Computational Vision
U. Minn. Psy 5036
Daniel Kersten
Lecture 14: Intro to Scenes from Images

Initialize

Spell check off

```
SetOptions[ArrayPlot, ColorFunction → "GrayTones",
  DataReversed → True, Frame → False, AspectRatio → Automatic,
  Mesh → False, PixelConstrained → True, ImageSize → Small];
SetOptions[ListPlot, ImageSize → Small];
SetOptions[Plot, ImageSize → Small];
SetOptions[DensityPlot, ImageSize → Small, ColorFunction → GrayLevel];
nbinfo = NotebookInformation[EvaluationNotebook[]];
dir = ("FileName" /. nbinfo /. FrontEnd`FileName[d_List, nam_, ___] :> ToFileName[d]);

Off[General::spell1];
```

Outline

Last time: Continued with discussion of the two views of the function of early visual coding

How should we think about the functions of V1 neuronal selectivites? In particular, the spatial filtering properties.

Efficient coding vs. edge/bar detection

--Efficient coding means fewer bits required to encode image

Examples: PCA->dimension reduction->quantization. Decorrelates filter outputs.
Filters localized in space and spatial frequency do too (e.g. wavelets).
Sparseness--high kurtosis histograms for filter outputs

--Edge/bar detection: local image measurements that correlate well with useful surface properties

Problems with edge detection

Noise & scale

Various scene causes can give rise to identical image intensity gradients
--no local information to "disambiguate" an edge

Alternative interpretations to V1's role in spatial processing?

Representations of texture/texture boundaries? saliency? Do we have a complete story of V1 selectivities? E.g. so-called "dictionary methods".

Today

Next homework

more filtering

Mathematica Demonstrations

Mathematica Demonstrations Illusions

Retina to V1 overview and review

Extrastriate cortex--overview

From images to objects, scenes and actions
scene-based modeling of images

Retina to V1: Review of form & function

(There are a number of web-based overviews, for example: http://www.sumanasinc.com/webcontent/anisamples/neurobiology/visualpathways.html).

Overview of pathways from eye-to-cortex

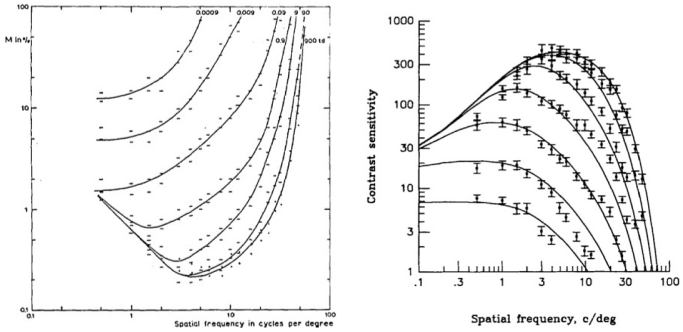
Roughly ten million retinal measurements are sent to the brain each second, where they are processed by some billion cortical neurons.
The primate retina has about 10^7 cones that send visual signals to the optic nerve via about 10^6 ganglion cells. The optic nerves from the two eyes meet at the optic chiasm where about half of the fibers cross over and the other half remain on the same side of the underside of the brain. Before synapsing in the lateral geniculate nucleus, about 20% of these fibers that make up the optic tract branch off to the superior colliculus--a structure involved with eye movements. Other fibers project to various other nuclei, but the majority of the optic tract fibers synapse on cells in the lateral geniculate nucleus. Cells in the lateral geniculate nucleus send their axons in a bundle called the optic radiation to layer IV (one of six layers) of primary visual cortex. A schematic representation of these pathways was shown in notes for an earlier lecture.

Retina

Much research has gone into understanding how gross behavioral sensitivity to contrast and spatial detail can be understood from retinal processing.

Spatial filtering

Earlier we noted that retinal ganglion cells have a characteristic center-surround organization with excitatory centers and inhibitory surrounds (or inhibitory centers and excitatory surrounds). We modeled the spatial output of the retina as a linear filter that convolves the input image with a kernel determined by the center-surround receptive field weights--a so-called single channel model, because the kernel is assumed to be the same shape and size at different locations. The spatial frequency bandpass characteristics of the retina are determined by just one kernel.



The left figure shows contrast thresholds for various light levels (from van Nes, & Bouman, M. A. (1967). Spatio modulation transfer in the human eye. J Opt Soc Am, 57(3), 401-406). The right figure is a replot of the left figure from: Atick, J. J., & Redlich, A. N. (1992). What does the retina know about natural scenes? Neural Computation, 4(2), 196-210. The solid lines show fits by Atick & Redlich based on an efficient coding model.

The retina's temporal processing can also be thought of as differentiation, but in time rather than space, and can be modeled as a band-pass temporal frequency filter (see Enroth-Cugell and Robson, 1966). Analogous to the spatial frequency selectivity, retinal ganglion cells pass the contrast of medium temporal frequencies more effectively than either low or high frequencies. For a retinal ganglion cell, contrast sensitivity as a function of temporal frequency is an inverted U, qualitatively similar to the spatial CSF. Humans are insensitive to temporal frequencies higher than the temporal cut-off (for humans about 50-80 Hz, depending on the mean light level). That is why TV frames (60 Hz interlaced) or computer displays (now usually >70 Hz) are not seen to be flickering. An extreme consequence of the low temporal frequency attenuation, is that an image that is held stationary on the retina disappears. A VLSI retina having similar spatial and temporal filtering properties was first built at Caltech by Mead and colleagues in the late 1980s (Mead, 1989).

At the retina, one begins to see evidence for multiple visual pathways for spatio-temporal information. In cats, ganglion X-cells have smaller receptive fields and poorer temporal resolution than Y-cells, suggesting that the X channel carries information important for fine spatial detail, and the Y-cell channel conveys coarse-scale spatial information quickly. There is a similar distinction in primates, the, so-called magno-cellular (homologous to Y-cells) and parvo-cellular (homologous to X-cells) cells and pathways.

Temporal filtering

Human temporal contrast sensitivity functions.

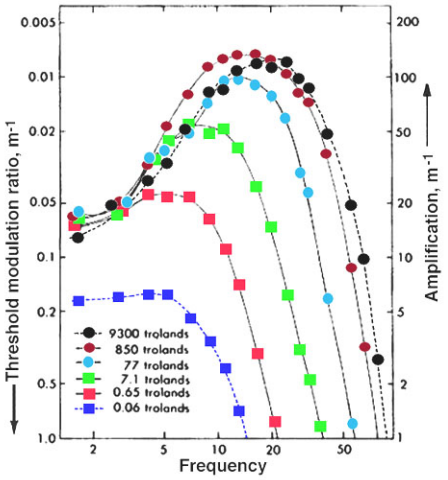
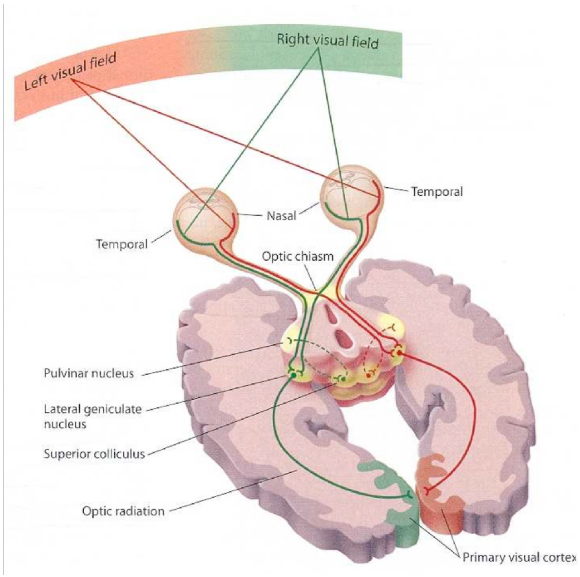


Fig. 11. Temporal Contrast Sensitivity Function (TSF) for various adapting fields. Kelly's data from Hart Jr, W. M., The temporal responsiveness of vision. In: Moses, R. A. and Hart, W. M. (ed) Adler's Physiology of the eye, Clinical Application. St. Louis: The C. V. Mosby Company, 1987.

But there may be much more to the information processing functions of the retina: Gollisch, T., & Meister, M. (2010). Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. *Neuron*, *65*(2), 150–164. doi:10.1016/j.neuron.2009.12.009

Functions of the optic chiasm and lateral geniculate nucleus (LGN)



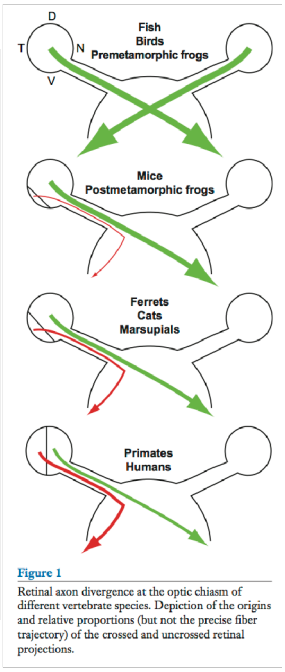


Figure 1
Retinal axon divergence at the optic chiasm of different vertebrate species. Depiction of the origins and relative proportions (but not the precise fiber trajectory) of the crossed and uncrossed retinal projections.

Figure on right from: Petros, T. J., Rebsam, A., & Mason, C. A. (2008). Retinal Axon Growth at the Optic

Chiasm: To Cross or Not to Cross. Annual Review of Neuroscience, 31(1), 295–315. doi:10.1146/annurev.neuro.31.060407.125609

The optic chiasm routes neuronal information so that information from corresponding points on the left and right eyes can come together at cortex for binocular vision, and in particular stereo vision. Typically animals with frontal vision have significant fraction of fibers that do not cross the midline, and animals with lateral eyes (e.g. fish) have few or no uncrossed fibers. In primates, the nervous system has gone to considerable length to bring information from the two eyes together early on. This suggests that certain kinds of cortical computations cannot easily be done "remotely", but require close connectivity between neurons, and the resulting topographic maps.

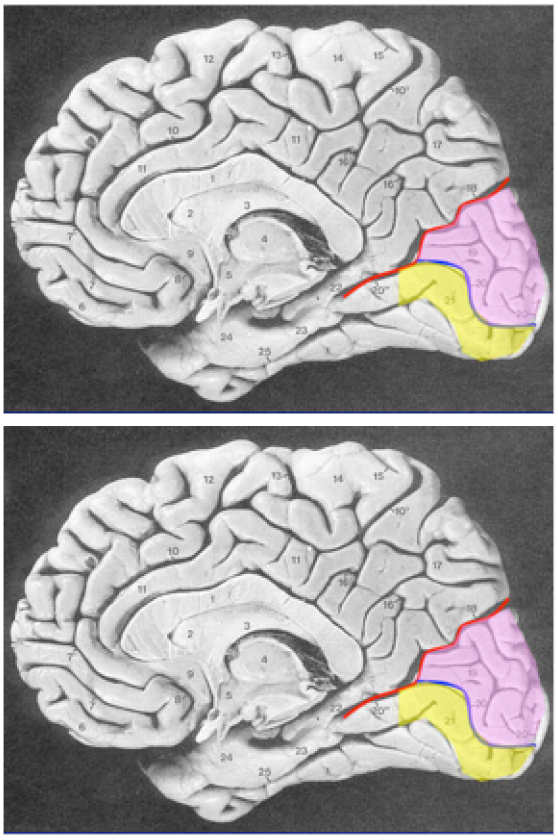
- How does the cross-over develop in infants?

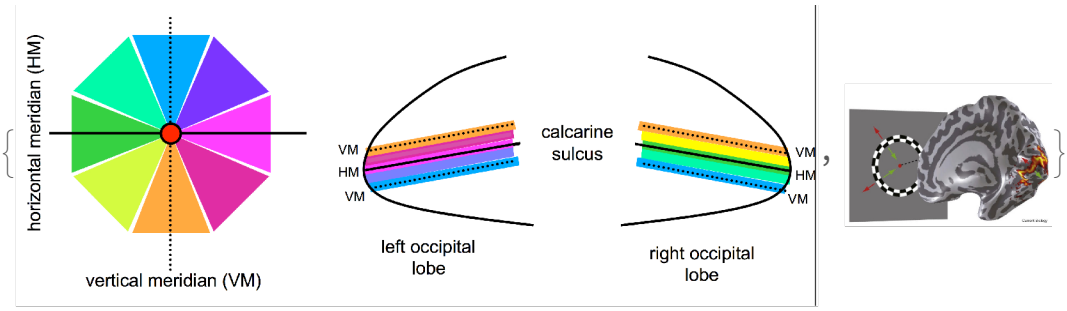
The neurons of lateral geniculate nucleus do more band-pass filtering, and the cells are characterized by fairly symmetrical center-surround organization like the ganglion cells. They show even less response to uniform illumination than ganglion cells. Despite the fact that neurons from the two eyes exist within the same nucleus, no binocular neurons are found in LGN. We have to wait until cortex to see binocular neurons. The X- and Y-cell division of labor continues with the so-called parvocellular (with corresponding retina input from P cells in monkeys, or X cells in cats), and the magnocellular (Y cells or M cells) pathways. Again the experimental measurements are consistent with the idea the the M pathway carries a fast, but coarse spatial representation of the image to the cortex, while the P pathway carries finer spatial detail but more slowly.

Although the LGN is often considered a relay station, feedback from cortex suggests possible role of attention mechanisms (see Crick, 1984 for a speculative neural network theory of LGN and reticular function; Mumford, 1991; Sillito et al., 1994). Although we will bypass a treatment of the superior colliculus, it has an important role is in the control of eye movements--a highly non-trivial problem requiring coordination of head and eye movements in the context of a constantly changing environment.

Anatomy and physiology of primary visual cortex

Large scale structure: V1 retinoptic map

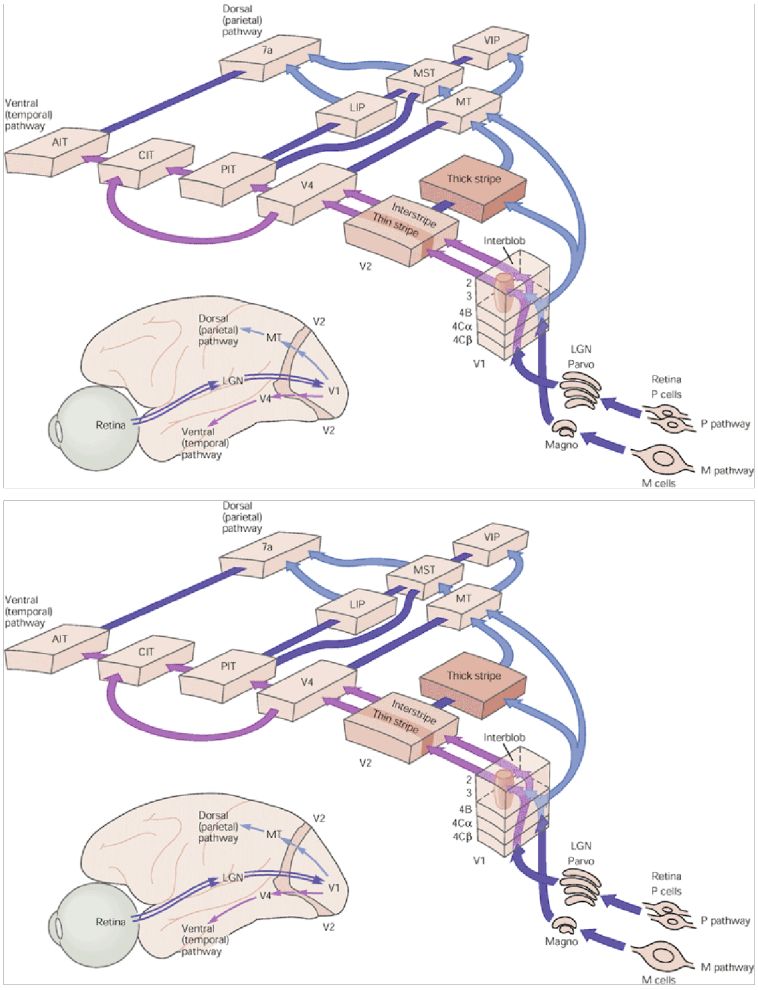




left visual field maps to right hemisphere
upper visual field maps to below the calcarine sulcus
horizontal meridian lies approximately along calcarine sulcus

Right figure from: Huk, A. C. (2008) Visual Neuroscience: Retinotopy meets Percept-otopy, Current Biology, 18, 21, R1005-1007.

Large scale structure: parallel pathways



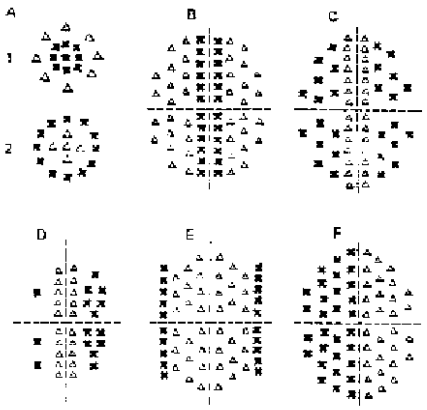
Neurons in the LGN send their axons (the optic radiation) to synapse on layer IV neurons of the primary visual cortex (also known as area 17 in cat, striate cortex or V1in monkeys and humans). Cortex is anatomically structured in layers, numbered from I (superficial) to VI (deep). The striate cortex is laid out as non-linear topographic map with 80% of cortical area devoted to about 20% of visual field, reflecting the higher acuity of foveal vision. Because of the cross-over at the optic chiasm, the left visual field (right retina) maps to right hemisphere. In monkey, many of the neurons in layer IV have receptive field properties similar to those in LGN.

However, in striking contrast with receptive field characteristics of earlier neurons, most cortical cells (other layers of V1) show:

- orientation selectivity
- spatial frequency selectivity, some with quite narrow tuning
- spatial phase selectivity (simple cells)
- binocularity
- motion selectivity

Apart from the spatial frequency selectivity, these properties were discovered in large part by the work over a couple of decades by Hubel, D. H., & Wiesel, T. N. (see 1968 reference). Hubel and Wiesel won the Nobel prize for this work. Below is a version of an earlier demonstration of local spatial filters tuned to spatial frequency and orientation.

Receptive field structure



29—7 Comparison of the receptive fields of neurons in the retina and in the lateral geniculate nucleus with those of simple cortical cells in area 17. A. Cells of the retina and lateral geniculate fall into two classes: on-center (1) and off-center (2). B–F. Neurons of the primary visual cortex also fall into two major classes: simple and complex. Each of these classes, moreover, has several subclasses. This is illustrated here for simple cells. Despite this variety, however, all simple cells are characterized by three features: (1) specific retinal position, (2) their discrete excitatory (x) and inhibitory (-) zones, and (3) specific axis of orientation. For simplicity, only receptive fields with a vertical axis of orientation from 12 to 6 o'clock are shown in this figure; each has a rectilinear configuration. In fact, each region of the retina is represented in area 17, not only for this but for all axes of orientation—vertical, horizontal, and various obliques. (Adapted from Hubel and Wiesel, 1962.)

Figures from Kandel & Schwartz

There are two main types of cells. The **simple cells** are roughly linear except for rectification, are spatially and temporally band-pass, and show spatial phase sensitivity. A first approximation model for simple cell response firing rate (in impulses/sec) is:

$$\sigma(\boldsymbol{w}.\boldsymbol{g}),$$

where **g** is the image vector, **w** the receptive field weighting function, and $\sigma(\cdot)$ is a rectifying function (e.g **If[#>0,#,0]&**).

Both the psychophysical and neurophysiological data could be accounted for, in part, by assuming the visual system performs a quasi-Fourier analysis of the image, the exact form determined by the receptive field weighting function **w**.

We've seen how one possible model assumes that the visual system computes the coefficients (or spectrum) of an image with respect to the following basis set, called a Gabor set (Daugman, 1988). The

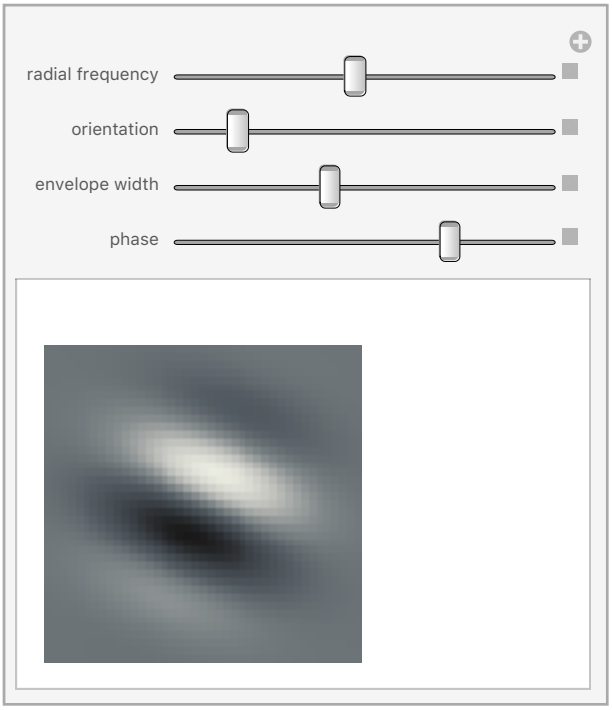
set $\{w_i\}$ is modeled as:

$\{e^{-\frac{(x^2+y^2)}{2\sigma^2}} \cos(2\pi(f_x x + f_y y + \phi))\}$, where i→$(f_x, f_y, \phi)$.

We will return to a more detailed discussion of the receptive field models of simple cells later in the section of functions of the visual cortex. The half-wave rectification operation, σ, sets negative values to zero, and is linear for positive values. The spectrum coefficients are represented by the firing rates of cells whose receptive field weights are represented by the above basis functions. In actuality, because simple cells behave more like linear filters followed by half-wave rectification, there should be two cells for each coefficient-- "on" and "off" cells). One difference between this basis set, and the Fourier basis set (i.e. the optical eigenfunctions) is that this set has a local spatial restriction because of the Gaussian envelope. A second difference, which has major implications for computation, is that the basis functions are, in general, not orthogonal.

You can view the demonstration either as a stimulus to test responses in cortical cells, or view it as a representation of the effective spatial weights of the underlying linear neural model that could account for the neuron's selectivities. If you open the phase slider, you can play a movie that also illustrates motion direction selectivity.

```
Clear[Grating,kern, GratingPatch];
Grating[x_,y_,fx_,fy_,phase_] := Cos[(2.0 Pi (fx x + fy y) + phase)];
GratingPatch[x_,y_,fx_,fy_,sig_,phase_] := Exp[-((x)^2 + (y)^2)/(2*sig^2)]*Grating[x,y,fx
kern[fx_, fy_, sig_,phase_] :=
  Table[GratingPatch[x, y, fx, fy, sig,phase], {x, -1, 1, .05}, {y, -1, 1, .05}];
  Manipulate[
GraphicsRow[{
ArrayPlot[kern[fr*Cos[theta], fr*Sin[theta],sig,phase]]}],{{fr,1,"radial frequency"},.1,2
{{theta,.4,"orientation"},0,Pi},{{sig,.4,"envelope width"},.001,1},{{phase,0,"phase"},.0,
```



The second major class of neurons is that of **complex cells**. Like simple cells, complex cells are spatially and temporally band-pass, show orientation and motion direction selectivity, but are insensitive to the phase of a stimulus such as a sine-wave grating. Rather than half-wave rectification, they show full-wave rectification. A model for complex cells would resemble the sum of the outputs of several subunits positioned at several nearby spatial locations. Each subunit would resemble simple cell with a linear spatial filter followed by a threshold non-linearity. One way of obtaining the phase insensitivity is to use subunits with cosine and sine phase receptive fields as above. The motion selectivity could be built in with appropriate inhibitory connections between subunits. Full-wave rectification could be built with subunit pairs that have excitatory and inhibitory receptive fields centers.

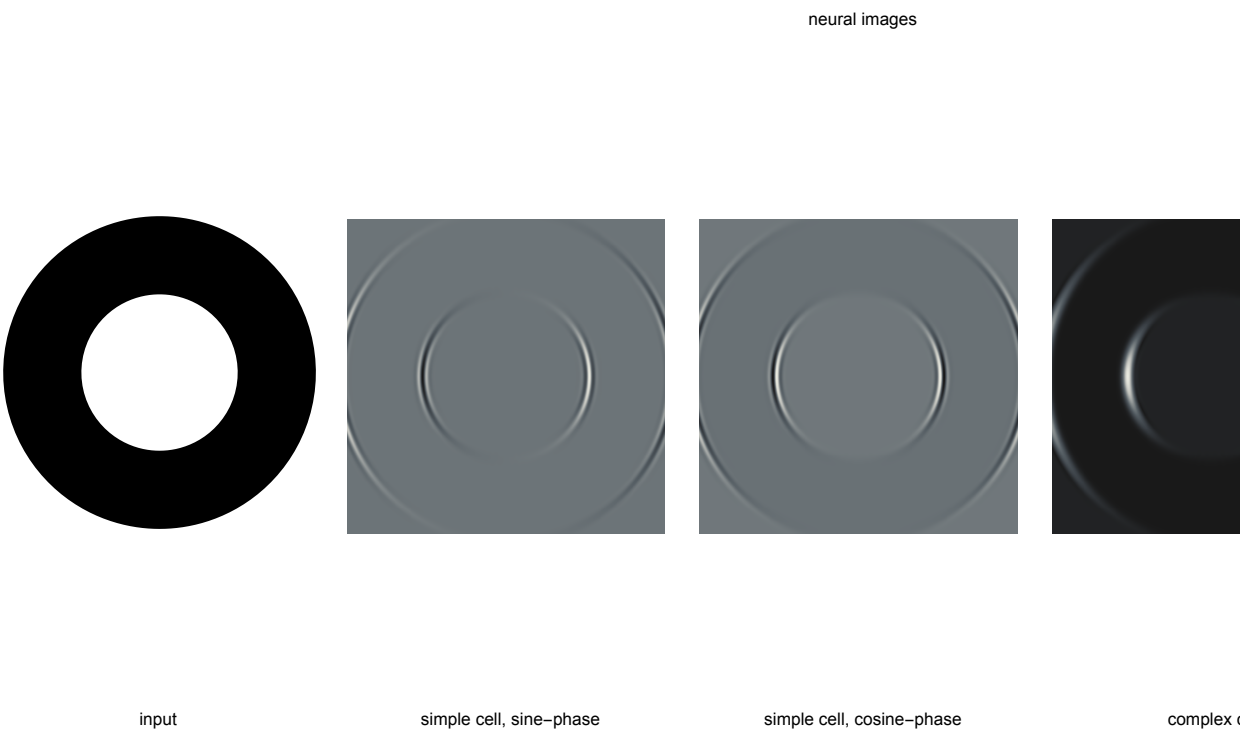
Quadrature pair model of complex cells

```
picture = ImageData[g9 = Graphics[Annulus[]]][[ ;; , ;; , 1]];
Image[ArrayPad[picture, 20, 1]];
```

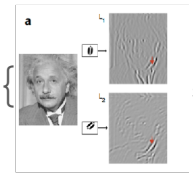
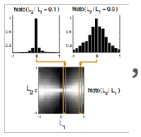
```
sgabor[x_, y_, fx_, fy_, sig_] :=
  N[Exp[(-x^2 - y^2) / (2 sig * sig)] Sin[2 Pi (fx x + fy y)]];
cgabor[x_, y_, fx_, fy_, sig_] :=
  N[Exp[(-x^2 - y^2) / (2 sig * sig)] Cos[2 Pi (fx x + fy y)]];

fsize = 32;
sfilter =
 Table[sgabor[(i - fsize / 2), (j - fsize / 2), 0, 1 / 8, 4], {i, 0, fsize}, {j, 0, fsize}];
sfilter = Chop[sfilter];
g10 = ArrayPlot[sfilter, Mesh → False, PlotRange → {-1, 1}, Frame → False];

fsize = 32;
cfilter =
 Table[cgabor[(i - fsize / 2), (j - fsize / 2), 0, 1 / 8, 4], {i, 0, fsize}, {j, 0, fsize}];
cfilter = Chop[cfilter];
g11 = ArrayPlot[cfilter, Mesh → False, PlotRange → {-1, 1}, Frame → False];

fspicture = ListConvolve[sfilter, picture];
g13 = ArrayPlot[fspicture, Mesh → False];

fcpicture = ListConvolve[cfilter, picture];
g14 = ArrayPlot[fcpicture, Mesh → False];
```

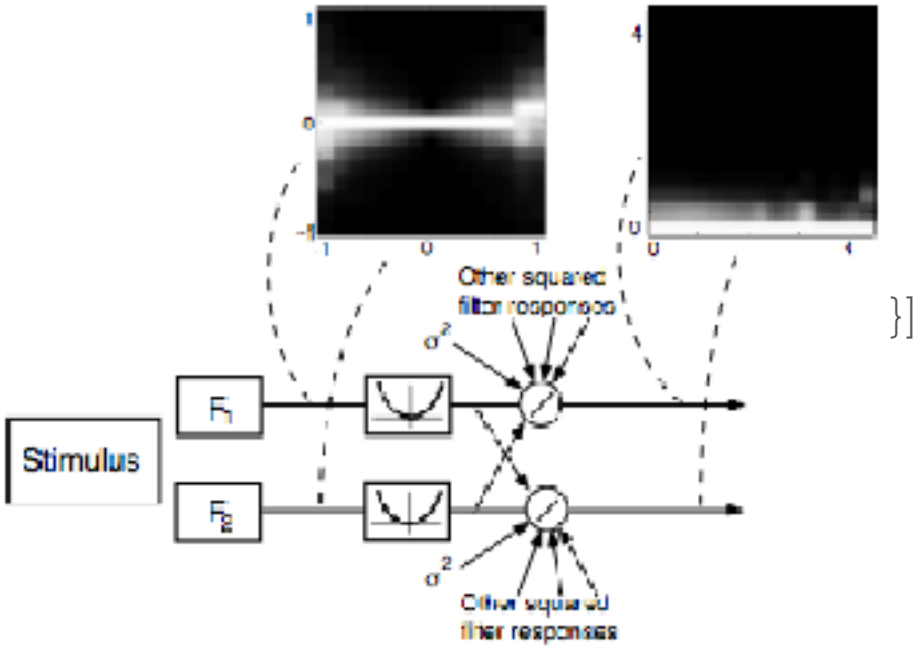
Look for peaks in local contrast energy

```
ss = Sqrt[fspicture^2 + fcpicture^2];
g15 = ArrayPlot[ss, Mesh → False];
GraphicsGrid[{{, , "neural images",}, {g9, g13, g14, g15}, {"input",
   "simple cell, sine-phase", "simple cell, cosine-phase", "complex cell"}}]
```

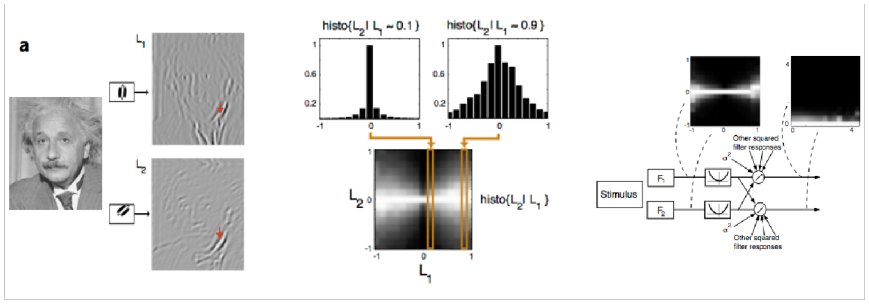


Both simple and complex cells show *contrast normalization*--an important feature not included in the above models. For a discussion of models of simple and complex cells, see: Heeger, D. J. (1991). Nonlinear model of neural responses in cat visual cortex. In M. &. M. Landy A. (Ed.), Computational Models of Visual Processing (pp. 119-133). Cambridge, Massachusetts: M.I.T. Press.

GraphicsRow[{  ,  ,



}]

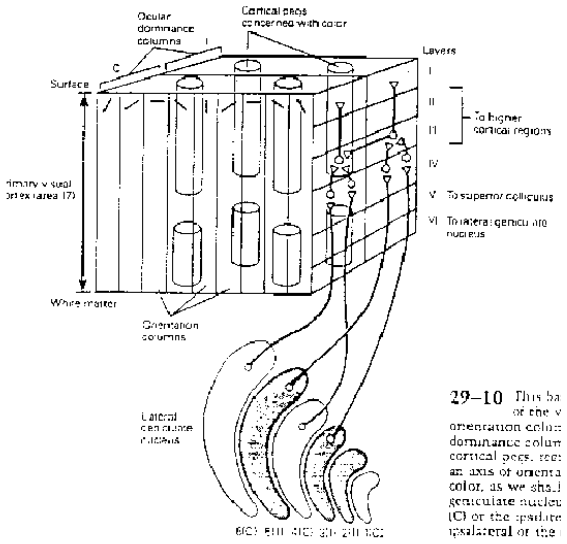


A third class of cells are the end-stopped (or "hyper-complex") cells that have an optimal orientation for a bar or edge stimulus, but fire most actively if the bar or edge terminates within the receptive field, rather than extending beyond it. It has been suggested that these cells act as "curvature" detectors. (Dobbins, A., Zucker, S. W., & Cynader, M. S., 1987).

But things aren't as necessarily as neat as they at first seem. "Hyper-complex" is seen as less of class, and instead cells can show "end-stopping". Further, see: Melcher and Ringach (2002) for a discussion of how the simple/complex cell distinction may be better thought of as continuous, rather than as a binary classification.

Columnar structure

In the cortex, we see for the first time binocular cells. The cells of the primary cortex are organized into columns running roughly perpendicular to the surface in which cells tend to have the same orientation preference and degree of binocularity. A "hypercolumn" is a group of columns spanning all orientations and both eyes



29–10 This ba... of the v... orientation colu... dominance colum... cortical pegs, rec... an axis of orienta... color, as we shall... geniculate nucleu... (C) or the pulvina... ipsilateral or the...

The receptive field organization of cortical cells is modifiable by experience. A number of models of self-organizing neural networks have been developed to account for this (Von der Malsburg, 1973; Bienenstock et al., 1982; Kohonen, 1981; and Linsker, 1988). Below we consider how efficient coding of natural image predicts how receptive field structure (Olshausen and Field, 1996; 2004).

Embedded in the cortical hypercolumns are cytochrome oxidase blobs in which are found opponent color cells that seem to lack strong orientation selectivity (Livingstone, M. S., & Hubel, D. H., 1984; Livingstone, M. S., & Hubel, D. H., 1987).

Functions of Primary Cortex

Local measurements

Basic idea:

V1 cortical cells measure local orientation-specific image contrast differences, that are correlated with spatial changes in surface/object depth, material (texture) and view-object and object-object changes (motion). Our challenge in the second half of the course will be to understand how local measurements can be used for global inference--e.g. as in object recognition.

Spatial frequency/orientation filtering: Psychophysics and physiology

Earlier, we looked at the psychophysical evidence for spatial frequency filtering in the experiment of Campbell and Robson, and the evidence for scale-invariance of the filters in the ideal-observer

experiments. These studies represent a small fraction of the psychophysics that has explored the properties of spatial frequency channels in human vision. Both adaptation and masking studies have also been used to infer properties of human spatial filters. The results of masking, adaptation, and other psychophysical studies of spatial and orientation frequency selectivity in human vision are surprisingly consistent in suggesting the basic form for a cortical basis set for images.

A discrete basis set model leaves several free parameters. Most models of detection and masking get by with about 6 spatial frequencies, about 12 orientations (specified by the ratio of horizontal and vertical spatial frequencies), and two phases (cosine and sine) at each retinal location. A subset of neurons representing a particular spatial frequency bandwidth makes up a spatial frequency channel. (Although there is neurophysiological evidence for pairs of V1 neurons having receptive fields with 90 deg phase shifted relative to each other, there is evidence against absolute phase--i.e. there is not a predominance of edge or bar type receptive fields. See Field and Tolhurst). One parameter still left unspecified is the standard deviation or spread of the Gaussian envelope. If large, this basis set approaches that of regular and global Fourier analysis. The psychophysical data suggest that the standard deviation be such that the Gaussian envelope is about one cycle (at the 1/e point) of the sine wave. One cycle corresponds to about 1.5 octaves spatial frequency bandwidth (an octave measure of width is: log to the base two of the ratio of the high to low frequencies.)

Why would the visual system have such a representation that combines orientation and spatial frequency selectivity? We have seen two types of explanations. One is that encoding over multiple spatial scales is important for subsequent processing that may involve edge detection, texture measurements, or stereoscopic matching, and so forth. Analogous pyramid schemes have been developed for computer vision. (See Adelson, E. H., Simoncelli, E., & Hingorani, R., 1987). The second explanation is in terms of economical or efficient encoding which we return to below (Simoncelli and Olshausen, 1999).

Stereo, or disparity measurements

As mentioned earlier, primary cortex brings together information from the two eyes in single neurons. This information is important for coordinated eye movements and stereo vision. Stereovision depends on the slight image differences, called disparities, that occur as a consequence of the two eyes having different views of the 3D world. Cells can be binocular without being sensitive to disparity. Although V1 cells are predominantly binocular, it was at first thought that disparity selectivity did not arise until V2 (Hubel and Wiesel, 1970). However, there is evidence for disparity selective cells in V1 and V2 (Poggio, G., F., & Poggio, T. ,1984). Disparity selectivity is a trivial task for single bar stimulus (in a uniform background), and it wasn't until relatively recently that neurons were found that effectively solve the problem of false matching (Poggio and Talbot, 1981). One possible algorithm for stereo vision is discussed here: Poggio, T. (1984). Vision by Man and Machine. Scientific American, 250, 106-115. Stereo vision has received a lot of attention in both computer and biological vision over the past several decades (Cumming, B. G., & DeAngelis, G. C. , 2001).

Motion measurements

The directional selectivity of cells in striate cortex provide a form of early motion detection, akin to that described for invertebrate and rabbit peripheral vision. This detection is only local and thus ambiguous. Cells early in visual processing suffer from the "aperture problem", and further computation is required to disambiguate object motion. Cortical cells are also selective for speed (Orban et al., 1983).

Both the motion selectivity and binocularity are consistent with our intuition of linking information likely to have a single environmental cause. We can add to that the the information should be in a formate that is useful for subsequent extra-striate processing. We will return to the computational theory of motion detection later.

In summary, basic image processing functions from eye to cortex are:

Retina

Spatio-temporal filtering attenuates low frequencies, wavelength/color coding

Chiasm

Begins grouping information from nearby points in the world to nearby anatomical locations.

Lateral geniculate nucleus (LGN)

More spatio-temporal filtering. Groups, but doesn't combine information from two eyes.

Primary visual cortex (V1, striate, 17)

Brings together local image measurements

--information that belongs together because of probable common cause

columnar structure

binocular vision and stereopsis

motion

edge & bar detectors

Spatial filtering by: Simple, complex, end-stopped cells

Why spatial filtering?

cortical basis set and efficient image representations

edge detection

Overview of extrastriate cortex

We've seen how to model the processing of spatial visual information in V1. Thirty years ago, one might have thought that a thorough understanding of primary visual cortex would produce a thorough understanding of visual perception. Not so. Since then, neurophysiologists have shown that primate visual processing has only just begun in V1. Much of this work is based on studies of the macaque monkey, but in the past decade and half, scientists have used brain imaging techniques to distinguish visual areas in the human cortex.

Extra-striate cortex

Primary visual cortex sends visual information to many other visually sensitive cortical areas (some estimates are about 30 visual areas in the macaque). These areas have been identified through anatomi-

cal, histological, and physiological techniques with the early work by Samuel Zeki at the University of London, and David Van Essen and colleagues. Areas have been delineated by:

Function: physiology, neurons in different brain areas selective for different aspects of image inputs

Architecture: laminar cytoarchitecture (e.g. cell size,cell density, density of axons, layering, discovered using different kinds of stains).

Connections: anatomical connections traced using retrograde and anterograde tracers.

Topography: retinotopic maps in each of several of the early visual areas (V1, V2, V3, V4...).

Primary visual cortex has a fairly precise topographic map of the visual field--nearby points in the image map to nearby cells in V1. Other areas have less precise topographic maps of the visual field.

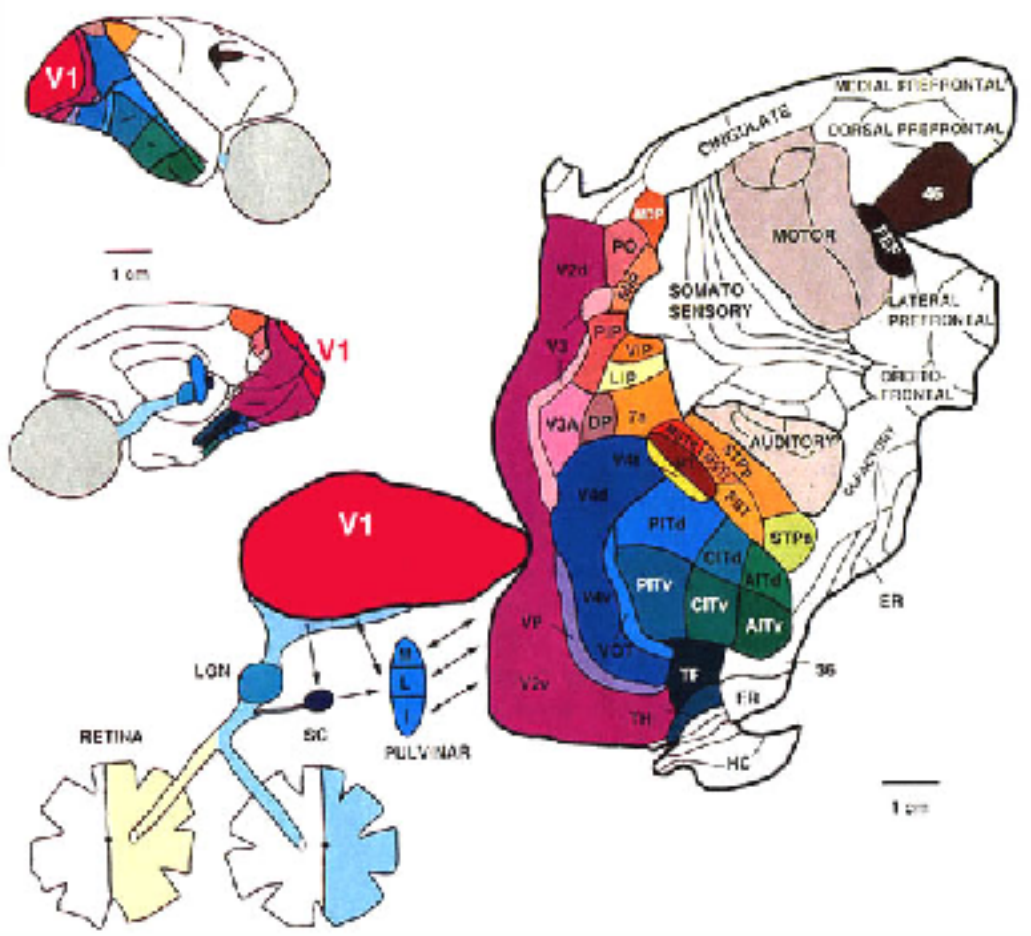
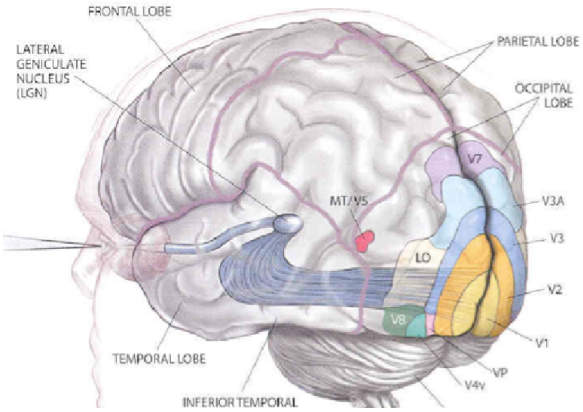
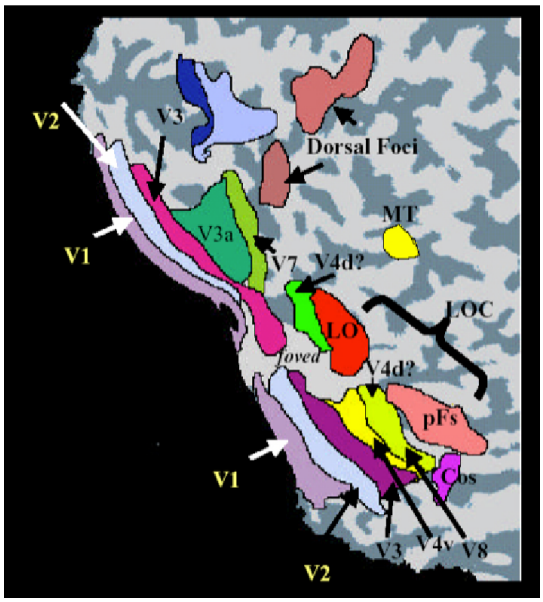


Figure 19. Much of V1 is located in the calcarine sulci and its relationship to other brain areas is best shown by unfolding the brain and showing it flattened open. The visually responsive areas of the macaque monkey are shown in color. From Van Essen et al. (1992).

From Van Essen et al. 1992

Human visual areas

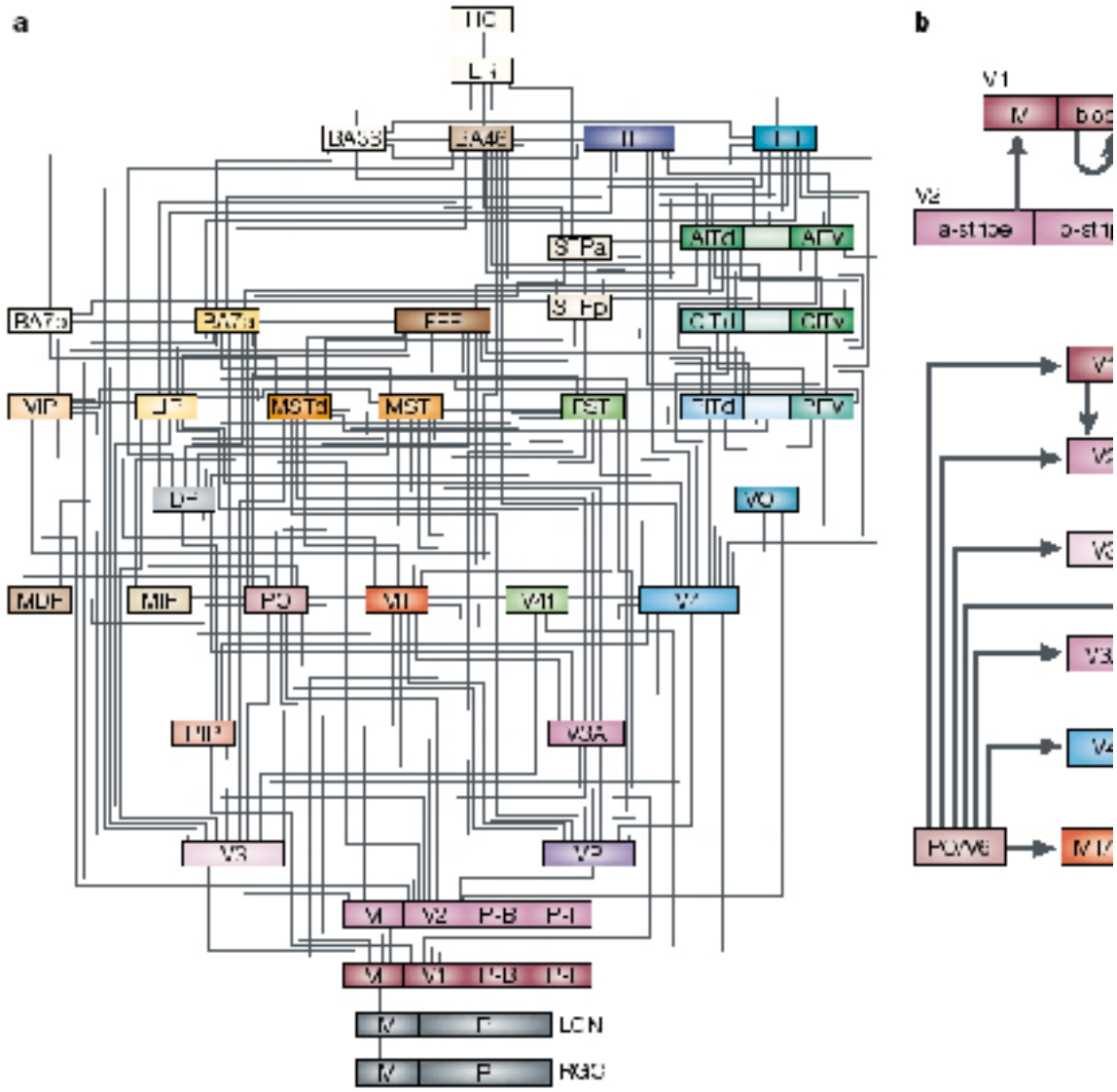


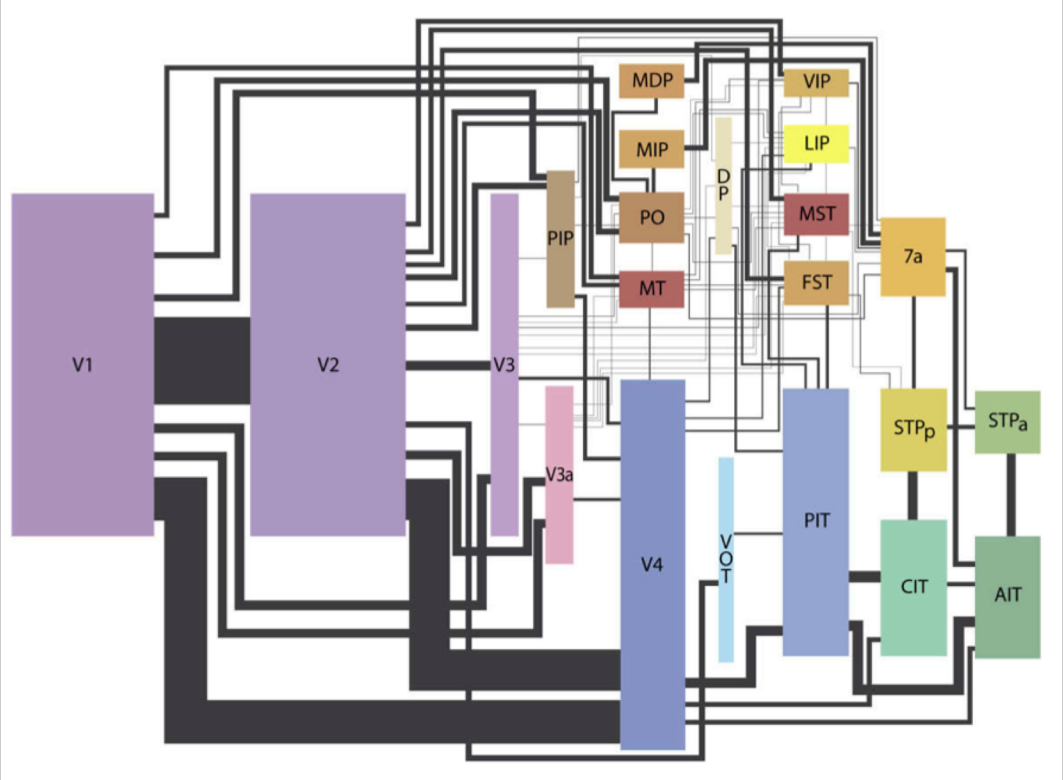


Left figure: From Scientific American.

Visual hierarchy

One of the remarkable discoveries about extra-striate cortex is that these areas are organized hierarchically (See Felleman and Van Essen, 1991; DeYoe and Van Essen, 1988; DeYoe et al., 1994), and involve multiple parallel pathways.


a
b
V1
V2
LCN
HCN



From Wallisch, P., & Movshon, J. A. (2008). Structure and Function Come Unglued in the Visual Cortex. Neuron, 60(2), 194–197.

A general pattern of connectivity between areas has emerged in which one sees:

• feedforward connections from superficial layers (I, II, III) to IV

• feedback connections originating in deep (V, VI) and superficial layers and terminating in and outside layer IV.

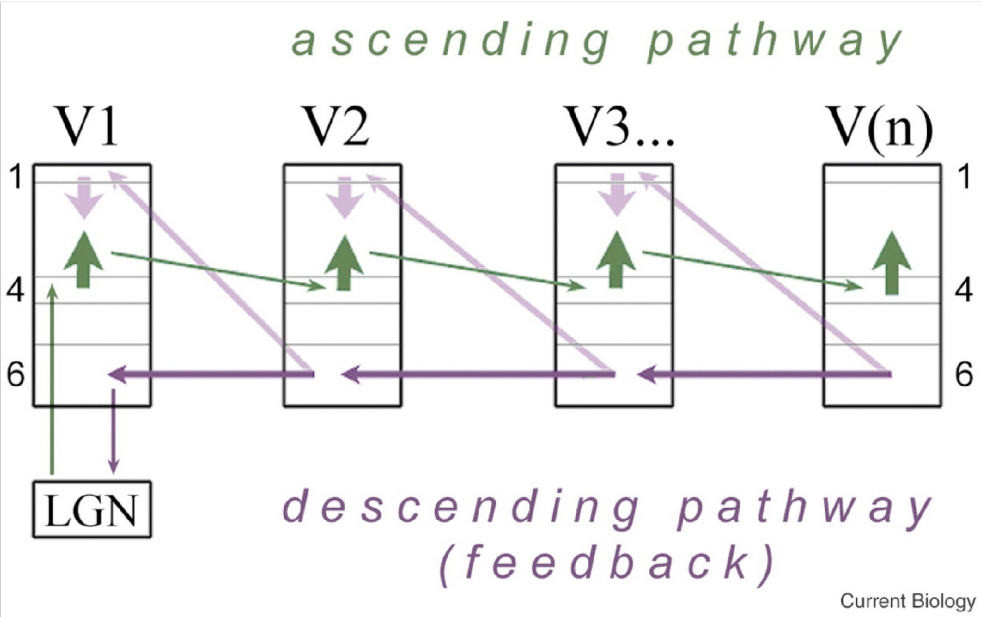
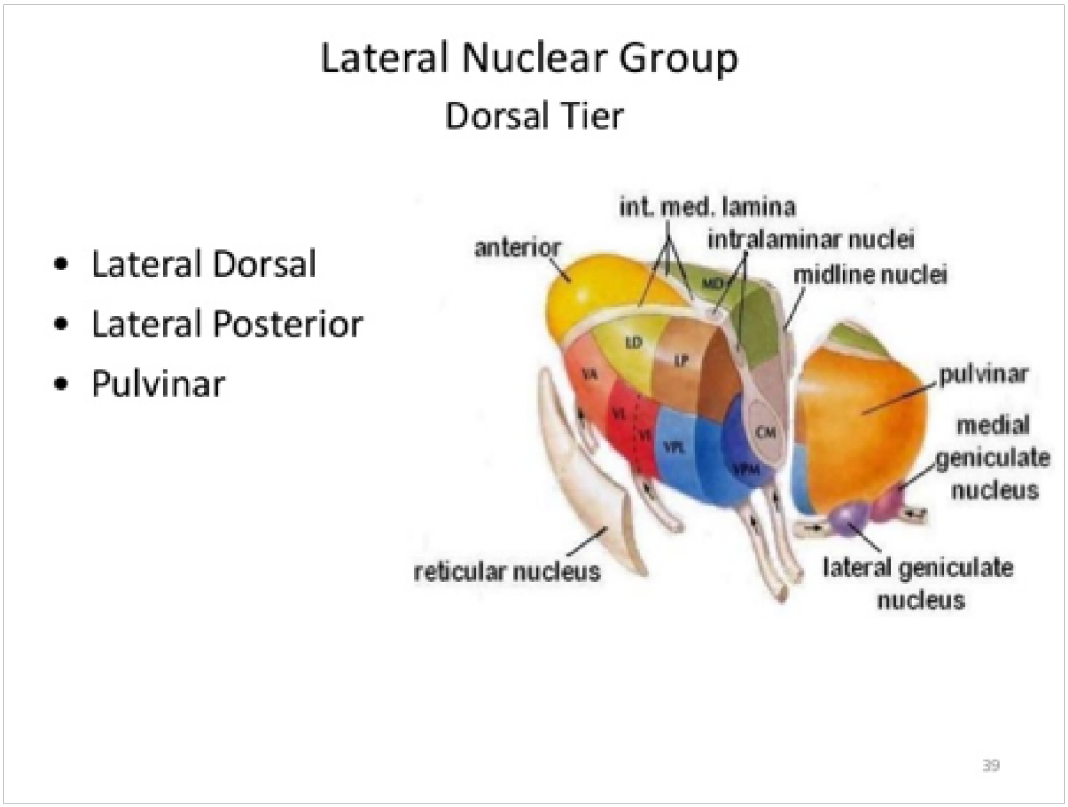


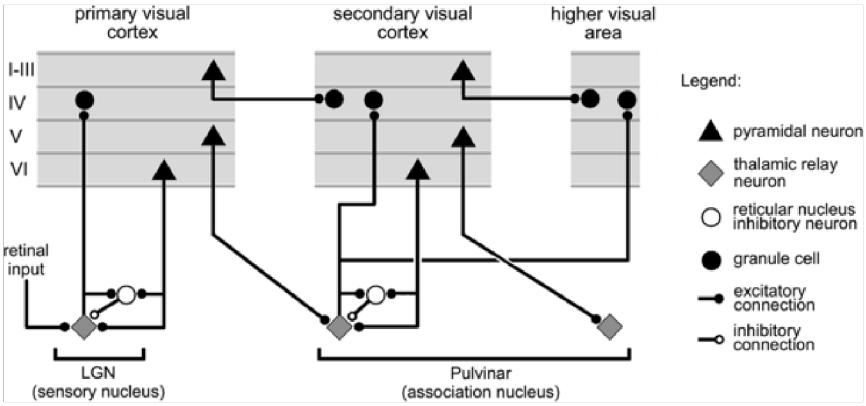
Figure showing dominant feedfoward and backward connections. From: Shipp, S. (2007). Structure and function of the cerebral cortex. Current Biology, 17(12), R443–9.

Connections to pulvinar and thalamus



Sherman, S. M. (2012). Thalamocortical interactions. Current Opinion in Neurobiology, 22(4), 575–579. http://doi.org/10.1016/j.conb.2012.03.005

Figure from: Gisiger, T., & Boukadoum, M. (2011). Mechanisms Gating the Flow of Information in the Cortex: What They Might Look Like and What Their Uses may be. Frontiers in Computational Neuroscience, 5, 1–15. http://doi.org/10.3389/fncom.2011.00001



Functions?

What are these extra-striate visual areas of cortex doing? At a general level, these areas turn image information into useful behavior, such as recognition, visuo-motor control, and navigation. Below we outline current views on two large-scale functional pathways. But it is also important to begin to look for detailed computations that extra-striate areas are doing. At the current time, we have only a few specific ideas, some of which we will look at in the lectures on motion perception, and object recognition.

For example, the very large receptive fields found in extra-striate areas (e.g. MT cells can have receptive fields as large as 100 deg!) bring together information from distant parts of the visual field. Again, the general idea is that information which likely belongs to edges at different locations on the same object is brought together.

Preview of computational problems

A few computational problems can be seen by taking a generative view--how scene properties affect local measurements:

stereovision

depth change causes different feature displacements in the two eyes.

motion disambiguation

object motion in one direction produces local motion signals going in different directions

color constancy

a constant gray surface can result in different intensities at different locations due to non-constant illumination

object contours & regions

An object's boundary produces a broad array of different local orientations

An object's interior can be made of many different samples of one texture, as well as many different textures.

The role of task--end-goals

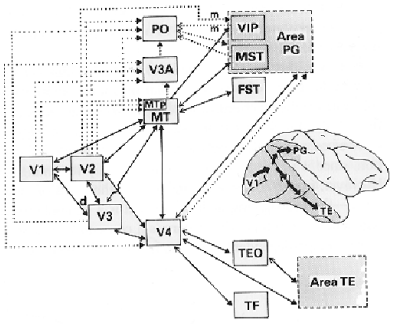
Large scale functional pathways

The flow of visual information follows two dominant streams. In the dorsal or parietal stream, information flows from primary cortex to parietal cortex. A substream that has been studied for motion processing is: V1 <-> MT <-> MST.

The temporal stream carries information from primary visual cortex to infero-temporal cortex. A

sub-stream which has been studies for object recognition is: V1 <-> V2 <-> V4 <-> IT.

Here is another figure illustrating the functionally distinct areas and their connections:



Smaller scale, parallel pathways

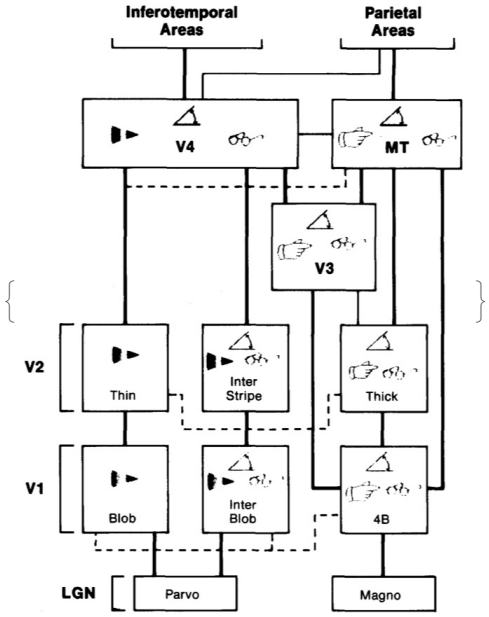


Fig. 3. Schematic diagram of anatomical connections and neuronal selectivities of early visual areas in the macaque monkey. LGN = lateral geniculate nucleus (parvocellular and magnocellular divisions). Divisions of Vl and V2: blob = cytochrome oxidase blob regions; interblob = cytochrome oxidase--poor regions surrounding the blobs; 4B = lamina 4B; thin = thin (narrow) cytochrome oxidase strips; interstripe = cytochrome oxidase-poor regions between the thin and thick strips; thick = thick (wide) cytochrome oxidasestrips; V3 = visualarea 3; V4 = visualarea(s) 4; MT = middle temporal area. Areas V2, V3, V4, MT have connections to other areas not expficitly represented here. Area V3 may also receive projections from V2 interstripes or thin stripes79. Heavy lines indicate robust primary connections, and thin lines indicate weaker, more variable connections. Dotted fines represent observed connections that require additional verification. Icons: rainbow = tuned and/or opponent wavelength selectivity (incidence at least 40%); angle symbol = orientation selectivity(incidenceatleast 20%); spectacles = binocular disparity, pointing hand = direction of motion selectivity (incidence at least 20%).

The icons signify selectivity for: wavelength/color (prism wedge), binocularity (spectacles), orientation (angle), and motion (finger pointing)

Here's another illustration from a different review:

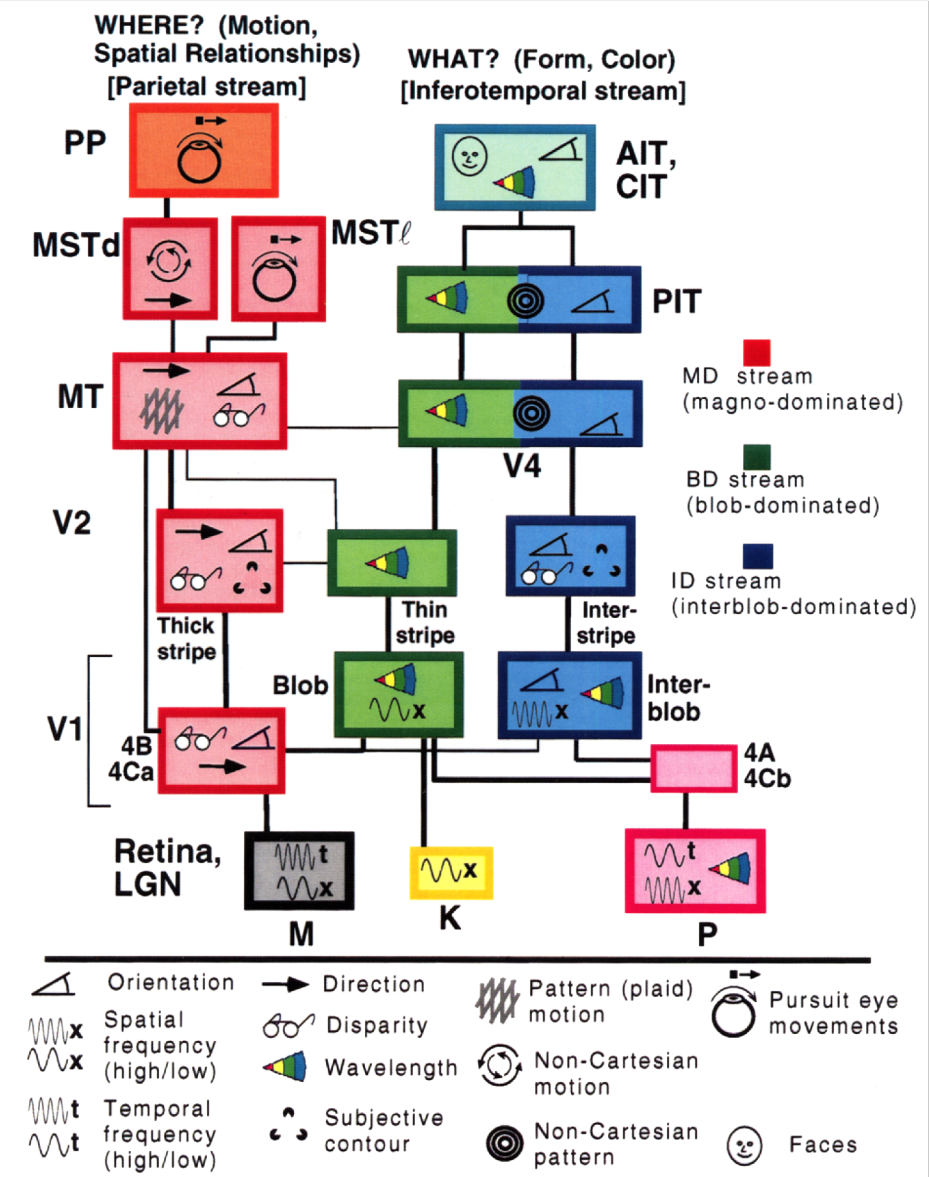


Figure 2. Hierarchical Organization of Concurrent Processing Streams in the Macaque Monkey
Boxes represent visual areas, compartments within an area, and subcortical centers; solid lines represent major connections between structures (usually reciprocal pathways); and icons represent characteristic neurophysiological properties. Subcortical streams in the retina and lateral geniculate nucleus (LGN) include the M, K, and P streams (grey, yellow, and pink, respectively). Cortical streams at early and intermediate stages include the MD, BD, and ID streams (red, green, and blue, respectively). The PP complex is shown in orange. The IT complex includes posterior inferotemporal areas (PIT), which are components of the BD and ID streams, and central and anterior areas (CIT and AIT). Pathways are based on information cited in Felleman and Van Essen, 1991, and Van Essen and DeYoe, 1994. Physiological specializations are based on information cited in Felleman and Van Essen, 1987, DeYoe and Van Essen, 1988, and in the present text.

Van Essen, D. C., & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. Neuron, 13(1), 1–10. http://doi.org/10.1016/0896-6273(94)90455-3

Functional streams and parallel pathways

Based on studies of the behavior of monkeys and man with lesions, and work using electrophysiological techniques, it is thought that the parietal stream has to do with navigation, and view-centered representations of the visual world. It is sometimes called the "where" system (Mishkin and Ungerlei-

der, 1983). Although it may more to do with "how" (Goodale & Milner 1992).

The temporal stream is sometimes called the "what" system. It is believed to be important for representations useful for object recognition--the representation of “intrinsic” object properties, that are insensitive to viewpoint variations. Form and color of objects is thought to be extracted by interacting modules in the temporal stream.

Current working hypotheses regarding function:

dorsal / parietal areas: e.g. V1 -> MT -> MST
"where out there?"
navigation, viewer centered representation
motion for layout, heading (MST)
...and for driving motor actions such as reaching

temporal: e.g. V1 -> V2 -> V4
"what is it?"
processing for non-viewer or object-centered representation
material color and shape & form
...and further downstream, temporal areas (IT) for object recognition

General Extra-striate Functions

The discovery of 30+ extra-striate visual areas in the macaque, together with a lack of ideas about what all of these modules are doing, suggests that it might be useful to step back and think about the computations that are required to perceive and act.

We will first focus on the idea that an intermediate goal of visual processing is to bring together local information/measurements from distant parts of the visual field likely to belong to same object, or have the same cause. Our study of edge detection shows that local ambiguity is a major computational challenge. So we will spend time understanding how to integrate local ambiguous measurements to arrive at useful representations of objects and their relationships to each other and to the viewer. Later we will try to understand how this intermediate-level processing leads to useful actions.

Note on terms: Low-level (early), intermediate-level (middle), and high-level vision.

One can think of overall visual organization in terms of the knowledge available at different levels.

Low-level--local measurements, simple grouping procedures. Low-level vision can be performed by a system which knows only about regularities of image patterns (e.g., that images typically consist of regions where the intensity changes slowly which are separated by edges where the intensity changes rapidly).

Intermediate-level--surfaces and surface-like representations, more global grouping processes, objects,..
Intermediate-level or mid-level vision processing knows about properties of geometric surfaces (e.g., that they tend to be spatially smooth) and that they can partially occlude each other.

High-level--functional tasks, object recognition, navigation, reaching,...High-level knows about objects, the relationships between them, and actions.

The flow of information from low- to high-level vision is from generic to specific.

From images to objects, scenes, and actions

What we have learned about the brain's very early processing of image information tells us rather little about how image information leads to useful behavior. Most of what we have studied shows how image information is coded into other forms that still has more to do with the image, than with what is out there, that is, the scene. But if much as 40-50% of visual cortex may be involved in visual processing, what is all this cortex for? In order to begin to answer this question, we ask a more general question of interest to both computer and biological vision scientists.

The role of computer vision

Visual function & tasks

So far, we've primarily addressed the issue of visual input, and have by and large ignored the analysis of functional visual behavior. Now it is time to ask: What are the goals of vision? The obvious answers are to gain information about the physical world useful for navigating, recognizing objects and planning future actions. In the 1940's, Kenneth Craik suggested that perception was a process in which the brain constructs a model of the physical world, in much the same way that in engineer builds (or perhaps simulates on a computer) a scale model of an airplane. The purpose of such a model is to test hypotheses about how it would function if actually built. This process of going from an image, which is a changing array of light intensities, to a model of the environment is a problem of image understanding. In order to gain an appreciation for what this process entails, let us look at some example questions of image understanding. But it is not necessarily the case that a 3D representation of the world is the best preliminary step to achieve a functional goal. There may be more direct processing strategies that are efficient in achieving a goal. In the first example below, a scalar summary statistic of optic flow can provide initial information for a collision. Further, evidence from human studies of visual attention show that people can be surprisingly "blind" to major changes between two images. This is the so-called phenomenon of "change blindness".

Nevertheless, no one disputes that vision must somehow convert image input to useful actions. Here are some examples.

• Given a dynamically expanding image on my retina, how long will it be before I collide with the object producing it? Here one would like to estimate time-to-contact from changing light intensities. One preliminary step may be to estimate optic flow, that is, compute the 2D projected velocity field of the 3D surface points. Computing optic flow itself is an inferential process. We will see later how a simple measure of optic flow expansion rate can be used to predict "time to contact".

• Given two slightly different images, one in the left eye and one in the right, what is the relative depth of the various objects causing the two images? This is the problem of stereopsis.

• Given spectral, spatial and temporal changes in the illumination falling on a particular object depending on time of day, how can I assign a relatively stable color to it? This is the problem of color constancy. Or when driving down the road, how do I avoid misinterpreting a large dark shadow for a turn off exit? Without direct measurements of the incident light, it is not immediately clear how to do this.

• Given contours, shading and/or texture pattern, how I can infer the shape of the object? This is the shape-from-X problem, where X is a local image measurement such as shading or texture gradients or motion flow.

These problems are so trivial for us as observers, they disguise the underlying difficulty of perception. Until the attempts over the last couple of decades to develop intelligent computer vision systems, it was not fully appreciated that many of the visual tasks that we as human observers accomplish so effortlessly are profoundly difficult to reproduce with a machine. We emphasized at the beginning of this course that to understand the biology and psychology of image understanding, one must also study the computational problems the biological substrate supports (Marr, 1982).

Many diverse goals suggests the importance of maintaining as much information as possible during early transmission stages perhaps through the kind of efficient coding models that we have studied. Some computer vision approaches have used the idea of "shared features". Succeeding stages preserve information, but with progressive selection aimed at the goals of the visual system. A major challenge is understanding the trade-off between selectivity and invariance in visual recognition (Geman, 2006).

The difficulties of developing image understanding models

What are the difficulties of image understanding? We've already spent considerable time thinking about how image inputs should be represented. Two major additional problems are:

• What is the output and how should it be represented?

• How can we compute scene-related outputs given an set of image measurements or representation?

Although the first input to vision can be represented as light intensity as a function of space and time, followed by spatial and temporal filtering, it is not at all clear how to represent the brain's visual "output".

One view is to model estimates of the scene parameters causing the image, as well as the relationships between features or parts, and the relationships between objects. Another (not necessarily exclusive) view is to more directly extract useful parameters for function (e.g. geometric shape dimen-

sions for object recogniiton, depth relationships bewteen viewer and object, time-to-contact for braking, or motor control variables for actions).

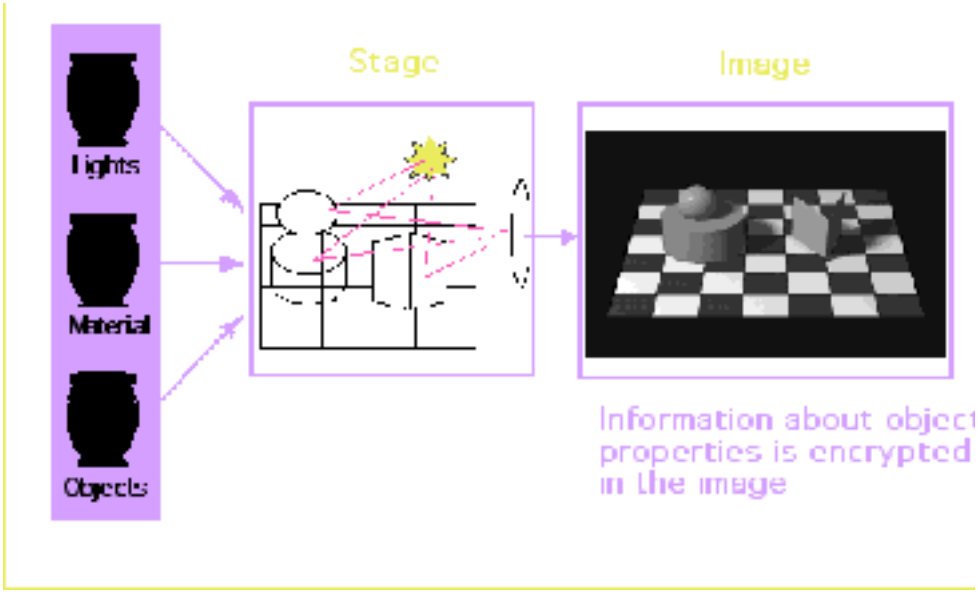
The role of scene-based image modeling

The image filtering approach can be thought of as primarily "image-based". The advantage of image-based modeling is that it is "closer to the input". Features are indexed spatially which makes sense given topographic representations. But as we begin to think about representing object properties, it may make more sense to think about indexing based on other measures of "closeness", such as viewpoint, or class membership.
When we consider visual tasks, it is useful to consider generative models that are "closer to the output" of vision. At first, this may sound counter-intuitive, so let's see what this means.

The first step of analysis is to understand the generative model of image formation in terms of the causal structure of the world. Here we can gain insight from 3D computer graphics. For example, here is a model of the image L(x,y):

$$L(x,y) = f(R(x,y), N(x,y), V, E)$$

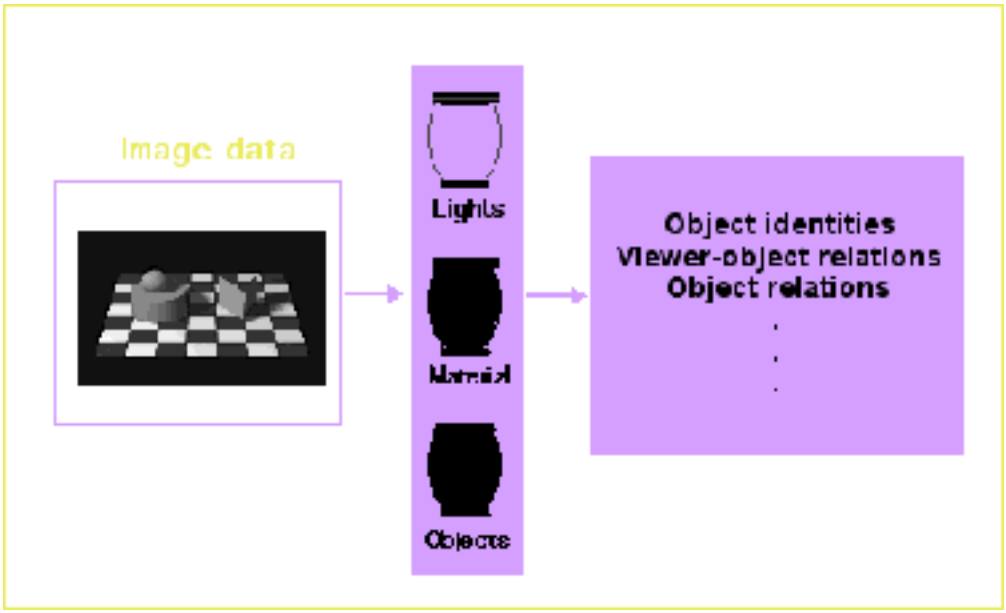
where **L** is the luminance, **R** is the albedo (surface reflectivity), **N** is a vector representation of the shape of the surface, **V** is the viewer angle, and **E** describes the lighting geometry (number, type and power of illuminants).



The inverse 3D graphics metaphor

One way to view vision is as the reconstruction of the scene or as the "decrypting" of the image to reveal the "message" that the world is sending. In this sense image understanding is a problem in

inverse optics or "inverse computer graphics". As an example, the forward optics problem may specify the luminance at each point of a surface as a function of the surface's albedo, its local geometry (or shape), the position of the viewer relative to the surface, and the lighting conditions:



The inverse problem is to take as input, **L**, and compute the scene causes **R** , **N**, **V** or **E**. Although it is unlikely that human vision exactly solves the inverse graphics problem even in small domains, the metaphor is useful to make explicit image ambiguities and to test functional goals and constraints utilized in human perception (Kersten, 1997). But there are strong limitations to the metaphor. One of them is that it doesn't make explicit the diverse set of tasks and requirements of flexible visual processing to accomplish those tasks.

Even if we could solve the inverse problem, how should one represent the mental homologues of shape, material properties, lighting or the geometrical relations between objects? For example, should depth be represented as absolute distance, relative distance, or perhaps not metrically at all, but rather in terms of ordinal relations? Or, should depth be a purely implicit in the neural transformatioins that take one from images to actions? Should shape be represented locally or globally? When is it important to compute depth, the first derivative of depth, or the second deriviative of depth? Each has a different utility, and the image information supporting inference can have a different relation to each. Despite the fact that the representation issue is so critical to arriving at a true account of biological visual functioning, it is often the most difficult to answer. Clues have to be sought in neurophysiological, psychophysical and computational studies. We will emphasize the computational approach to these problems and often will proceed with only a guess as to what the visual system is computing, and then look at how one can get from the input data to the desired output.

The second major problem is specifically that the image data, L(x,y,t) does not make explicit any of the parameters representing the scene.
We run into two sub-problems. First, as was emphasized in the context of edge detection, any local image measurement is often a function of more than one cause. For example, an intensity change is a

function of material and illumination change. Further, even when given multiple sources of visual information (e.g. motion parallax and stereo views), one has to somehow combine this information to yield a unitary percept. This combination should be done in a common "language", with some measure of the reliability of each source. Second, even a single cause may be ambiguous. For example, many 3D "wire" objects map to the same 2D line-drawing. The image data mathematically underconstrains the solution--the inference or estimation problem is sometimes said to be "ill-posed".

The role of ideal observers & Bayesian decision theory

At the beginning of the course, we showed the advantages starting off with a formal statement of what an ideal image understanding system should be doing, and then investigate the ways in which one might approach this ideal.

An earlier lecture provided a preview of how Bayesian decision theory could be used to develop a framework for estimating scene properties from images.

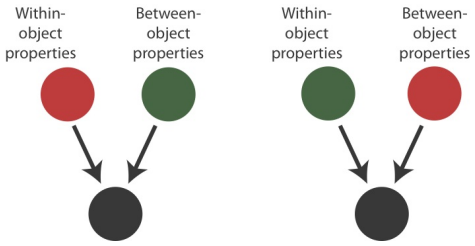
In particular, the ideal observer can be modeled as a Bayesian estimator of scene parameters, given image data. E.g. the MAP observer would pick the most probable scene given a fixed set of image measurements based on the posterior probability

p(scene | image measurements)

This formulation can be used to cast many of image understanding problems in terms of finding minima of high dimensional "cost" or "energy" functions. We can run into problems with multiple minima, and it becomes difficult to find the right one, which in general is the lowest one. One can either improve the descent methods (e.g. simulated annealing, or multi-grid techniques), re-shape the topography of the cost function appropriately, or change the representational architecture of the problem. This involves choosing the right input and output representations, and raises questions like: Should one use raw light intensities for input, or some other primitives like edges or local Fourier transforms? What purpose is gained by 2D preprocessing or filtering of the image? We can get some insight into these questions by studying what is known about the psychology and physiology of vision. A Bayesian approach adds an additional and arguably important twist by placing an emphasis on the reliability of multiple sources of interacting information--a competent visual inference device doesn't just proceed by passing the estimate at one stage on to the next, it should also pass information regarding the reliability of its estimates.

Choosing an efficient algorithm for finding the right solution depends on both the computational problem and on the hardware available for implementation. We will see that neurons have limited dynamic range, limited metabolic resources, limited dendritic connectivity and spread, and so forth. Efficiency has to be evaluated relative to both computational and hardware constraints.

The selection and processing of information will differ depending on task. For example, the Bayesan decision theory perspective is consistent with the ideas of ventral and dorsal stream processing involving mechanisms that select and discount information appropriate for the distinctly different tasks of extracting intrinsic object properties vs. deciding their spatial relationships.



- 1. How would the computational problems of size estimation differ for the tasks of grasping an object vs. recognizing an object?
- 2. How does the inverse optics or graphics view differ from efficient coding?

References

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci, 17*(21), 8621-8644.

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science, 264*(5163), 1333-1336.

DeYoe, E. A., & Van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. Trends in Neuroscience, 11(5), 219-226.

DeYoe, E. A., Felleman, D. J., Van Essen, D. C., & McClendon, E. (1994). Multiple processing streams in occipitotemporal visual cortex. Nature, 371(6493), 151-154.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex, 1(1), 1-47.

Friston, K. (2005). A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci, 360(1456), 815-836.

Geman, S. (2006). Invariance and selectivity in the ventral visual pathway. J Physiol Paris, 100(4), 212-224.

Goodale, M. A., & Milner, D. A. (1992). Separate visual pathways for perception and action. Trends in Neuroscience, 15(1), 20-25.

Guillery, R. W., & Sherman, S. M. (2002). Thalamic relay functions and their role in corticocortical communication: generalizations from the visual system. Neuron, 33(2), 163-175.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proc Natl Acad Sci U S A, 93*(2), 623-627.

Kersten, D. & Schrater, P. R., (submitted). Pattern Inference Theory: A Probabilistic Approach to Vision. In R. Mausfeld, & D. Heyer (Ed.), Perception and the Physical World. Chichester: John Wiley & Sons, Ltd. http://vision.psych.umn.edu/www/kersten-lab/papers/KerSch99.pdf

Kersten, D. (1997). Inverse 3-D graphics: A metaphor for visual perception. *Behavior Research Methods, Instruments, and Computers*., **29**, 37-46.

Kersten, D. & Madarasmi, S. (1995) The Visual Perception of Surfaces, their Properties, and Relationships, Proceedings of the DIMACS Workshop on Partitioning Data Sets: With Applications to Psychology, Vision and Target Tracking - 1993.

Knill, D. C., & Kersten, D. K. (1991). Ideal Perceptual Observers for Computation, Psychophysics, and

Neural Networks. In R. J. Watt (Ed.), Pattern Recognition by Man and Machine MacMillan Press.
Maunsell, J. H. R., & Newsome, W. T. (1987). Visual Processing in Monkey Extrastriate Cortex. Annual Review Neuroscience, 10, 363-401.
Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. Trends in NeuroSciences, 6, 414-417.
Eero P Simoncelli and Odelia Schwartz (1998) Modeling Surround Suppression in V1 Neurons with a Statistically-Derived Normalization Model . Advances in Neural Information Processing Systems 11. ed. M.S. Kearns, S.A. Solla and D.A. Cohn, pp. 153-159, May 1999. © MIT Press, Cambridge, MA.
E P Simoncelli. Statistical models for images: Compression, restoration and synthesis. In 31st Asilomar Conf Signals, Systems and Com-puters, pages 673–678, Pacific Grove, CA, November 1997. Available from http://www.cns.nyu.edu/~eero/publications.html
Tolhurst, D. J., & Heeger, D. J. (1997). Comparison of contrast-normalization and threshold models of the responses of simple cells in cat striate cortex. *Vis Neurosci, 14*(2), 293-309.
B Wegmann and C Zetzsche. Statistical dependence between orientation filter outputs used in an human vision based image code. In Proc SPIE Visual Comm. and Image Processing, volume 1360, pages 909–922, Lausanne, Switzerland, 1990.
See: http://www.cns.nyu.edu/~eero/ABSTRACTS/simoncelli98d-abstract.html