# *Lecture 19.*
# *Depth Perception*
# *Reading Assignments: TMB2 7.1*

## Fine Grain Parallelism

In vision, such low-level processes as

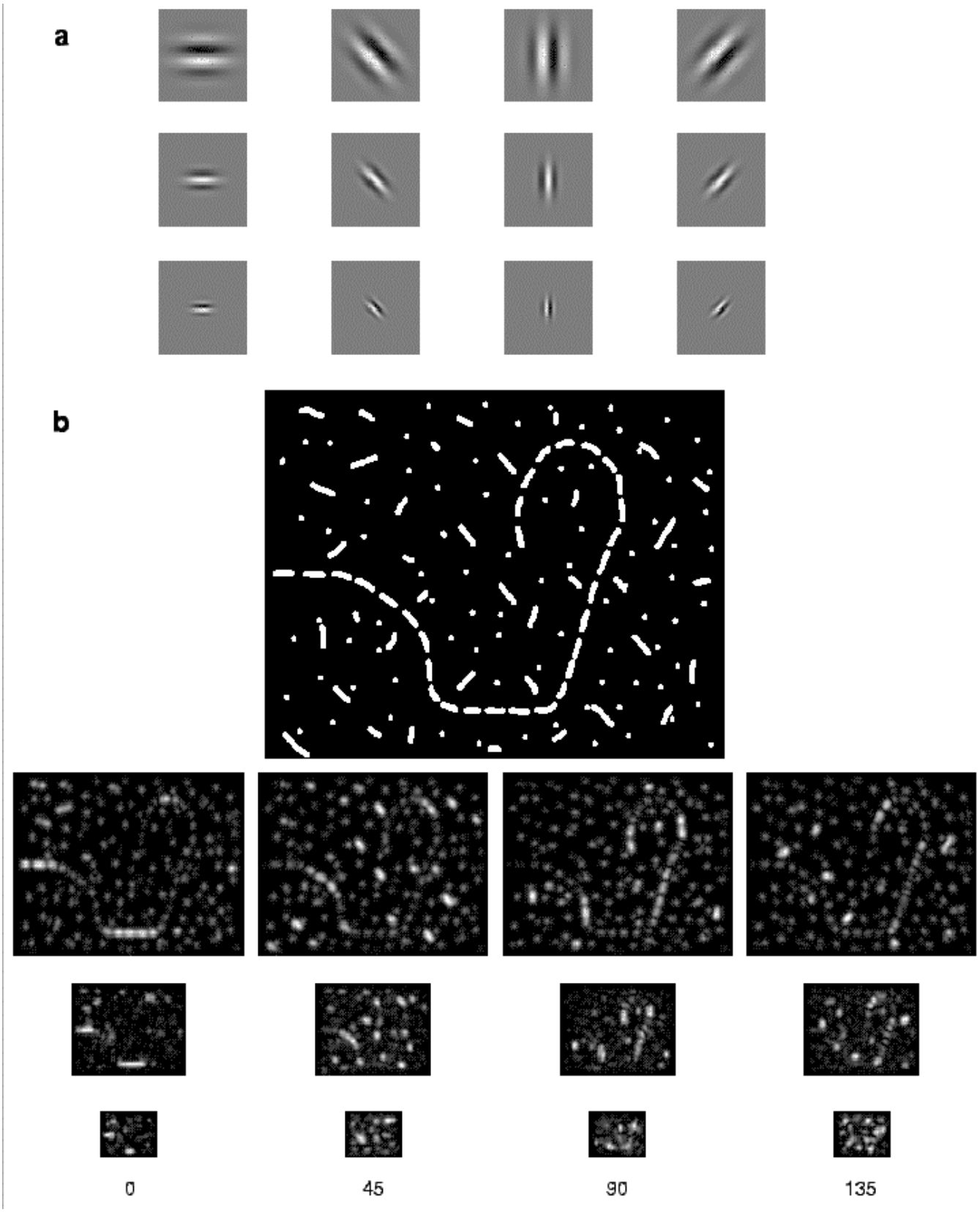
segmentation,

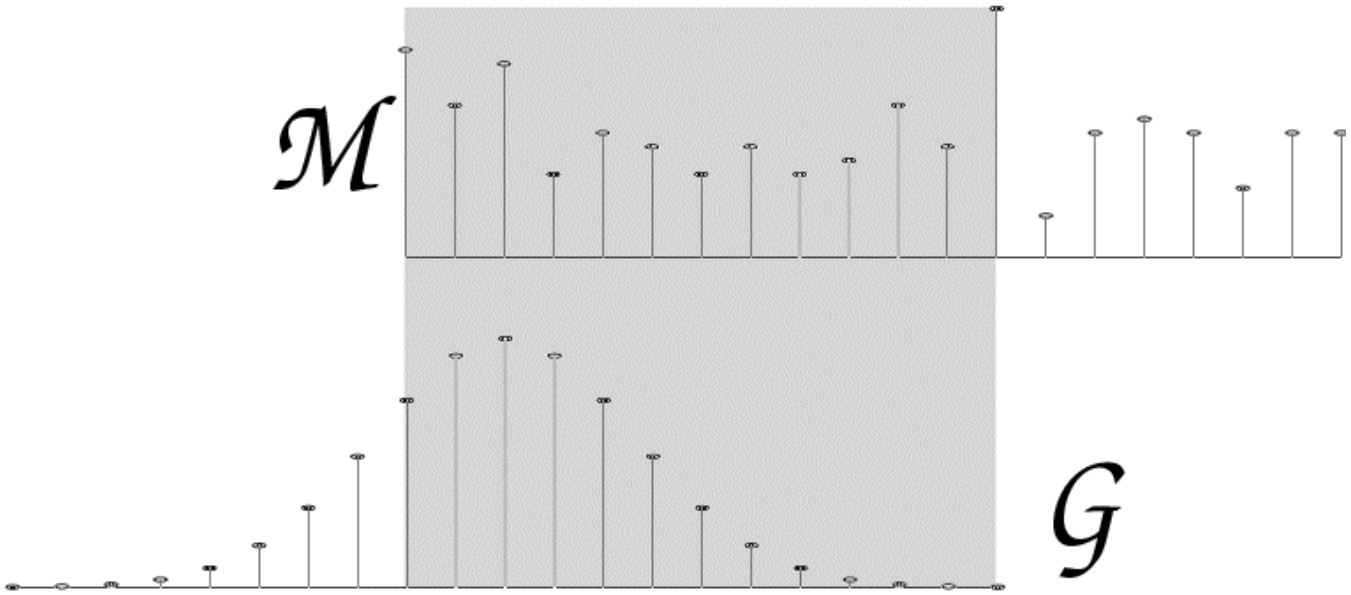
depth mapping and

texture extraction

involve an array of identical processes applied over an entire image: they iterate back and forth until the desired transformation of the image is completed.

Similarly, in our neural studies, we analyze regions of the brain in terms of a small number of layers, each layer being made up of a large number of similar neurons.


a
b
0
45
90
135

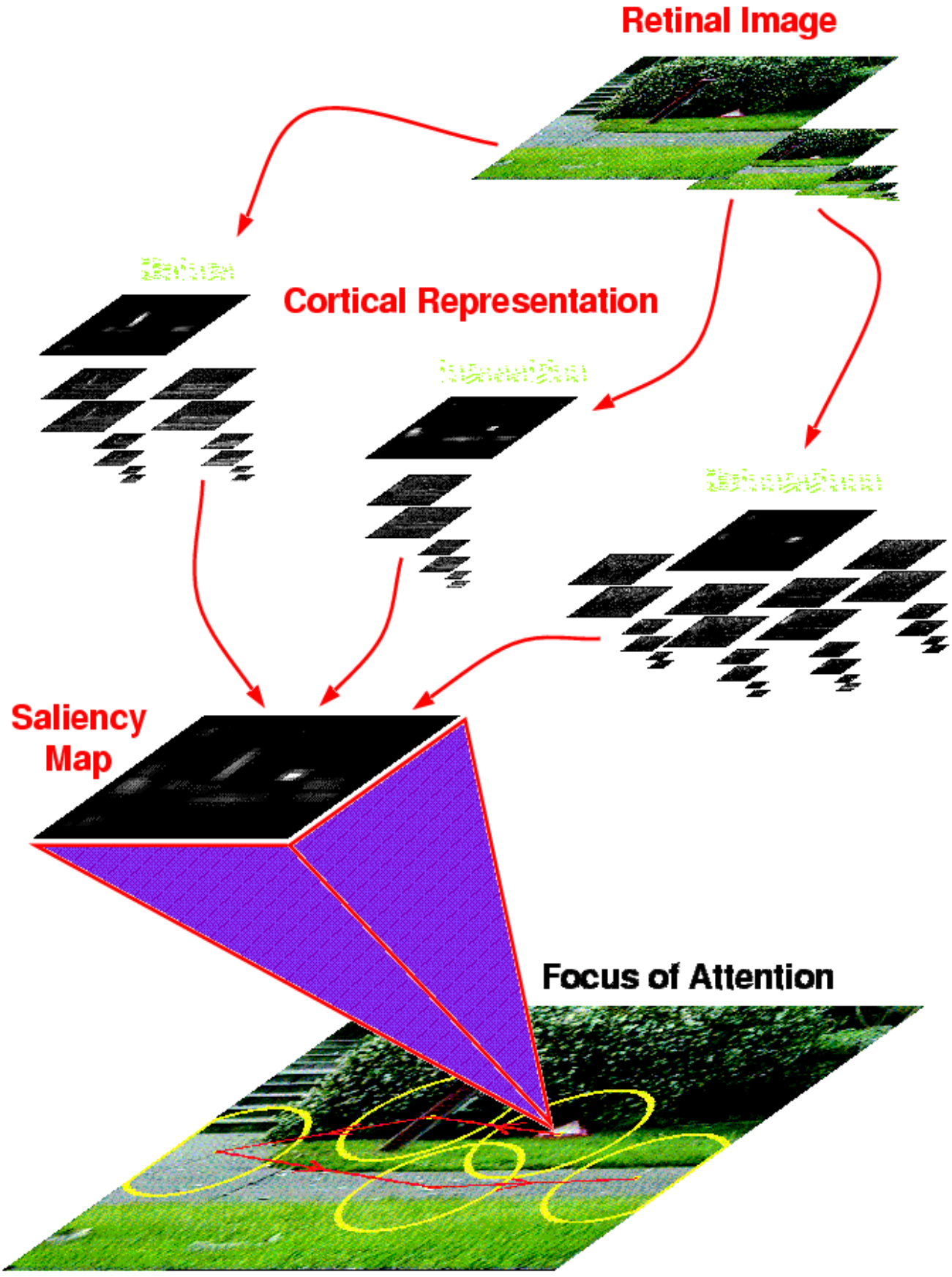



A key concept in layered visual computation is retinotopy: Due to overlap of receptive fields, retinotopic "maps" are not simple point-by-point transmissions of arrays of stimulation:

- "bug detectors" and "enemy detectors" in frog retina
- "edge detectors" in the visual cortex of cats and monkeys

These maps provide input variables for

- controlling behavior and
- perceiving aspects of the world


Retinal Image
Cortical Representation
Saliency Map
Focus of Attention

## Brain structure

Distinguish

- retinotopic representations in parts of the brain from
- the more abstract representations which seem to be associated with object recognition

## Machine vision

- Low-level vision:

  processing done to recode information without using object knowledge, using parallel array processing
- High-level vision: knowledge-intensive processes

  better modeled by distributed processing.

We here study basic "modules" for low-level vision: those for stereopsis and depth perception more generally.

# Depth Perception

One candidate for a natural functional module which subserves part of the task of vision is **depth perception**

which enables us to analyze the world in terms of aspects located at various distances from us.

From a single eye, we can determine the direction in space of various features of the world.

To locate the features in depth:

- stereopsis uses cues provided by correlating the visual input to two spatially separated eyes
- optic flow uses the information provided to the eye at moments separated in time
- accommodation determines what focal length will best bring an object into focus

A machine vision system can also make use of direct measurements, e.g., laser or sonar rangefinders.

**www-sop.inria.fr/robotvis/demo/f-http/html/**

will compute the stereo correspondence between two images

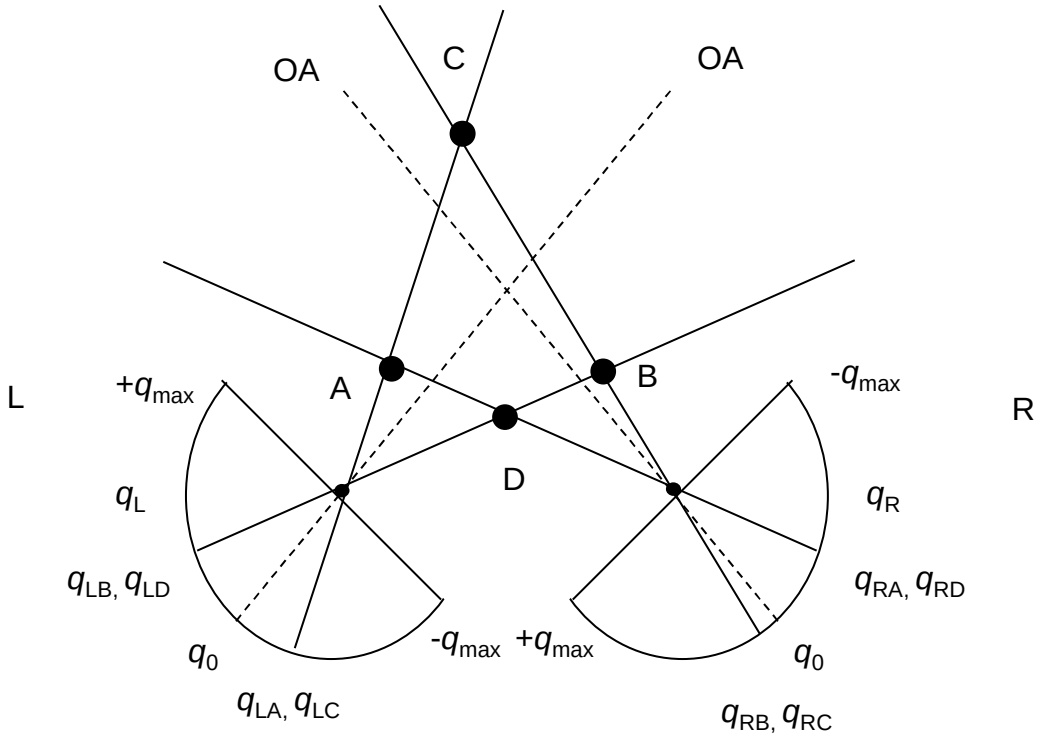
# Stereopsis

Here we concentrate on stereoposis,

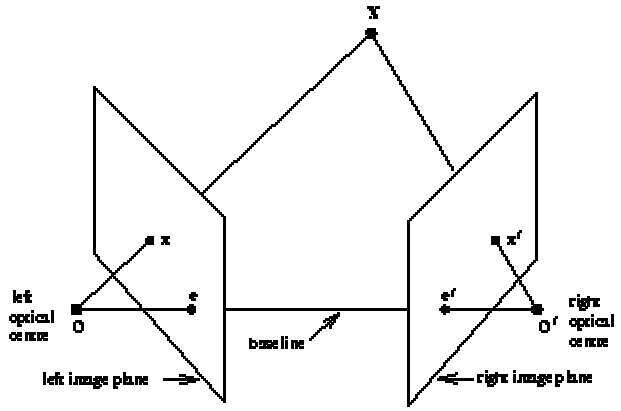
and then see how accomodation may be used cooperatively to improve the depth estimates provided by a stereo-based system.

The Correspondence Problem: To match corresponding points on the 2 retinas to be able to infer depth of the stimulus of which both features are the image.

Why is it a problem? "Ghosts"

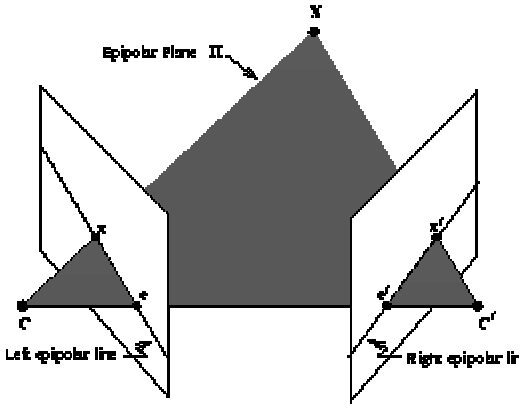


# Epipolar Geometry




baseline: line joining both eye's optical centers

epipole: intersection of baseline with image plane




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.*

# What is the challenge then?

Finding correspondences between image locations, under the constraint of epipolar geometry.

# A classic question in Visual Psychology:

Does depth perception comes before or after pattern recognition?

- 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,

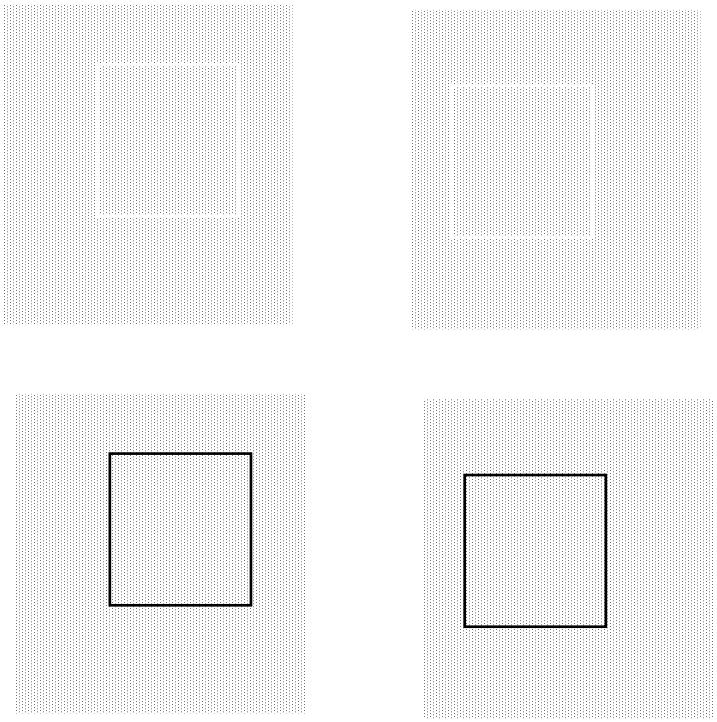
  thus building up a depth map of surfaces and small objects in space which provides the input for perceptual recognition?

## Cyclopean Perception

Bela Julesz 1971 developed the method of

**random dot stereograms**

stereo pairs each of which contains only visual noise but so designed that the visual noise was correlated.



Here I have used regular dot patterns to generate the stereo pair (shown above) - the edges (shown below) are *not* part of the stereo pair but inserted to show the mode of construction.

In *random dot* stereograms each of the "pieces" has dots placed randomly.

Julesz found that human subjects were in fact able to solve the correspondence problem, carrying out the appropriate matching to see surfaces stippled with random patterning at varying depths in space

The formation of a depth map of space could precede the recognition of pattern.

## Neural Network Solutions

Julesz offered a model of this process in terms of cooperative computation involving an array of magnetic dipoles connected by springs — anticipating the current interest of physicists in **spin glasses** as a statistical mechanics analog of neural nets. (cf. §4.3 and Chap. 2 of Hertz, Krogh and Palmer)

For the brain theorist the issue was thus raised:

"Could the depth map be computed by a cooperative process involving realistic neurons?"

In 1971, data on what constituted realistic neurons was provided by the work of Barlow et al 1967.

More current data are reviewed by Poggio & Poggio 1985.

Arbib, Boylls and Dev 1974 built a neural net cooperative computation model ("the Dev Model") for building the depth map

"guided by the plausible hypothesis that our visual world is made up of relatively few connected regions".

# The Dev Stereopsis Model

x: the coordinate on the left retina

x + d: the coordinate on the right retina - d is disparity

## Input to the excitatory array

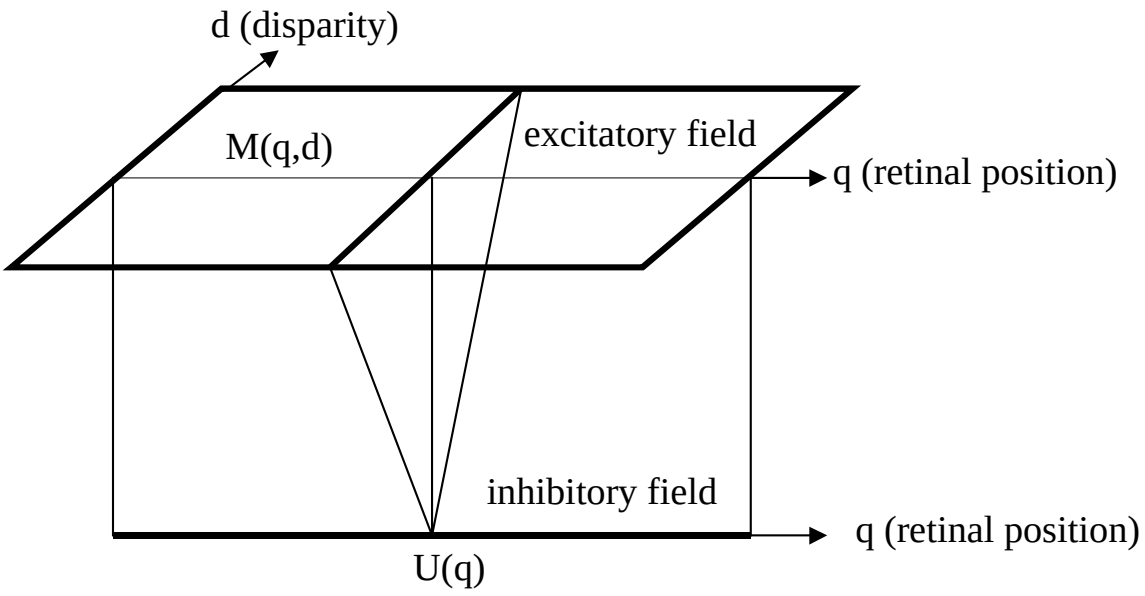
s(x,d) measures the correlation between retinal features at x on lefty retina and x+d on right retina.

## Key idea

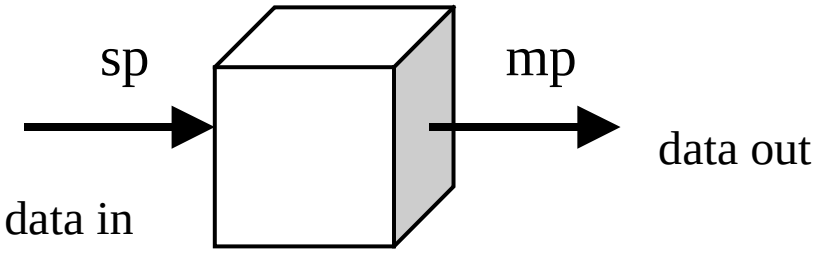
- Use a maximum selector for each x

  neurons corresponding to the same visual direction but different depths are connected via mutual inhibition

- But, to favor surface formation, provide excitatory coupling for similar features at nearby visual directions and approximately equal depths.



The simulation in the NSL book is based on the **Dev** disparity module. It contains an input port **sp** that receives external data and an output port **mp** that generates output data:




Dev dev

$$\tau_m \frac{\partial m_{ij}}{\partial t} = -m_{ij} + w_m * f\left(m_{ij}\right) - w_u * g\left(u_j\right) - h_m + s_{ij}$$

$$\tau_u \frac{\partial u_j}{\partial t} = -u_j + \sum_i f\left(m_{ij}\right) - h_u$$

$$f\left(m_{ij}\right) = step\left(m_{ij}, k\right)$$

$$g\left(u_j\right) = ramp\left(u_j\right)$$

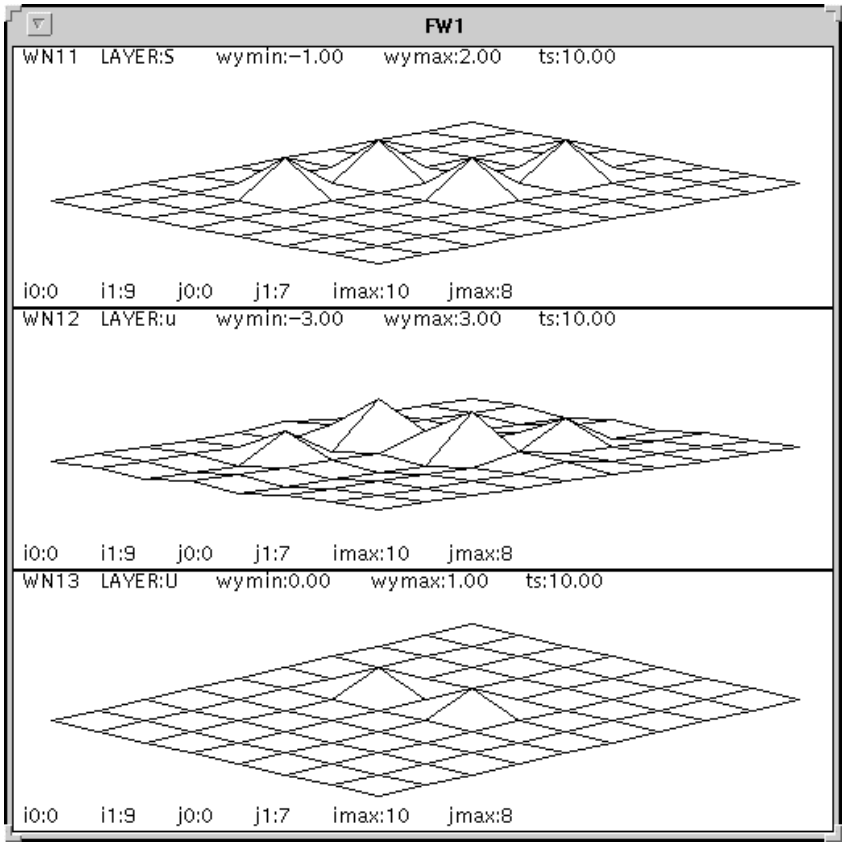
Input $s$ is computed by calculating disparity between left $r_L$ and right $r_R$ retina mappings. $r_L(q)$ is set to 1 if some object projects to point $q$ on the left retina, $r_L(q)$ is set to 0 otherwise; and similarly for $r_R(q)$. Stereo input is then defined as

$$s_d(q) = R_L(q)\, R_R(q+d)$$

which is 1 only if there is an object at position $q$ on the left retina as well as at $q+d$ on the right retina, and is otherwise 0.

(Note that a more subtle version of the model would require similar local features, rather than mere presence of an object, at $q_L$ and $(q+d)_R$.)

An overview of the suppression of ghost targets:

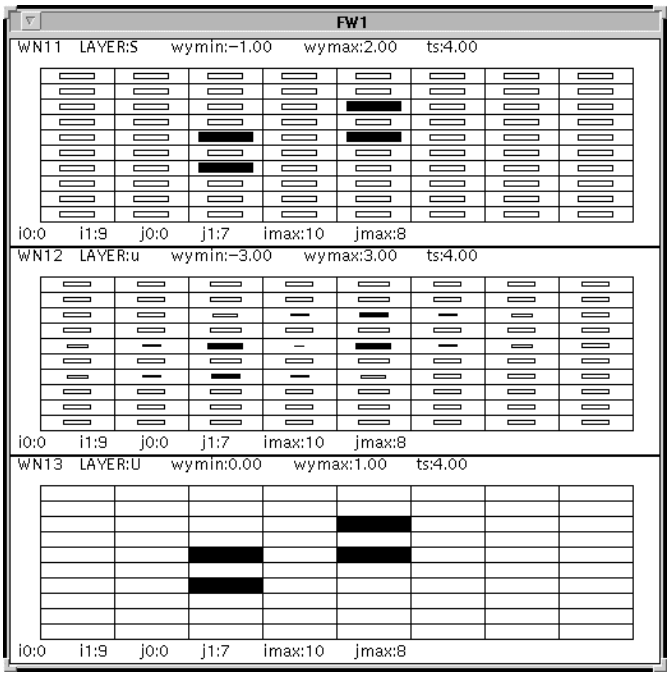


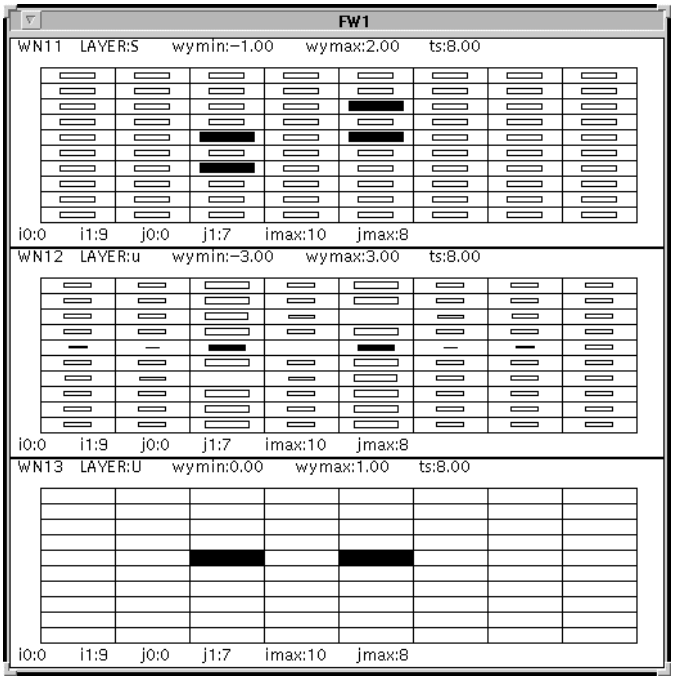

The Dev model favors targets A and B as the "real" targets, and exorcises C and D as "ghost targets", even the retinal data were neutral as to the choice of (A,B) versus (C,D).

This is because the design of the Dev model meets the constraint that the world is made up of surfaces, and thus favors a choice consistent with nearby points of similar disparity over other choices.

The House model (see below) exploits accommodation as well as disparity cues.

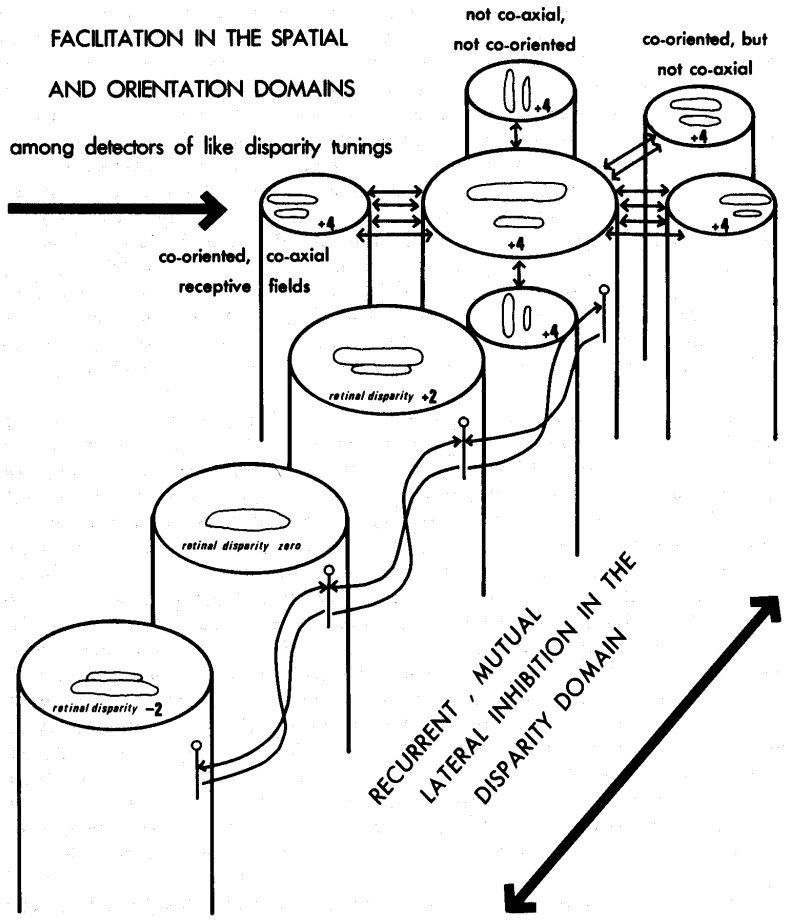
You can check that if C and D are the "real" inputs, then the new model will verify this, whereas the Dev model will not.


FW1
WN11 LAYER:S wymin:−1.00 wymax:2.00 ts:4.00
i0:0 i1:9 j0:0 j1:7 imax:10 jmax:8
WN12 LAYER:u wymin:−3.00 wymax:3.00 ts:4.00
i0:0 i1:9 j0:0 j1:7 imax:10 jmax:8
WN13 LAYER:U wymin:0.00 wymax:1.00 ts:4.00
i0:0 i1:9 j0:0 j1:7 imax:10 jmax:8


FW1
WN11 LAYER:S wymin:−1.00 wymax:2.00 ts:8.00
i0:0 i1:9 j0:0 j1:7 imax:10 jmax:8
WN12 LAYER:u wymin:−3.00 wymax:3.00 ts:8.00
i0:0 i1:9 j0:0 j1:7 imax:10 jmax:8
WN13 LAYER:U wymin:0.00 wymax:1.00 ts:8.00
i0:0 i1:9 j0:0 j1:7 imax:10 jmax:8

Sperling 1970 and Nelson 1975 gave related models.

Nelson's model took advantage of edge detectors.




Later, a variant of the Dev model was published by Marr and Poggio 1977.

In subsequent writings Marr took the plausible hypothesis that our visual world is made up of relatively few connected regions, and showed how it could be developed into an elegant mathematical theory relating the structure of a depth perception algorithm to the nature of surfaces in the world.

Much further work on stereopsis algorithms has followed:

HBTNN: Stereo Correspondence and Neural Networks

# In machine vision

computation of depth from stereo images usually proceeds in two steps:

1) establishment of a correspondence between selected points found, e.g., using an interest operator in the two images to provide a disparity map followed by the computation of the depth at the points; and

2) application of some type of interpolation process which extends the depth map to all points in the images followed by the discovery, extraction, and description of the surfaces in the 3-dimensional scene.

## The problem of false minima

Consider a picket fence:

An initial mismatch could co-opt the possible choices of neighbors, and end up with a high confidence estimate that the fence was at a different depth from that at which it actually occurred to provide a false local

energy minimum

Need algorithms that avoid (some of) these false minima.

## Marr and Poggio 1979

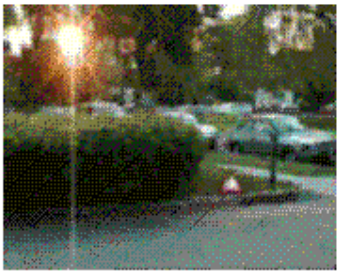
used

- the idea of spatial frequency channels; and
- the idea of pyramids or processing cones

1. With hardly any cooperative computation, a fairly confident rough depth estimate for different surfaces could be made using the low spatial frequency channel.

2. This rough model was then used to control vergence eye movements, sculpting a more detailed spatial map on the first approximation through disparity information provided via channels of higher spatial frequency.

[cf. the two stages of map formation for self-organizing feature maps - but there it was on a "developmental" time schedule.]





## Mayhew and Frisby 1980

conjectured that the matching processes are integrated with the construction of a primitive binocular-based description of image intensity changes.

They then offered an algorithm, STEREOEDGE:

1. The intial stages used local piecewise binocular grouping of adjacent, similar, **zero crossings** or peak matches, while

2. Later stages used a relaxation process (Section 4.2).

The earlier stages were akin to Marr-Poggio, while the later stages exhibited the cooperative processing that Marr and Poggio 1979 sought to exclude.

## Prazdny 1987

rejected the Marr and Poggio 1979 proposal that

"false matches may be avoided by trading off resolution for disparity range using a coarse-to-fine matching strategy."

He notes that low and high spatial frequencies are often informationally orthogonal.

*For example, if grass is viewed through a picket fence, there is no reason why the disparities of the fence should be related in any way to the disparities of the grass surface.*

Prazdny offered a specific laboratory test using a random-dot stereogram in which the background plane is transparent,

and where two depths, one from low and one from high spatial frequencies, can be observed simultaneously.

He concludes that patches of the visual field may be fused and then held "locked" by some form of hysteresis as proposed by Julesz 1971.

## Cooperation between Stereopsis and Accomodation

In the Dev scheme,

- those neurons which represent similar features at nearby visual directions and approximately equal depths excite each other, whereas

- those neurons which correspond to the same visual direction but different depths are (via interneurons) mutually inhibitory.

In this way, neurons which could represent elements of a surface in space will cooperate, whereas those which would represent paradoxical surfaces at the same depth will compete. The result is that, in many cases, the system will converge to an adequate depth map.

However, such a system may need extra cues.

Recall the paling fence example.

In animals with frontal facing eyes, such ambiguity can be reduced by the use of vergence information to drive the system with an initial depth estimate.

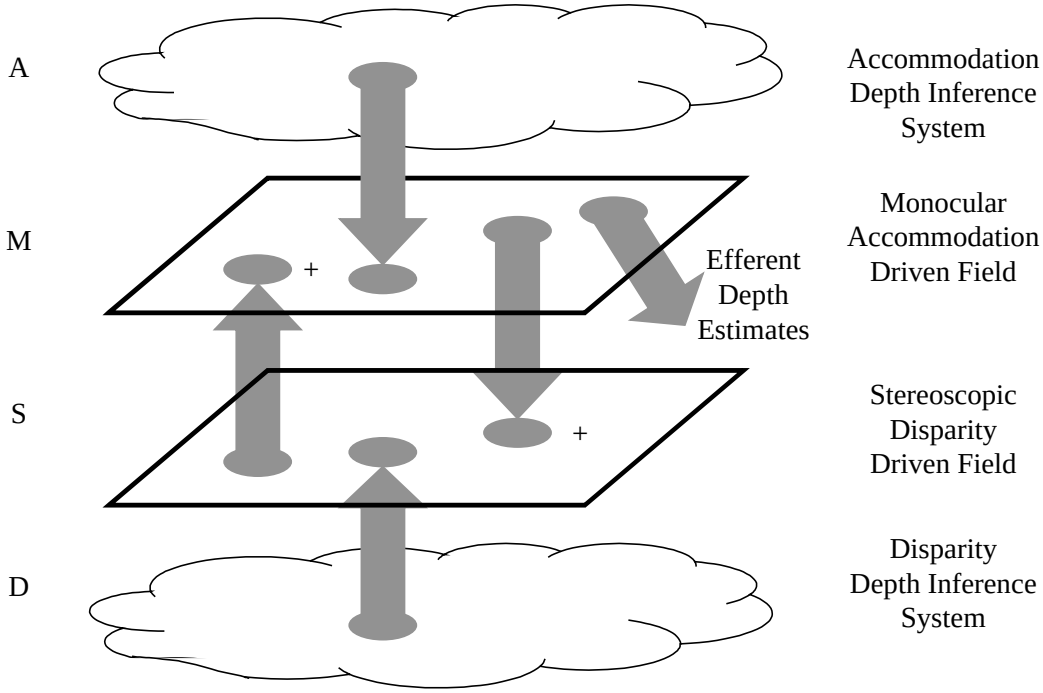
Another method is to use accommodation information to provide the initial bias for a depth perception system; this is more appropriate to the amphibian, with its lateral-facing eyes.

Ingle 1976 had observed that a monocular frog can snap fairly accurately at prey in its monocular field.

Collett 1977 used experiments with prisms and lenses placed in front of the eyes of the toad to show that, in its binocular field, the toad relied mainly on stereopsis, but that the monocular toad did make depth judgments based on accommodation.

## The Cue Interaction Model

(House 1984) uses two systems to build a depth map, each based on Dev's stereopsis model:




S (for STEREO) is driven by disparity cues,

M (for MONO) by accomodation cues,

M's input is provided by an accomodation measure which represents the sharpness of the image in direction x when the focal length of the left eye is set at d.

[Seems to need a working memory.]

• Corresponding points in the two maps have excitatory cross-coupling.

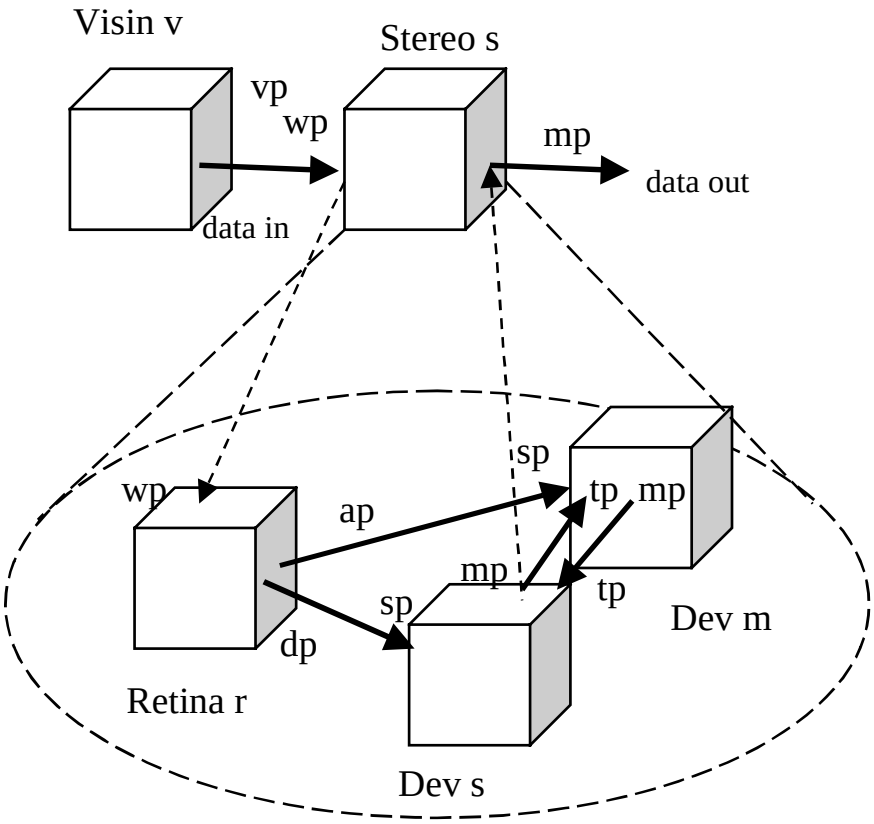
The model is so tuned that binocular depth cues predominate where available, but monocular accomodative cues remain sufficient to determine depth in the absence of binocular cues.

Since we explicitly designed the model, we know that one layer represents accommodation information, while another represents disparity.

However, during the ongoing behaviour of the system, both surfaces represent pooled information based on the interaction between the layers, rather than representing information directly supplied by sensory systems.

This clearly indicates the dangers of experimentation based on feature analysis without related high-level modelling.

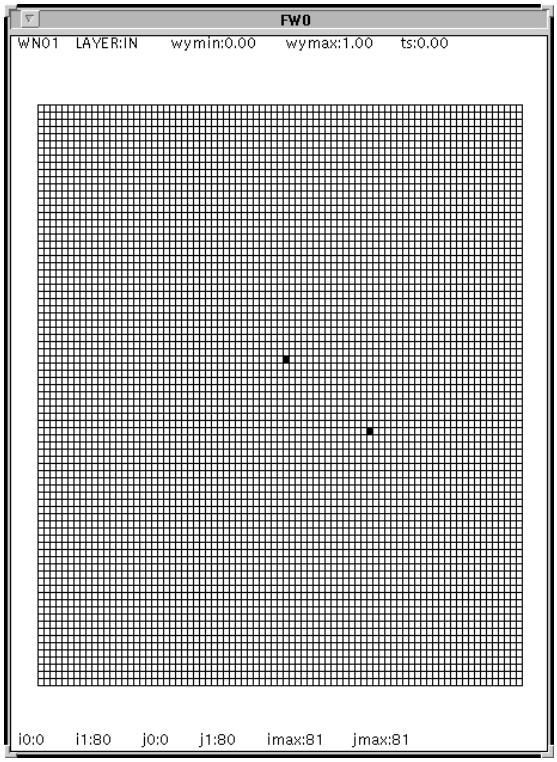
## Hierarchical View of NSL Modules in the Implementation

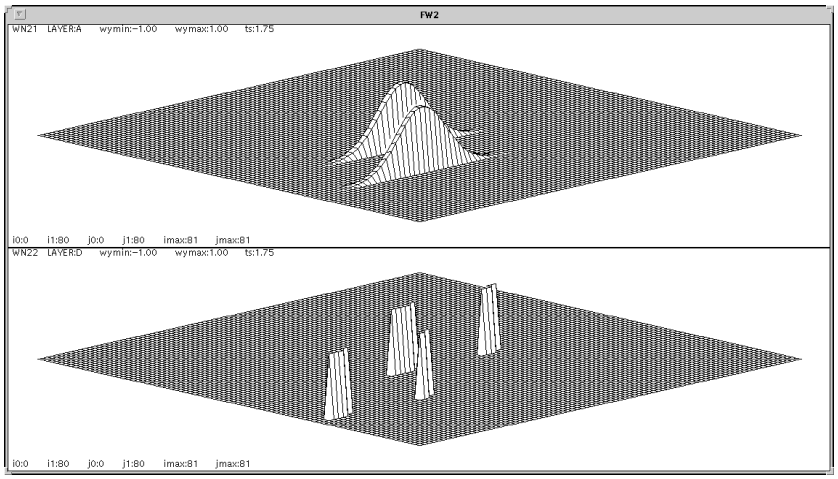



**Visin** is the input module

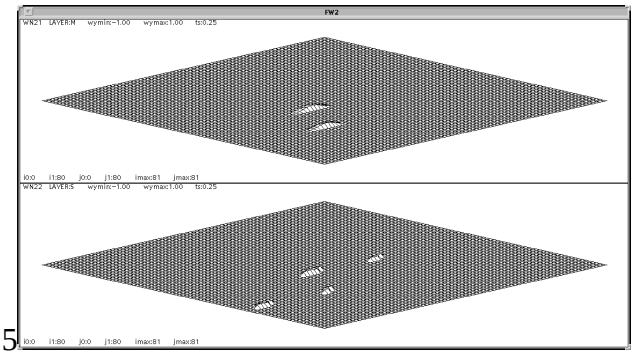
**Stereo** assemblage comprises a **Retina** module, and two instances of a (slightly modified) **Dev** module

The new **Dev** module is implemented by neural networks similar to those defined in the original Dev module. These equations are extended to enable cross-coupling between Dev modules

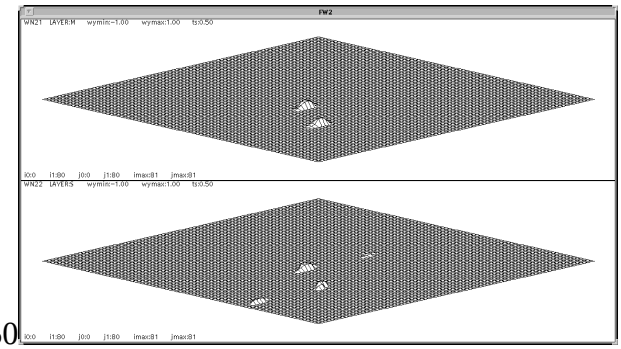


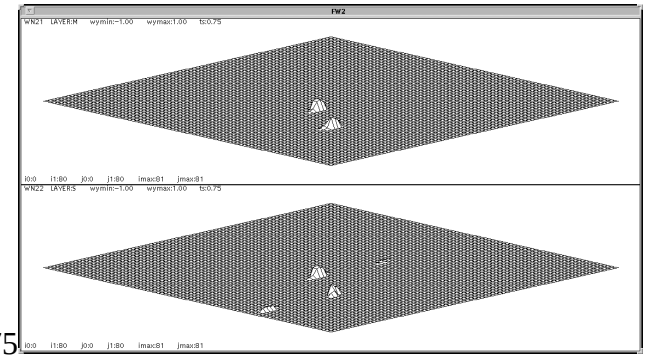

World input for processing: Input array *in* and output arrays *a* and *d*, from the **Retina**.



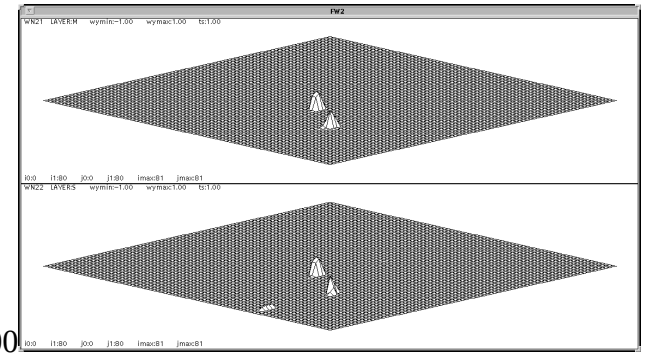
0.25



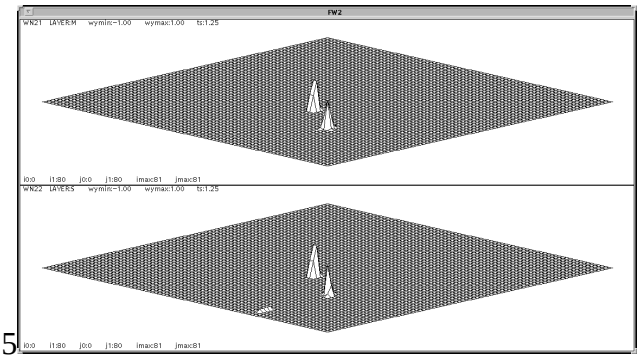
0.50



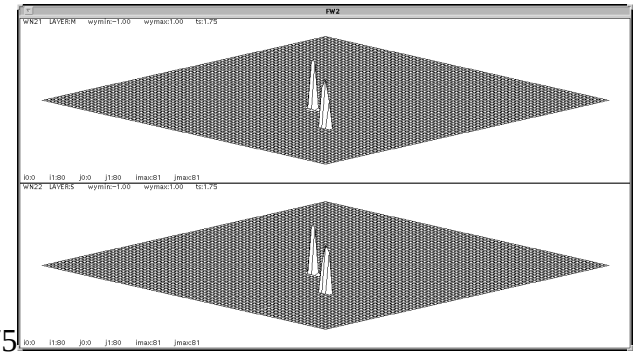
0.75



1.00




1.25




1.75

## Conclusion

We have shown how to take advantage of the features in NSLM to modularly extend the original Dev disparity model into the House depth perception model. The ability to extend models makes NSL a very powerful simulation language.