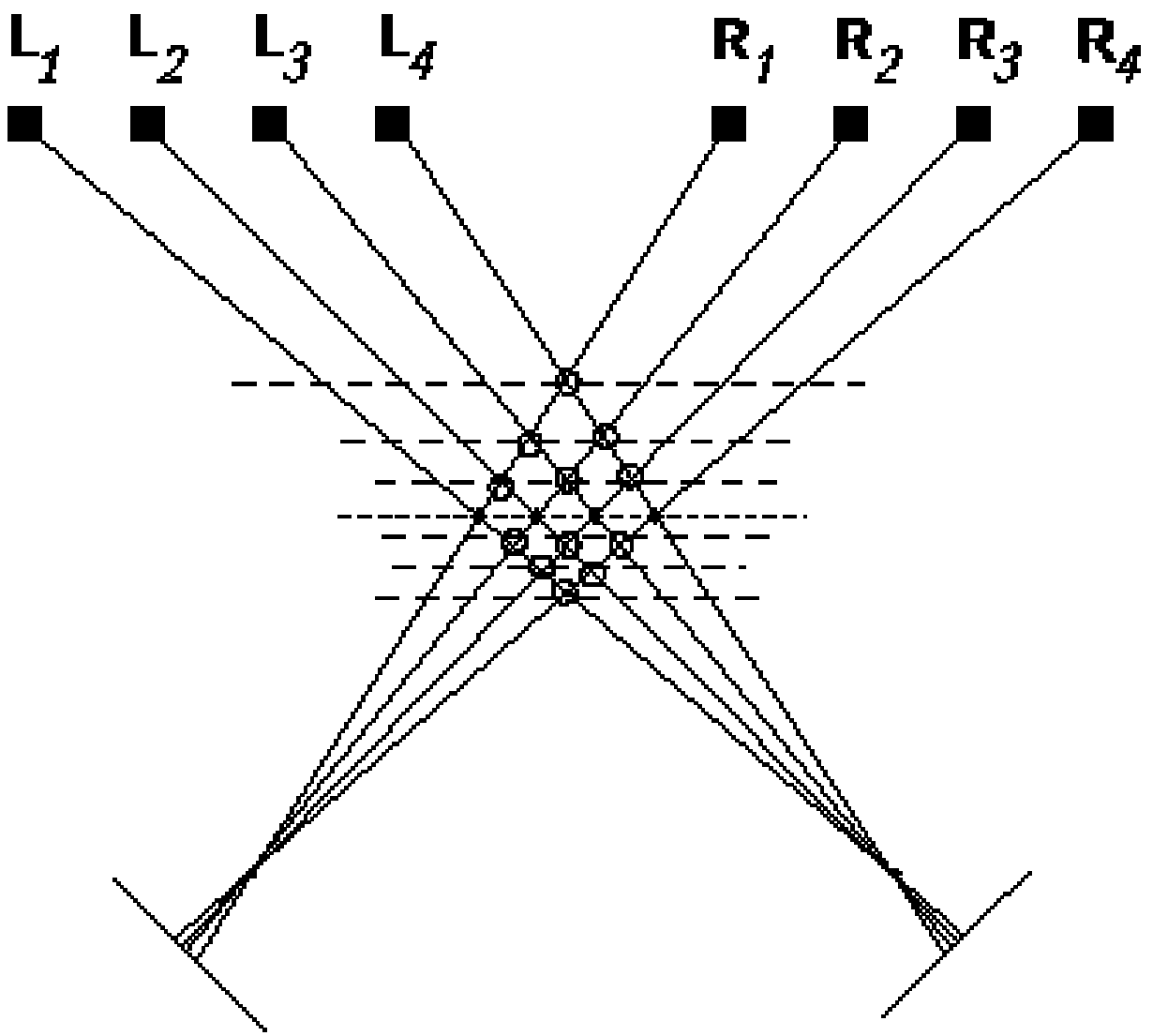
4) Interpolation/denoising/filling-in: from recovered depth at locations of features, infer dense depth field over entire images.

The Correspondence Problem

16 possible objects…

but only 4 were actually present…

Problem: how do we pair the L_i points to the R_i points?



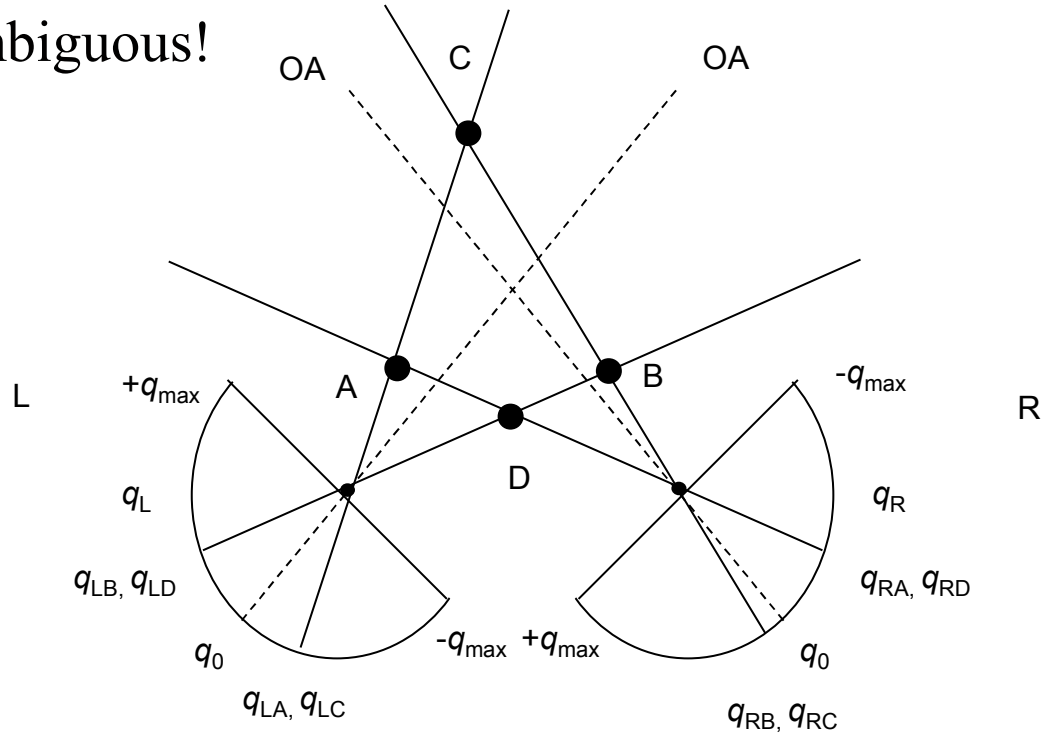
The Correspondence Problem

The correspondence problem: to match corresponding points on the two retinas such as to be able to triangulate their depth.

why a problem? because ambiguous!

presence of "ghosts:"

A scene with objects
A and B yields exactly the
same two retinal views
as a scene with objects
C and D.



Given the two images, what objects were in the scene?

Computing Correspondence: naïve approach

- extract “features” in both views.

- loop over features in one view; find best matching features by searching over the entire other view.

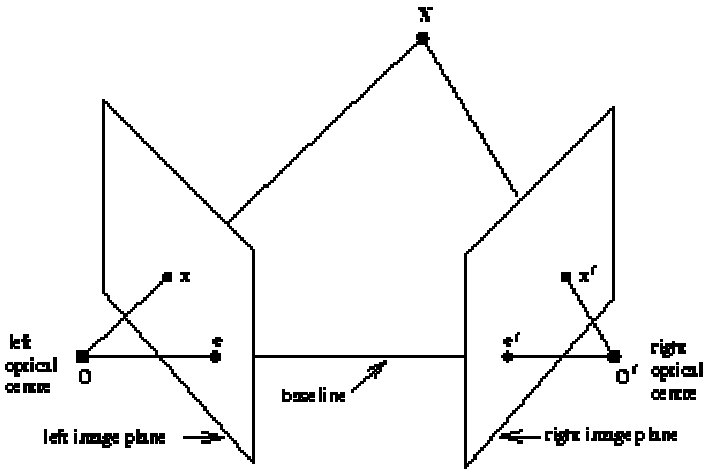
- for all paired features, compute depth.

- interpolate to whole scene.

Epipolar Geometry

baseline: line joining both eyes' optical centers

epipole: intersection of baseline with image plane



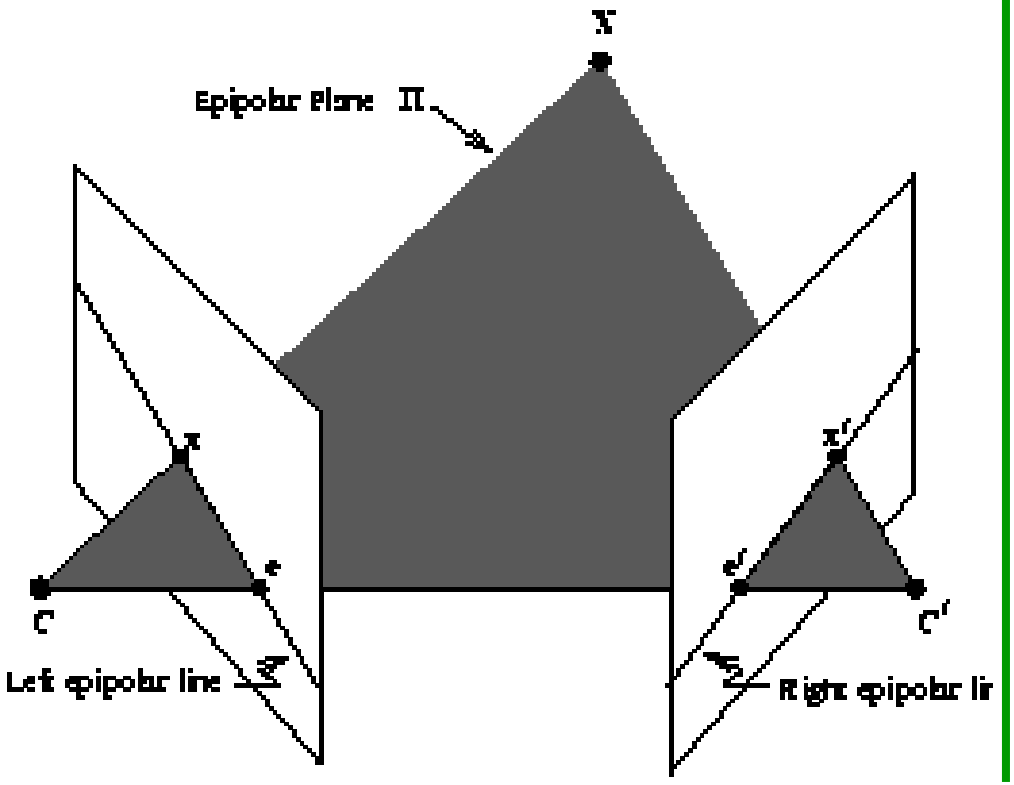
Epipolar Geometry

epipolar plane: plane defined by 3D point and both optical centers

epipolar line: intersection of epipolar plane with image plane

epipolar geometry: given the projection of a 3D point on one image plane, we can draw the epipolar plane, and the projection of that 3D point onto the other image plane is on that image plane's corresponding epipolar line.

So, for a given point in one image, the search for the corresponding point in the other image is 1D rather than 2D!



Feature Matching

Main issue for computer vision systems: what should the features be?

- edges?
- corners, junctions?
- "rich" edges, corners and junctions (i.e., where not only edge information but also local color, intensity, etc are used)?
- jets, i.e., vector of responses from a basis of wavelets; textures?
- small parts of objects?
- whole objects?

How about biology?

Classical question in psychology:

do we recognize objects first then infer their depth, or can we perceive depth before recognizing an object?

Does the brain take the image from each eye separately to recognize, for example, a house therein, and then uses the disparity between the two house images to recognize the depth of the house in space?

or

Does our visual system match local stimuli presented to both eyes, thus building up a depth map of surfaces and small objects in space which provides the input for perceptual recognition?

Bela Julesz (1971) answered this question using

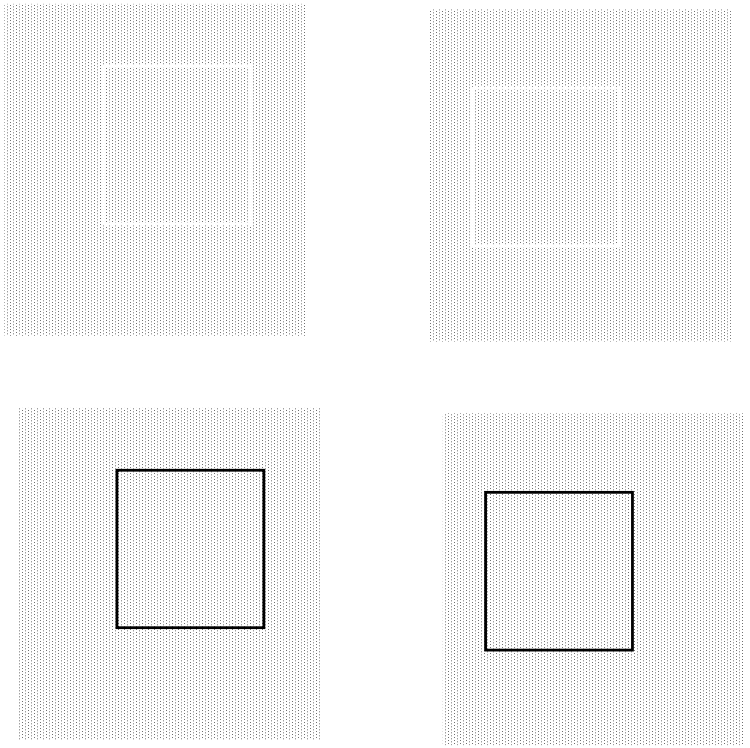
random-dot stereograms

Random-dot Stereograms

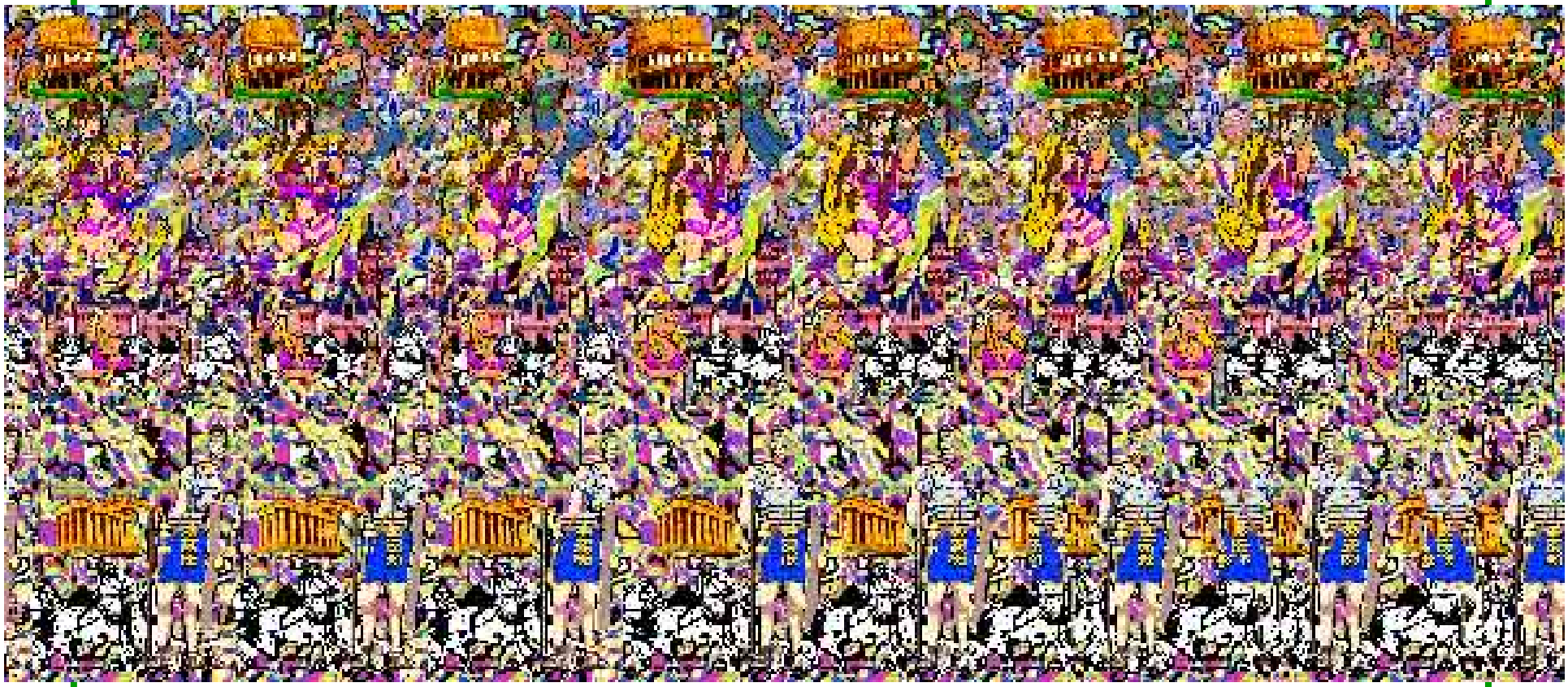
- start with a random dot pattern and a depth map

- cut out the random dot pattern from one eye, shift it according to the disparity inferred from the depth map and paste it into the pattern for the other eye

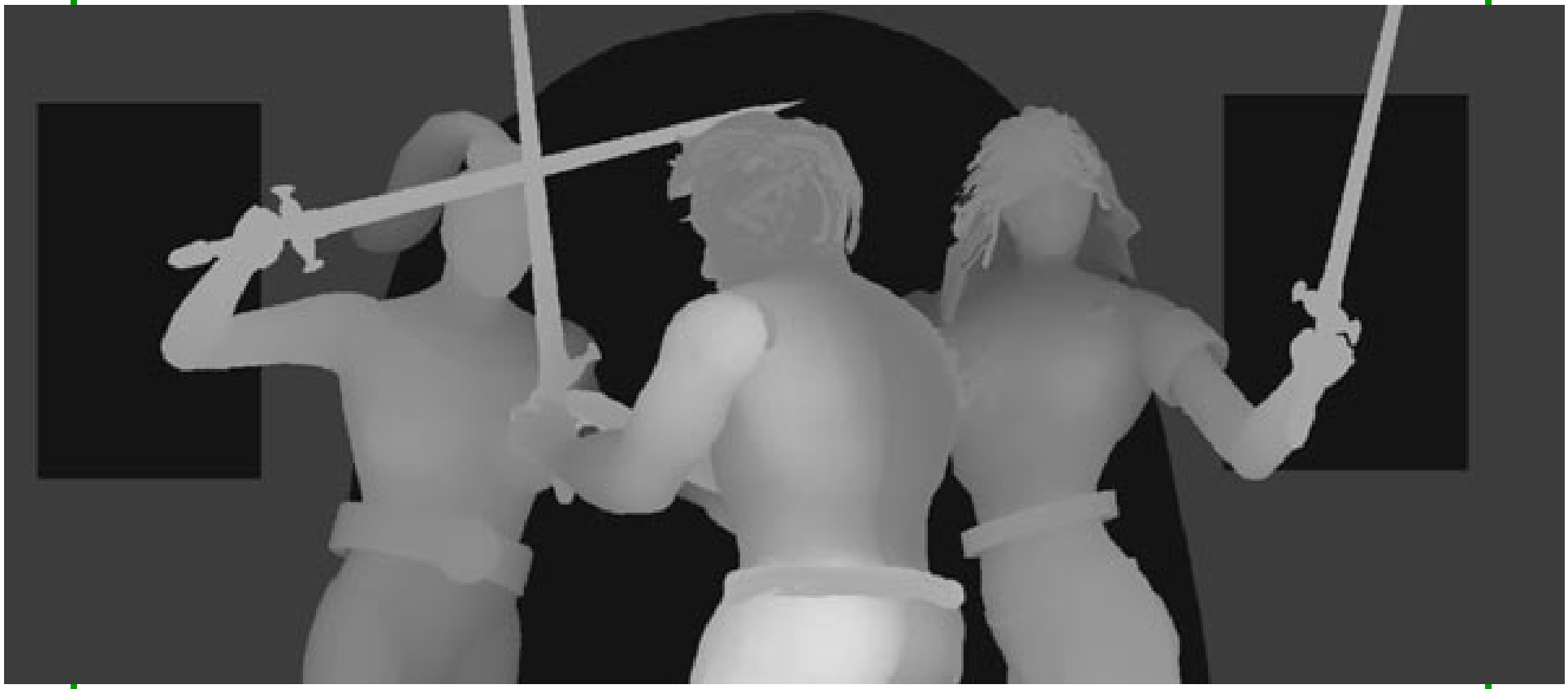
- fill any blanks with new randomly chosen dots.



Example Random-Dot Stereogram



Associated depth map



Conclusion from RDS

We can perceive depth before we recognize objects.

Thus, the brain is able to solve the correspondence problem using only simple features, and does not (only) rely on matching views of objects.

Reverse Correlation Technique

Simplified view:

Show random sequence of all possible stimuli.

Record responses.

Start with an empty image; add up all stimuli that elicited a response.

Result: average stimulus profile that cause the cell to fire.

I. Ohzawa, G. DeAngelis, & R. Freeman --- Encoding of Binocular Disparity by Simple Cells

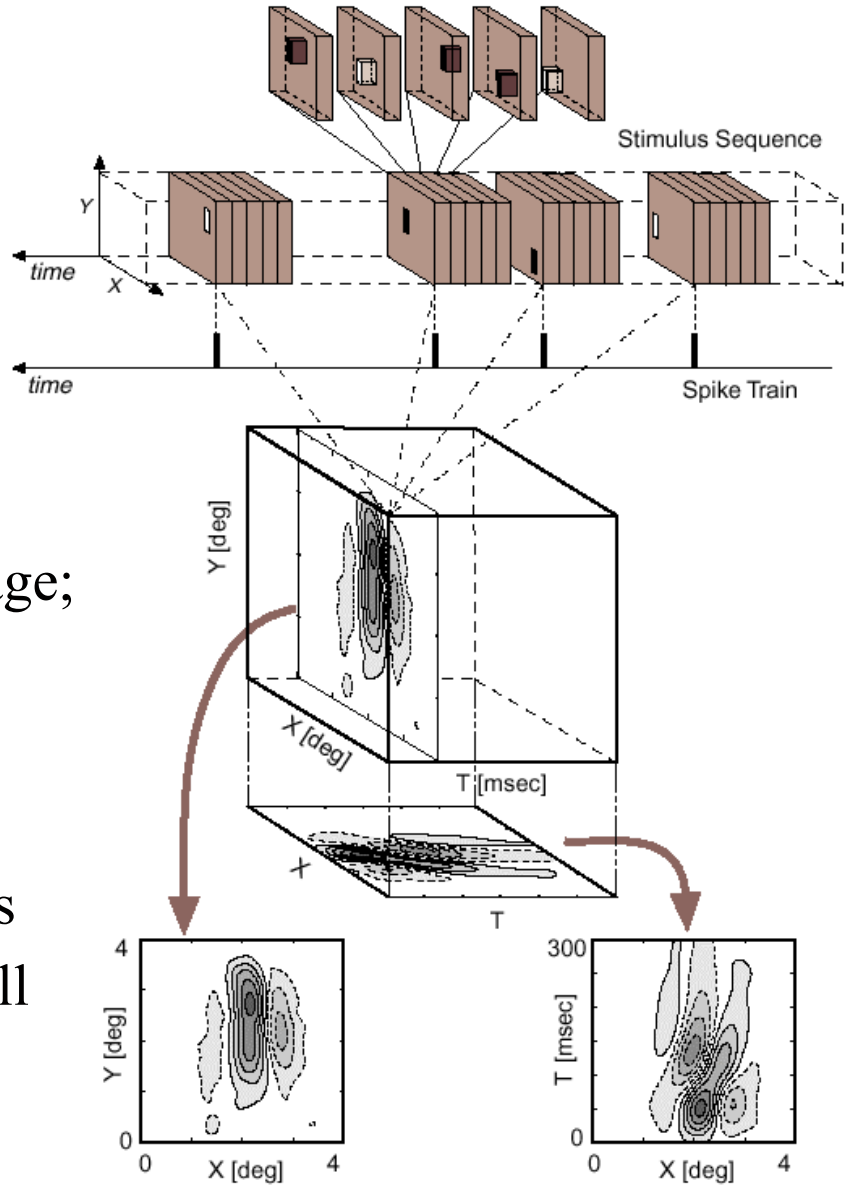


Fig. 3 A reverse correlation technique is used to measure space-time RFs of simple cells. A section of a stimulus sequence is illustrated at the top. A part of it is shown in an exploded view to reveal individual stimuli that are typically 30-50 msec in duration. Each stimulus is a bright or dark bar presented on a gray background at one of 20x20 grid points. The position and the polarity (bright or dark) are randomized. Although stimuli are presented with no interstimulus interval, not all will elicit a discharge from a neuron as indicated by the relatively sparse spike train. The reverse correlation technique may be understood as a procedure for obtaining the *average stimulus profile* that has caused a cell to fire. This is achieved by summing sections of the stimulus sequence that immediately precede each spike (shown by shaded cubes in the stimulus sequence) for all spikes generated. The result is a three-dimensional map of X, Y and T (two of space and one of time) that indicates the effectiveness of a stimulus flashed at position (X, Y) in causing the cell to fire T msec after the flash. This is the definition of a RF in space and time. Various cross-sections and projections of the 3-D data are used for clarity of presentation as shown at the bottom.

Spatial RFs

Simple cells in V1 of cat.

Well modeled by Gabor functions with various preferred orientations (here all normalized to vertical) and spatial phases.

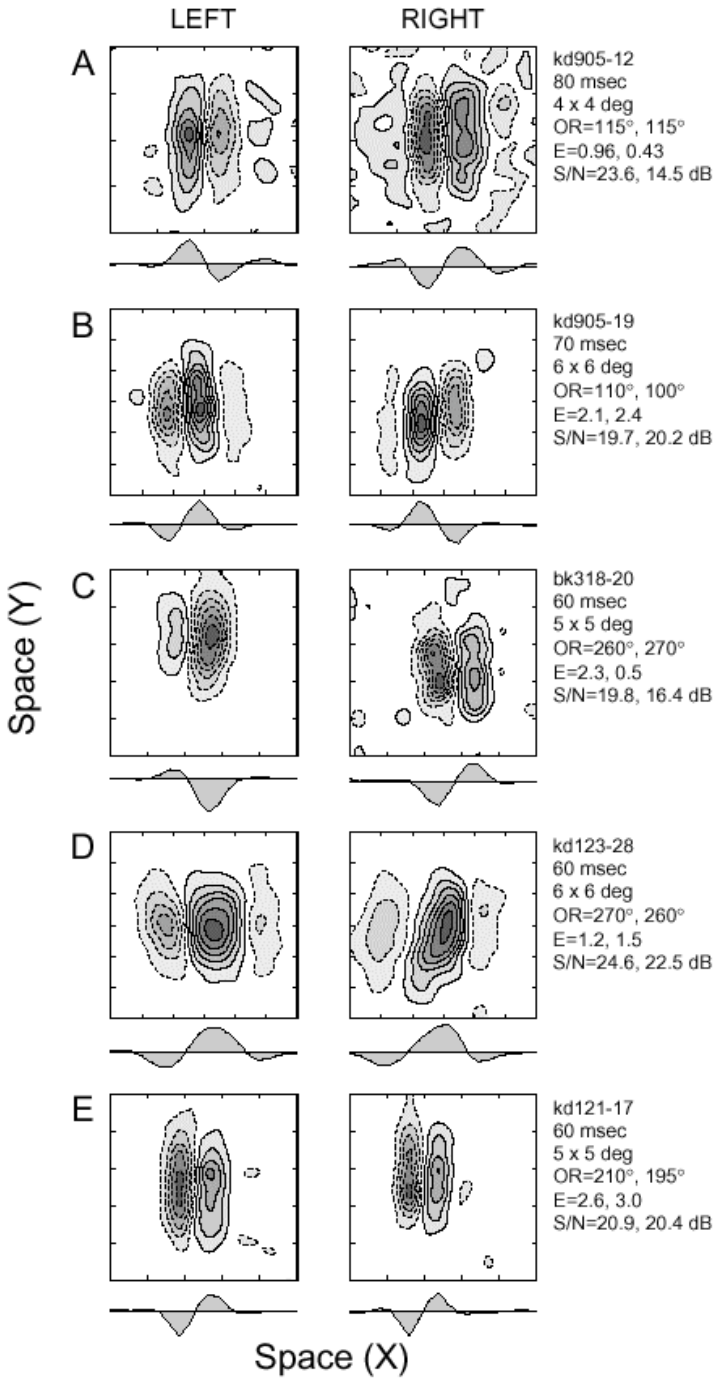


Fig. 4 Two-dimensional spatial (X-Y) RFs are shown as contour plots for left and right eyes for 5 simple cells (**A-E**). The X dimension of the RF, which is perpendicular to the preferred orientation in all cases, is represented on the horizontal axis. In this and all subsequent figures, solid and dashed contours represent bright and dark excitatory subregions, respectively. The shading is proportional to the strength of the response. A 1-dimensional profile, obtained by integrating the 2-D profile along the Y dimension, is also shown below each 2-D profile. Details of recording are indicated to the right of the profiles in the following order: animal code-cell number, correlation delay at which spatial profiles are obtained, size of the stimulus grid, preferred orientation (OR; 0 , 180 =horizontal; 90 , 270 =vertical) for left and right eyes, *efficiency* (E) which is the average number of spikes generated by a single flash at the highest peak of the spatial profile, and signal-to-noise ratio (S/N) of the 1-D profile for the two eyes (see text). Because the mapping grid has been rotated to match the preferred orientation of the RF, subregions are always elongated along the vertical.

RFs are spatio-temporal!

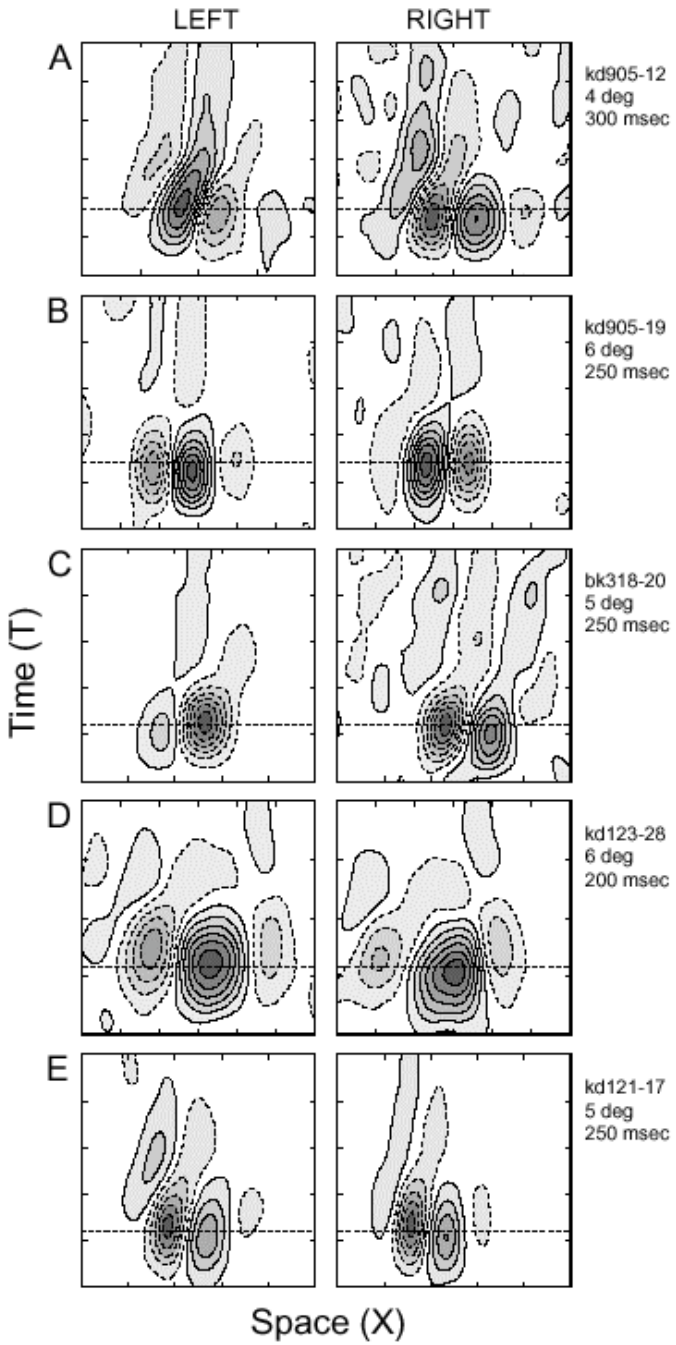


Fig. 5 Space-time (X-T) RFs are shown for the 5 simple cells (**A-E**) presented in Fig. 4. Horizontal and vertical axes represent space (X) and time (T) dimensions, respectively. T=0 at the bottom of each profile, and is maximum at the top. Details of experimental conditions are shown to the right of the profiles in the following order: animal code-cell number, size of spatial domain, and the maximum correlation delay. Horizontal dashed lines indicate the delay at which the spatial RF profiles in Fig. 4 were obtained (i.e. for time values at which there was a maximum response).

Parametrizing the results

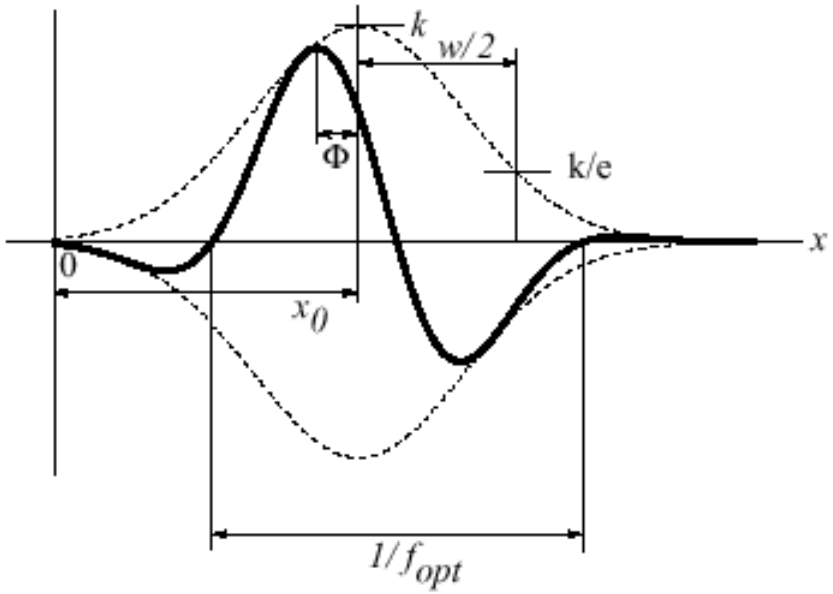


Fig. 7 Five parameters that define a Gabor function (solid curve) are illustrated. Parameters k, w, and x_0 are the amplitude, full width at 1/*e* of the peak, and center position of the Gaussian envelope, respectively, and f_{opt} and Φ are the spatial frequency and phase of the sinusoid, respectively (see eq. 1 in text). Dashed curves represent the Gaussian envelope. The phase is referenced to the center of the Gaussian envelope. This function is used to fit the spatial RF profiles of simple cells, and to extract five key parameters that characterize the RF.

Binocular-responsive simple cells in V1

Cells respond well to stimuli presented to either eye.

but the phase of their RF depends on the eye!

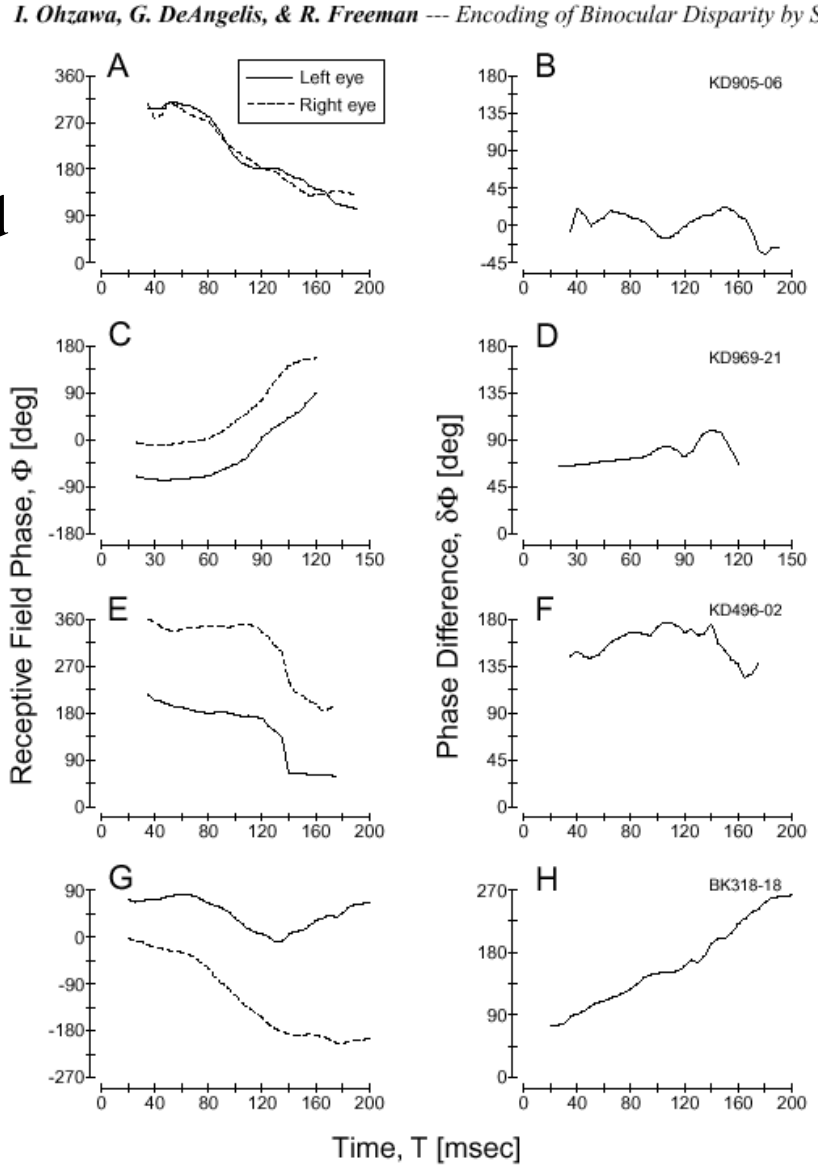


Fig. 10 Time course of the RF phase parameter is shown for 4 additional cells. Other parameters are not shown. A, C, E, and G depict the phase parameter for the two eyes for cells indicated by the animal code-cell number at the right. B, D, F, and H depict the phase difference between the two eyes for the same cells.

Ohzawa et al, 1996

Space-Time Analysis

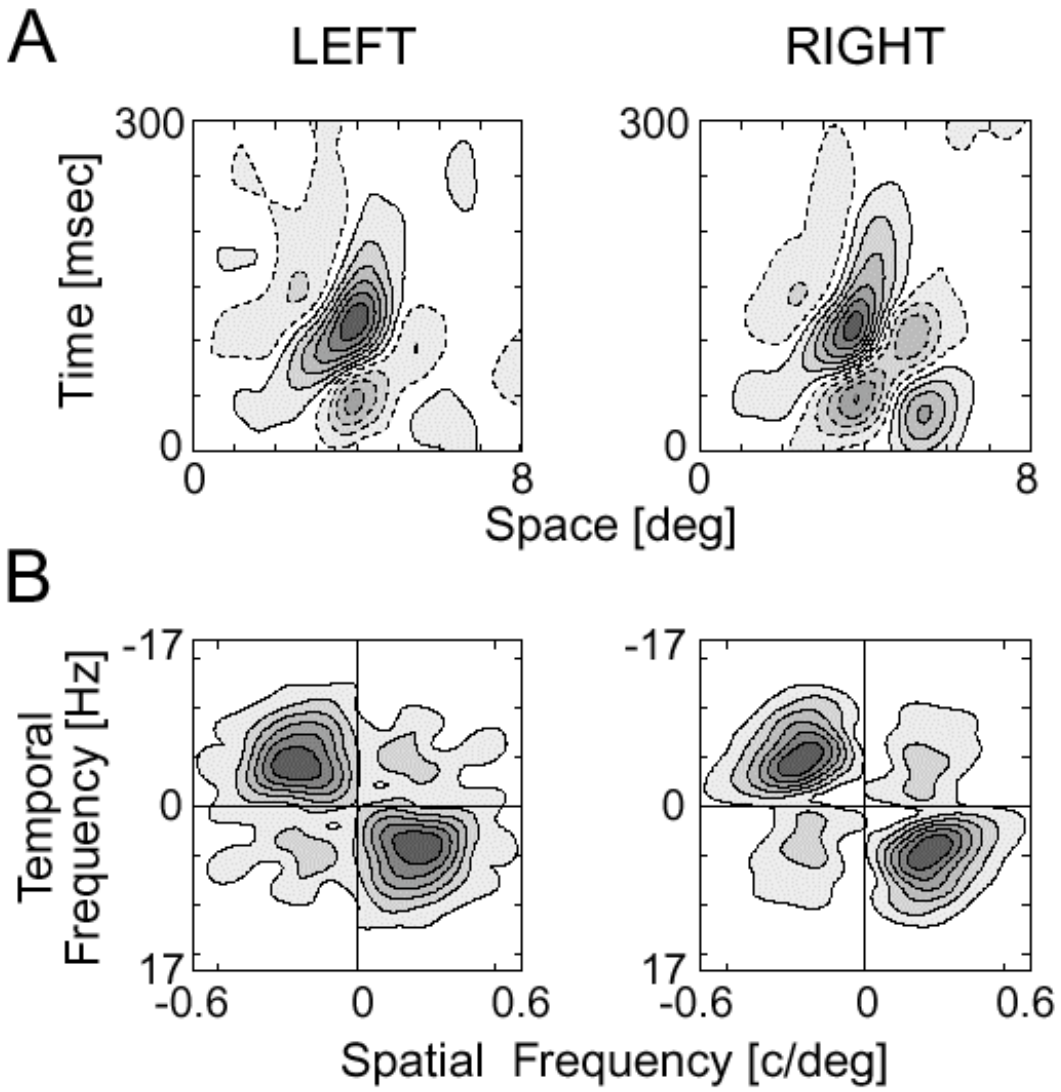


Fig. 13 Space-time RFs and their amplitude spectra in the spatio-temporal frequency domain are shown for a binocular simple cell. **A**: Space-time RFs are shown in the same format as that of Fig. 5. The RFs of this cell are clearly inseparable in the space-time domain as indicated by the tilted orientation of the subregions. **B**: Amplitude spectra for the two RFs are shown in the spatial and temporal frequency domain. The spectra are symmetric about the origin because of the nature of the Fourier transform. Positive and negative frequencies may be interpreted as representing responses to a sinusoidal grating moving in forward and reverse directions, respectively. A large amplitude difference for the peaks in the first and fourth quadrants indicates direction selectivity. Preferred orientations were 75 degs for both eyes. Bar stimuli used in reverse correlation mapping had dimensions of 2 x 0.75 degs . Stimulus duration was 52.8 msec (4 frames).

Summary of results

Approximately 30% of all neurons studied showed differences in their spatial RF for the two eyes.

Of these, nearly all prefer orientations between oblique and vertical; hence could be involved in processing horizontal disparity.

Conversely, most cells found with horizontal preferred orientation showed no RF difference between eyes.

RF properties change over time, but in a similar way for both eyes.

Main issue with local features

The depth map inferred from local features will not be complete:

- missing information in uniform image regions
- partial occlusions (features seen in one eye but occluded in the other)
- ghosts and ambiguous correspondences
- false matches due to noise

typical solution: use a regularization process to infer depth in regions where its direct computation is unclear, based on neighboring regions where its computation was unambiguous.

The Dev Model

Example of depth reconstruction model that includes a regularization process: Arbib, Boylls and Dev's model.

Regularizing hypotheses:

- the scene has a small number of continuous surfaces.
- at one location, there is only one depth

So,

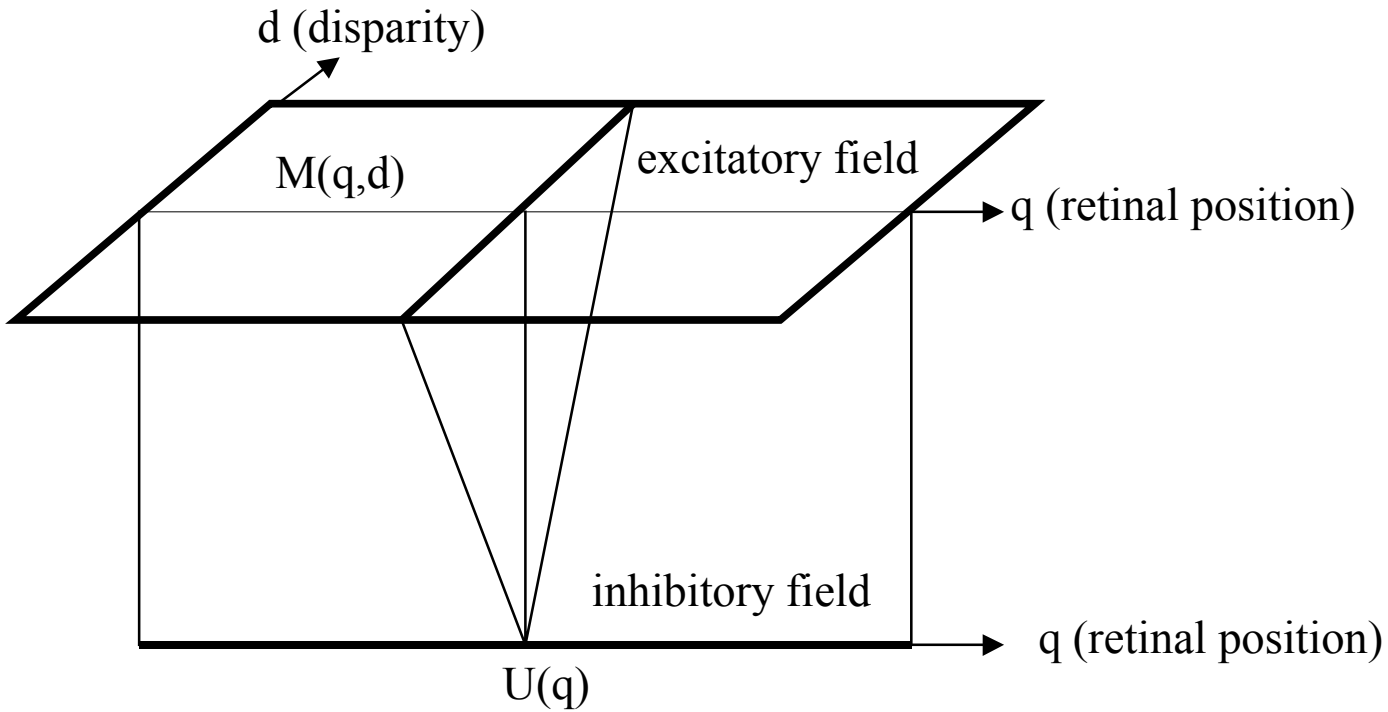
- depth at a given location, if ambiguous, is inferred from depth at neighboring locations
- at a given location, multiple possible depths values compete

The Dev Model

consider a 1D input along axis “q”; object at each location lies at a given depth, corresponding to a given disparity along the “d” axis.

along q: cooperate & interpolate through excitatory field

along d: compete & enforce 1 active location through winner-take-all



Regularization in Biology

Regularization is omnipresent in the biological visual system (e.g., filling-in of blind spot).

We saw that some V1 cells are tuned to disparity

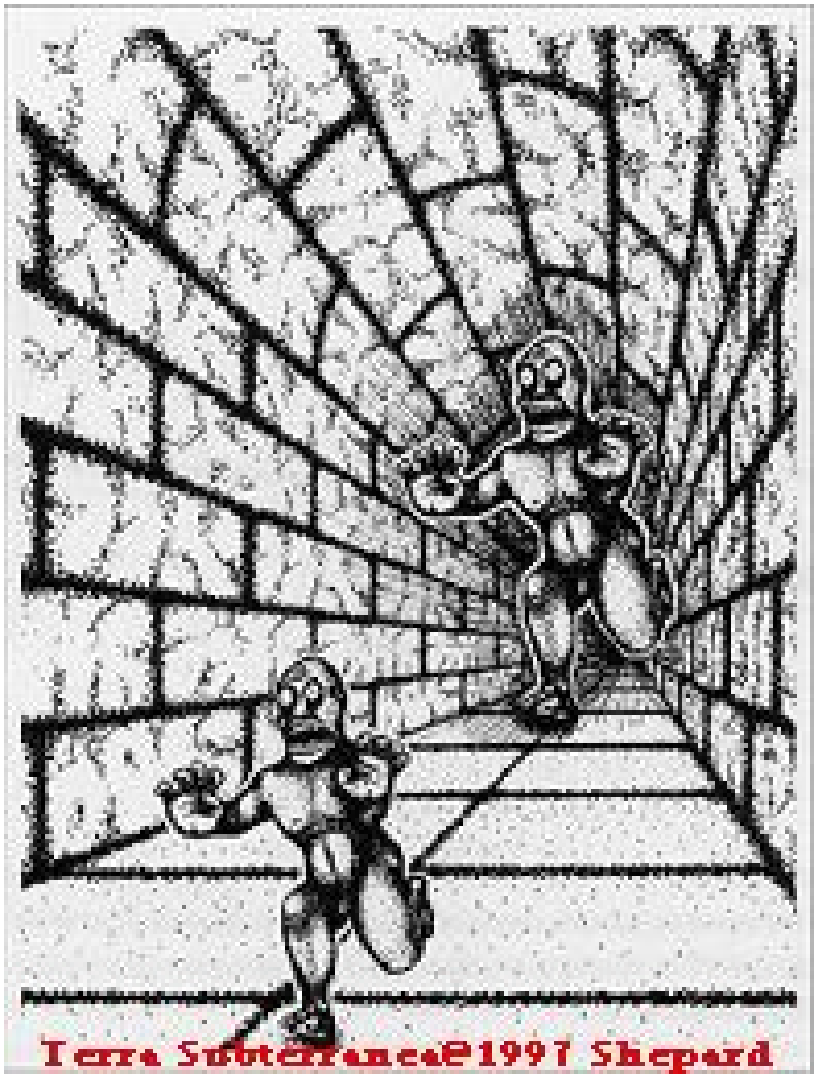
We saw (last lecture) that long-range (non-classical) interactions exist among V1 cells, both excitatory and inhibitory

So it looks like biology has the basic elements for a regularized depth reconstruction algorithm. Its detailed understanding will require more research ;-)

Low-Level Disparity is Not The Only Cue

as exemplified by size constancy illusions:

when we have no disparity cue to infer depth (e.g., a 2D image of a 3D scene), we still tend to perceive the scene in 3D and infer depth from the known relative sizes between the various elements in the scene.



More Biological Depth Tuning

Dobbins, Jeo & Allman, Science, 1998.

Record from V1, V2 and V4 in awake monkey.

Show disks of various sizes, on a computer screen at variable distance from animal.

Typical cells:

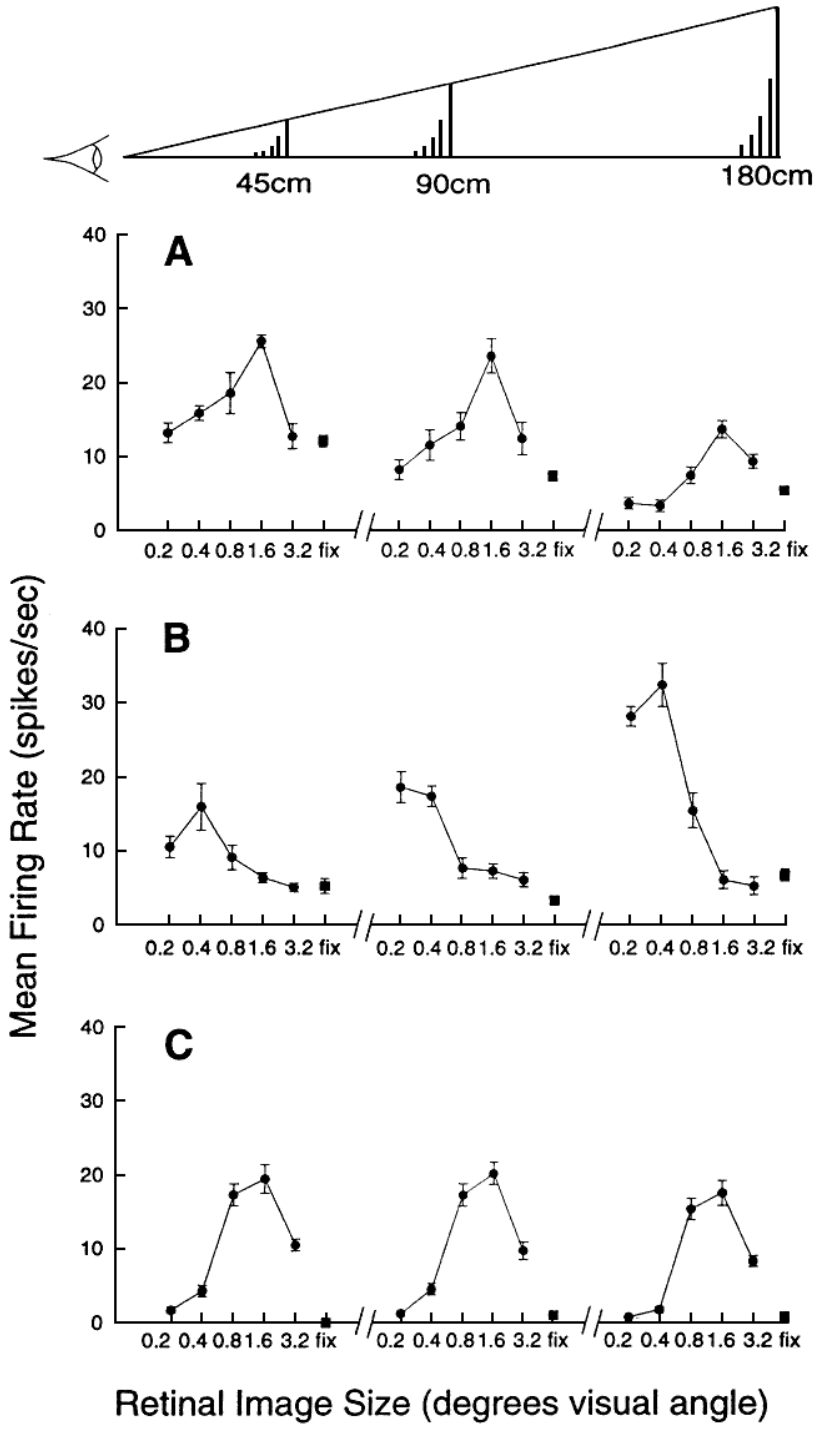
- are size tuned, i.e., prefer the same retinal image size regardless of distance;
- but their response may be modulated by screen distance!

Distance tuning

A: nearness cell (fires more when object is near, for same retinal size);

B: farness cell

C: distance-independent cell



Outlook

Depth computation can be carried out by inferring distance from disparity, i.e., displacement between an object's projection on two cameras or eyes

The major computational challenge is the correspondence problem, i.e., pairing visual features across both eyes

Biological neurons in early visual areas, with small RF sizes, are already disparity-tuned, suggesting that biological brains solve the correspondence problem in part based on localized, low-level cues

However, low-level cues provide only sparse depth maps; using regularization processes and higher-level cues (e.g., whole objects) provides increased robustness