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

Initialize

■ Spell check off

```
In[1]:=   Off[General::spell1];
```

Outline

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

■ Oriented filters: 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

Today

■ **Next week's homework**

■ **Adaptive filtering**

■ **Extrastriate cortex--overview**

■ **Scenes from images, scene-based modeling of images**

Adaptive spatial filters & V1 cells

■ **Simultaneous contrast**

A classic brightness illusion demonstrates what we might expect of a spatial filter that adapts its response to the response of its neighbors. The circle on the left appears to be darker than the circle on the right; however, the intensity is the same for both (**ming**). One way to explain this is that the respone of a single unit that signals intensity gets divided by the responses of neighboring units that also signal intensity.

```
width = 110; radius = width / 5;
maxg = 0.85; ming = 0.5; maxg2 = 0.55;
d1 = Table[If[(i - width / 2) ^ 2 + (j - width / 2) ^ 2 < radius ^ 2, ming, maxg],
    {i, 1, width}, {j, 1, width}];
g1 = ListDensityPlot[d1, Mesh → False, PlotRange → {0, 1}, Frame → False,
    DisplayFunction → Identity];
d2 = Table[If[(i - width / 2) ^ 2 + (j - width / 2) ^ 2 < radius ^ 2, ming, maxg2],
    {i, 1, width}, {j, 1, width}];
g2 = ListDensityPlot[d2, Mesh → False, PlotRange → {0, 1}, Frame → False,
    DisplayFunction → Identity];
Show[GraphicsArray[{g1, g2}, GraphicsSpacing → .1],
  DisplayFunction → $DisplayFunction];
```


■ Simultaneous "Contrast of Contrast"

The above illusion involved manipulating mean light level. What if we manipulate contrast (or variance) and keep the means fixed?

```
width = 110; radius = width / 5;
maxg := 0.8 * (Random[] - 0.5) + 0.5;
ming := 0.25 * (Random[] - 0.5) + 0.5;
maxg2 := 0.15 * (Random[] - 0.5) + 0.5;
d1 = Table[If[(i - width / 2) ^ 2 + (j - width / 2) ^ 2 < radius ^ 2, ming, maxg],
    {i, 1, width}, {j, 1, width}];
g1 = ListDensityPlot[d1, Mesh → False, PlotRange → {0, 1}, Frame → False,
    DisplayFunction → Identity];
d2 = Table[If[(i - width / 2) ^ 2 + (j - width / 2) ^ 2 < radius ^ 2, ming, maxg2],
    {i, 1, width}, {j, 1, width}];
g2 = ListDensityPlot[d2, Mesh → False, PlotRange → {0, 1}, Frame → False,
    DisplayFunction → Identity];
Show[GraphicsArray[{g1, g2}, GraphicsSpacing → .1],
  DisplayFunction → $DisplayFunction];
```
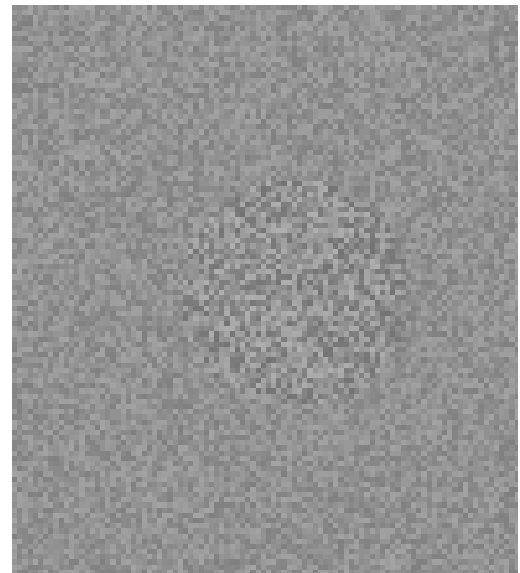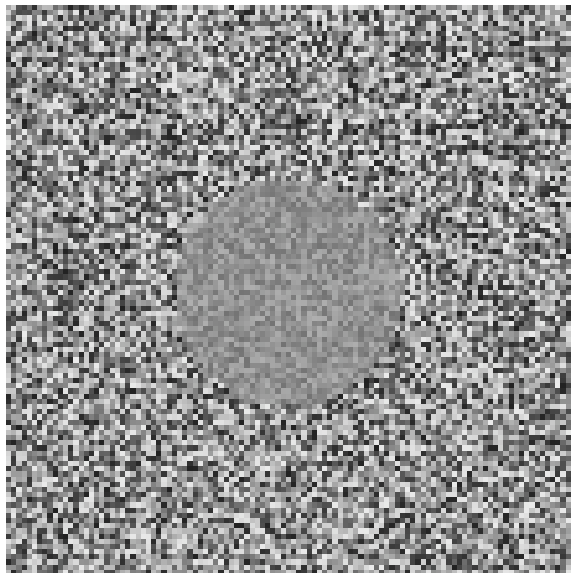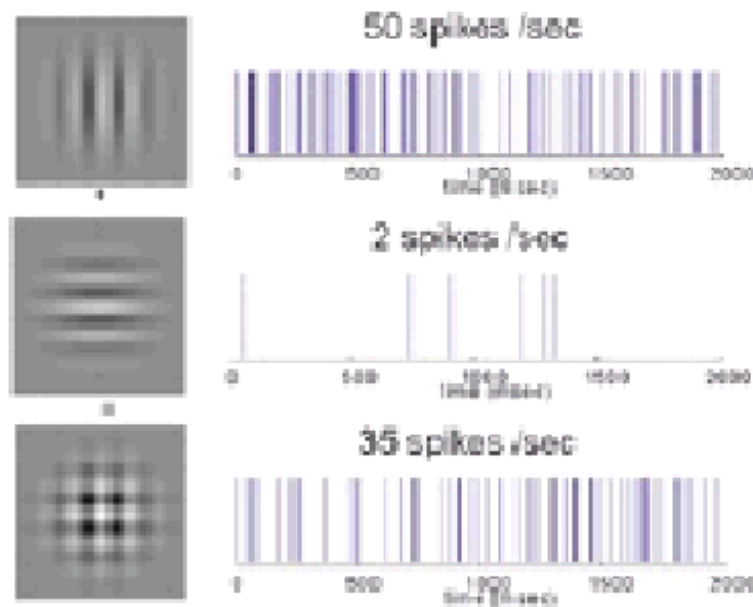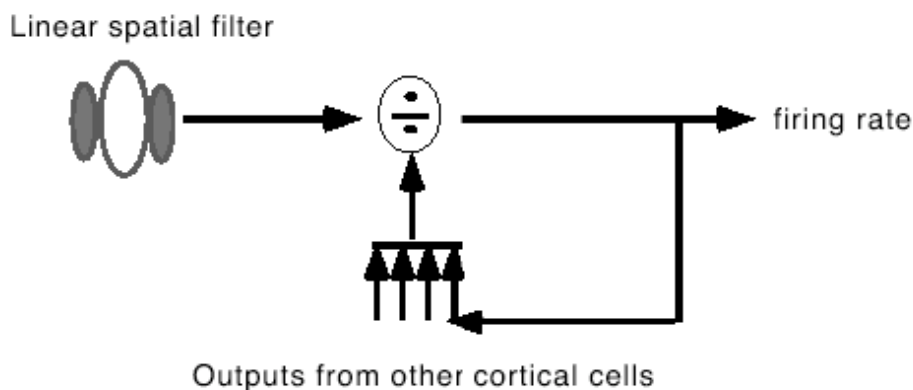

■ Contrast normalization

It turns out that neurons in V1 show an analogous response to your own perception of contrast.

The linear spatial receptive field model for a V1 neuron says that that response should scale linearly with contrast. But simple cells don't show this property--instead, the response begins to saturate at high input contrasts (e.g. for a drifting sinusoidal grating matching the orientation, spatial frequency and motion direction preferences of the cell). Time-wise, the response also begins to occur sooner as the stimulus contrast is increased. Another break-down is seen in the response of a cell to the combination of a horizontal and vertical sinusoidal gratings. Linearity predicts the response to the sum should be the sum of the responses, but it isn't. Instead neurophysiologists find "cross-orientation inhibition". Interestingly enough, a cell that prefers say the vertical grating will typically show zero response to the horizontal one; yet, the presence of the horizontal one still inhibits the cell's firing to the vertical.

A model that quantitatively accounts for the responses of simple cells is called the "contrast normalization" model (Heeger, 1997 ; Carandini, M., & Heeger, D. J. (1994)).

■ Contrast normalization & image statistics

We've noted that a number of image basis sets yield image coefficients that are essentially uncorrelated for natural images. Fourier, and PCA bases all tend to produce uncorrelated outputs for natural images. Wavelets and gabor-filters also tend to produce uncorrelated outputs for natural images. Wavelets may be close to providing an independent components representation for natural images. Independent Components Analysis (ICA) seeks basis sets for which $p(a,b) = p(a)p(b)$, where a and b are coefficients. This is in contrast to PCA which seeks bases for which $E(ab) = E(a)E(b)$ (Recall that independence implies no correlation, but not the reverse).

Consider a standard model for V1 cells consisting of a set of gabor-like filters, or more specifically a wavelet basis set selective for several orientations and spatial scales, spread across the image domain. Despite the fact that the responses are decorrelated, gathering image statistics shows that the wavelet coefficients are not statistically independent.
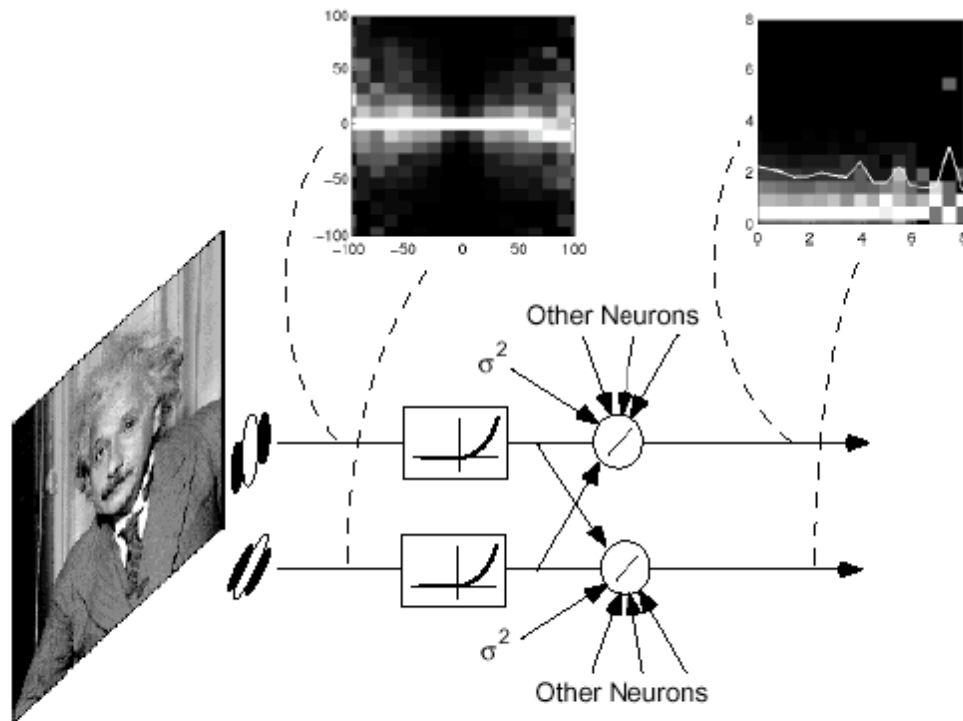
Figure 1: Illustration of image statistics as seen through two neighboring receptive fields. Left image: Joint conditional histogram of two linear coefficients. Pixel intensity corresponds to frequency of occurrence of a given pair of values, except that each column has been independently rescaled to fill the full intensity range. Right image: Joint histogram of divisively normalized coefficients (see text).

Figure from: 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. See: http://www.cns.nyu.edu/~eero/-ABSTRACTS/simoncelli98d-abstract.html

The figure shows that the expected value of the ordinate coefficient is about zero regardless of the abscissa value, i.e. the correlation is about zero. However, the variance of the ordinate increases with the absolute value of the abscissa--the responses are not independent of each other. Simoncelli & Schwartz (1998) report this pattern for pairs of coefficients at neighboring spatial positions, orientation and scales. Remarkably, the statistical dependence goes away if the image filter responses are normalized by the sum of responses of nearby filters. In other words, the contrast normalization model (above) could be the consequence of an efficient coding principle.

(The Appendix of **12.SpatialCodingEfficiency.nb** has code for calculating a 2D histogram plot of joint spatial filter statistics for an image.)

Overview of extrastriate cortex

We've seen how to model the processing of spatial visual information in V1. Twenty 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.

■ Extra-striate cortex

Primary visual cortex sends visual information to many modules (current estimates are about 30 visual areas). These areas have been identified through anatomical, histological, and physiological techniques starting with the early work of Samuel Zeki at the University of London, and the extensive work by David Van Essen and colleagues.

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

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

Connections: anatomical connections are traced using retrograde and anterograde tracers.

Topography:retinotopic maps in each of several of the early visual areas (V1-V8).

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

Maunsell & Newsome, 1987
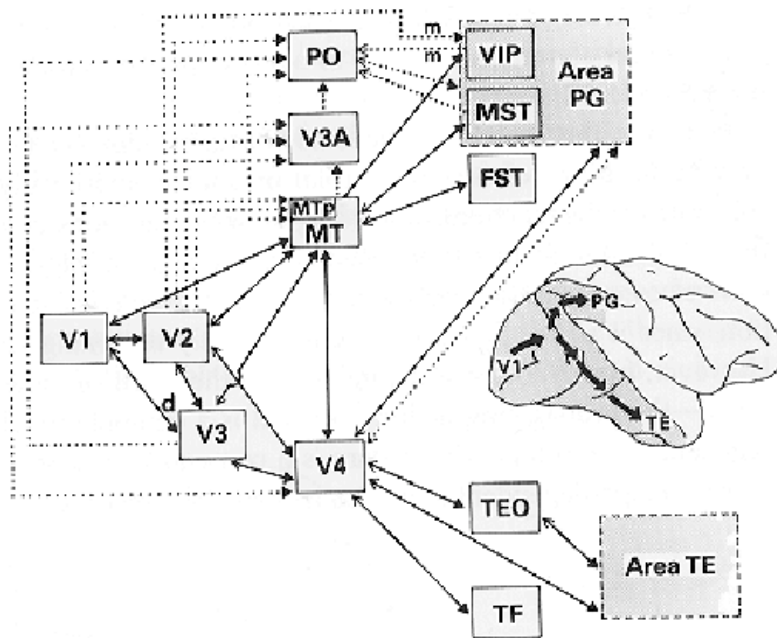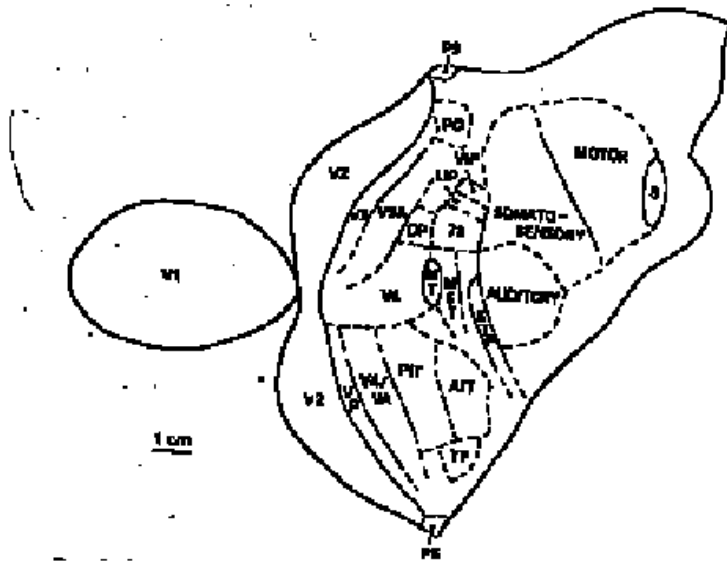
■ More complicated--"plumbing diagram"
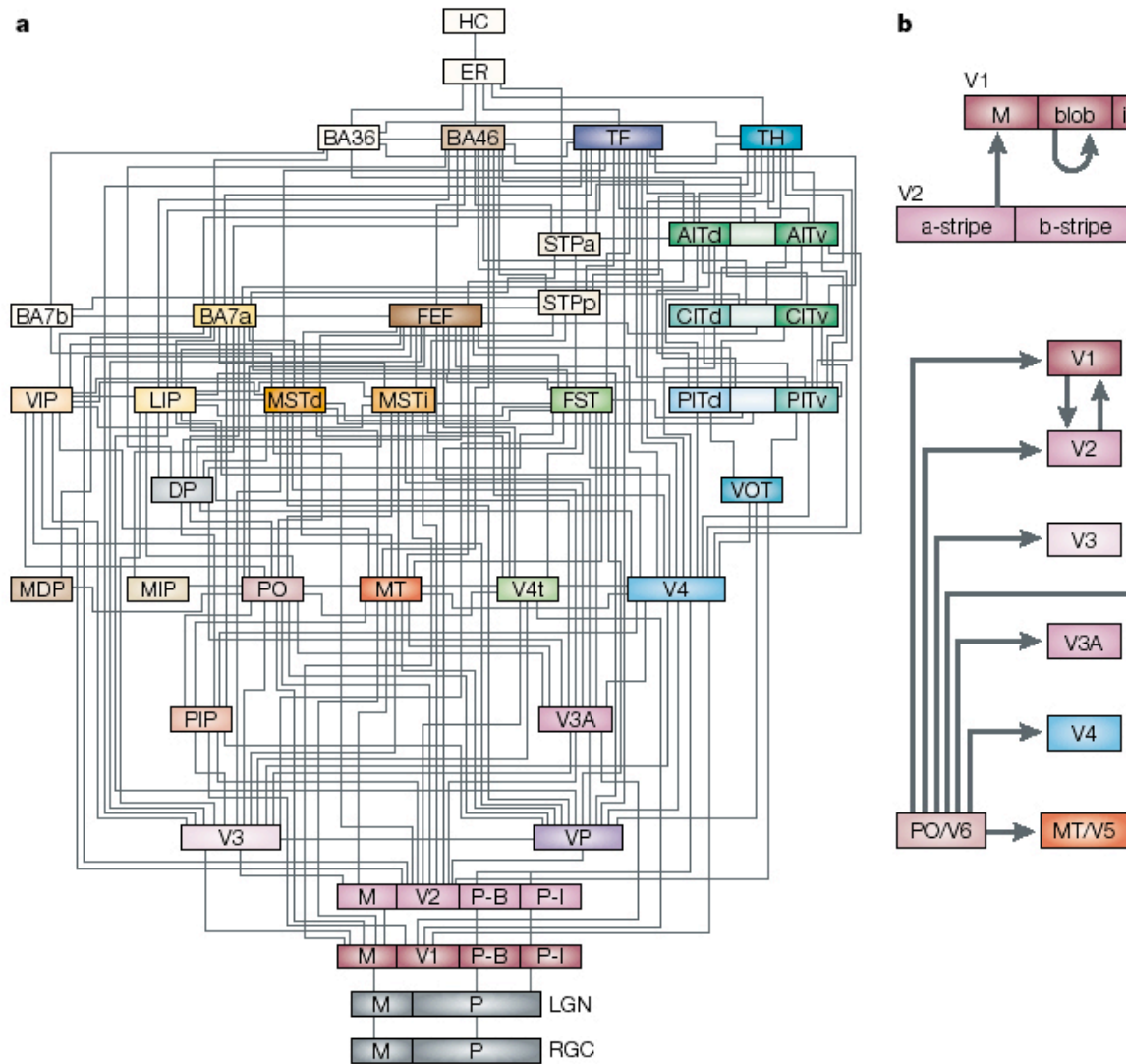
■ Pathways

The flow of visual information through extra-striate cortex is though follow two dominant streams. In the dorsal or parietal areas, information flows from primary cortex to parietal cortex. An example stream would be from areas V1 <-> MT <-> MST.

The temporal stream takes information from primary visual cortex to infero-temporal cortex. One specific sub-stream of information, for example, is: V1 <-> V2 <-> V4 <-> IT.
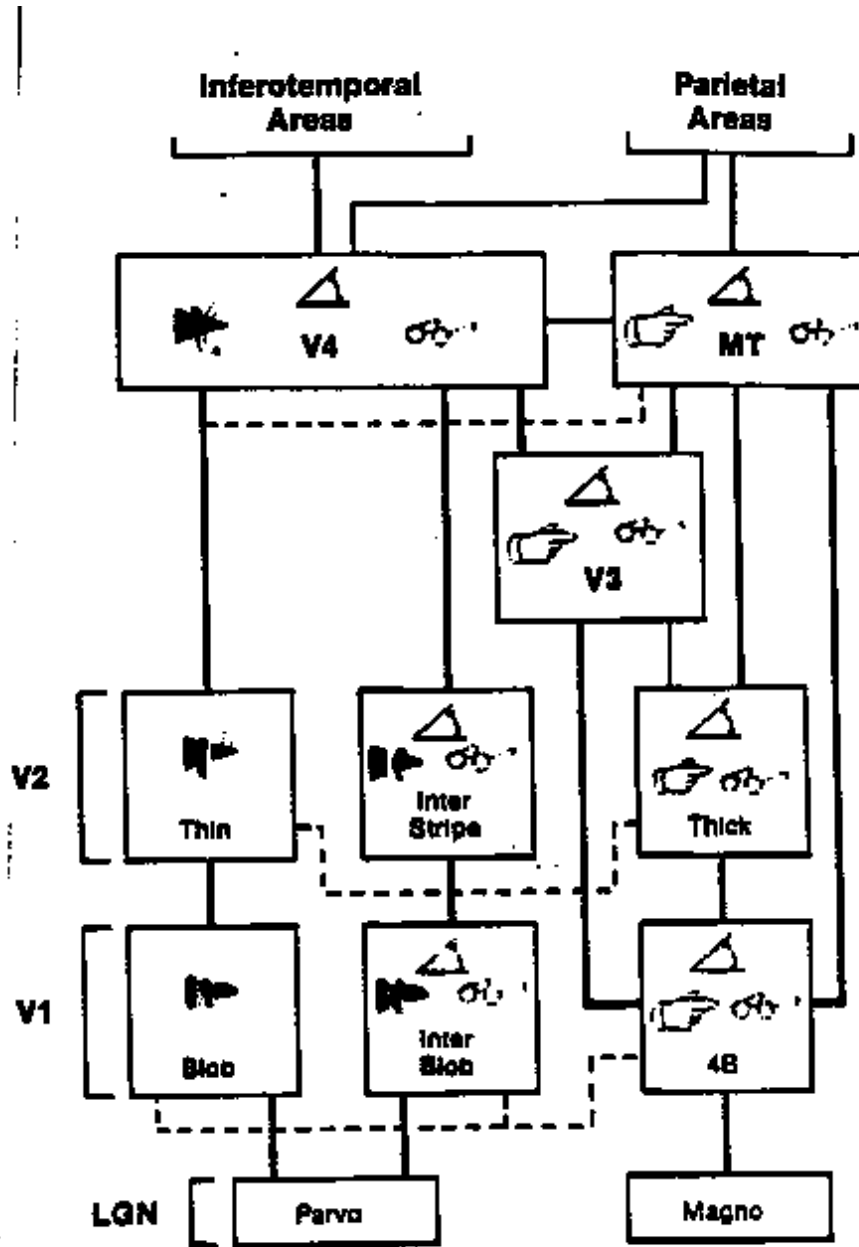
Fig. 3. *Schematic diagram of anatomical connections and neuronal*

One of the remarkable discoveries about extra-striate cortex is the identification of hierarchical organization and multiple parallel pathways (See Felleman and Van Essen, 1991; DeYoe and Van Essen, 1988; DeYoe et al., 1994). Hierarchical organization is identified using anatomical tracing techniques. A general pattern has emerged in which there tend to be:

> • 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.

The figures show examples of some of the interconnections between areas.

What are these extra-striate visual areas of cortex doing? Generally, we have seen that it is involved with recognition and navigation. But it is important to begin to look for detailed computations that extra-striate areas are doing. At the current time, we have only a few ideas, some of which we will look at in the lectures on motion perception.

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. One idea is that information which likely belongs to same object, or have the same causeis which being brought together. Specific problems are: stereovision, motion disambiguation, and color constancy. Some indication of possible function is provided by the icons for color (prism

What are these extra-striate visual areas of cortex doing? Generally, we have seen that it is involved with recognition and navigation. But it is important to begin to look for detailed computations that extra-striate areas are doing. At the current time, we have only a few ideas, some of which we will look at in the lectures on motion perception.

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. One idea is that information which likely belongs to same object, or have the same causeis which being brought together. Specific problems are: stereovision, motion disambiguation, and color constancy. Some indication of possible function is provided by the icons for color (prism wedge), binocularity (spectacles), orientation (angle), and motion (finger pointing) in the figure above.

The fascinating discoveries of 30+ extra-striate visual areas, together with a lack of ideas about what all of these modules are doing, suggests that it might be useful to step back, take a look at what we know so far, and then think about the computations that are required to perceive and act. The next section gives an outline of early visual function, summarizing much of what we have covered, and the last section provides an introduction to the computational problem of going from image to scene representations.

■ Dominant functional streams

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 Ungerleider, 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 non-view centered representations useful for object recognition. 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 eventually 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 eventually, temporal areas (IT) for object recognition

■ General Extra-striate Functions

Bring together local information/measurements from distant parts of the visual field likely to belong to same object, or have the same cause

e.g. stereovision for surfaces, object motion, color constancy of surfaces, ...

■ **Major computational theme**

Suggests a major computational theme: How to integrate local ambiguous measurements to arrive at functional representations of objects and their relationships to each other and to the viewer?

Scenes from images

We know quite a bit about the brain's early processing of image information. But what we have learned tells us rather little about perception. 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 as much as 40-50% of visual cortex may be involved in visual processing, including both the primary visual area and extra-striate areas. 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 and processing strategies that are efficient in achieving a goal. In fact, 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 output. 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. 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, how can I assign a relatively stable color to it? This is the problem of color constancy. In particular, 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 a shading 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. Succeeding stages preserve information, but with progressive selection aimed at the goals of the visual system.

■ **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. Another (not necessarily exclusive) view is to more directly extract useful parameters for function (e.g. geometric shape dimensions for object recogniiton, or time-to-contact for braking).

■ **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".

When we consider visual tasks however, it is useful to consider generative models that are "closer to the output" of vision.

So one 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).
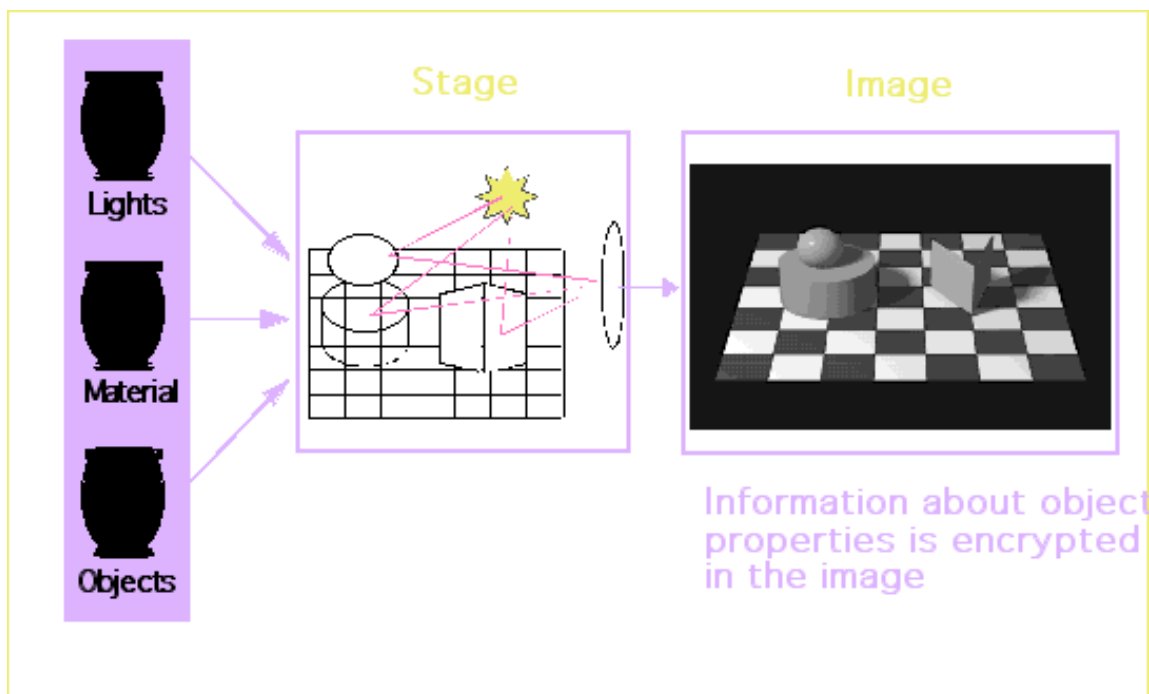
■ The inverse 3D graphics metaphor

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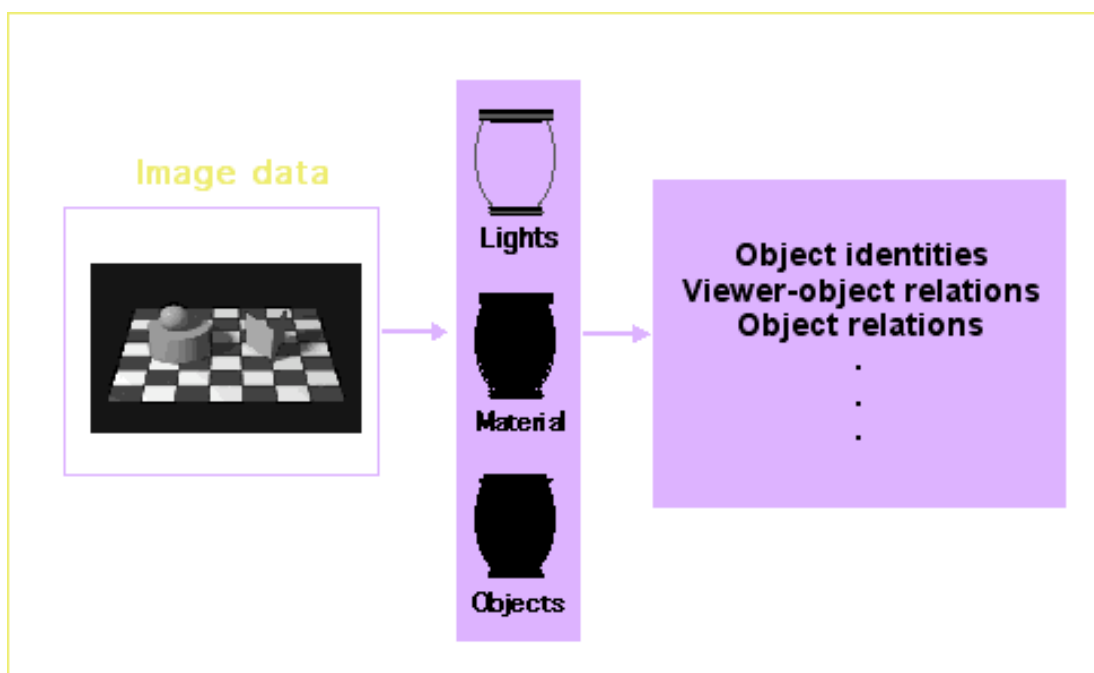
The inverse problem is to take as input, **L**, and compute the scene causes **R** , **N**, **V** or **E**. Although it is unlikely that human vision exactly solves the inverse graphics problem even in small domains, the metaphor is useful (Kersten, 1997) to make explicit image ambiguities and to test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

Preview of the role of ideal observers & Bayesian decision theory

It is useful to start 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 (Knill and Kersten, 1991; Kersten and Madarasmi). 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 casts many of our 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.
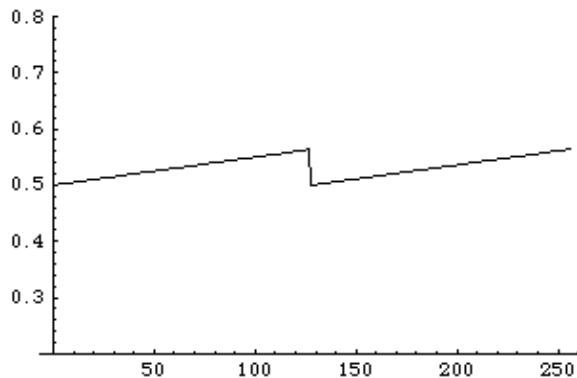
How does the inverse optics or graphics view differ from efficient coding?

Knill & Kersten revisited

■ Land & McCann's "Two squares and a happening"

The left half looks lighter than the right half. But, let's plot the intensity across a horizontal line:


The two ramps are identical...tho' not too surprising in that that is how we constructed the picture. How can we explain this illusion based on what we've learned so far?

One explanation is that the visual system takes a spatial derivative of the intensity profile. Recall from calculus that the second derivative of a linear function is zero. So a second derivative should filter out the slowly changing linear ramp in the illusory image. We approximate the second derivative with a discrete kernel (-1,2,-1).

The steps are: 1) take the second derivative of the image; 2) threshold out small values; 3) re-integrate

■ **Knill & Kersten's "Two cylinders and no happening"**
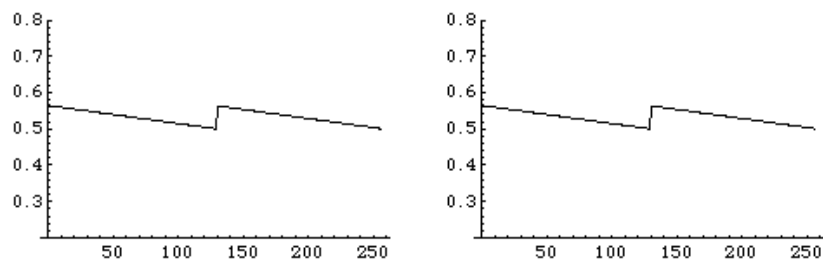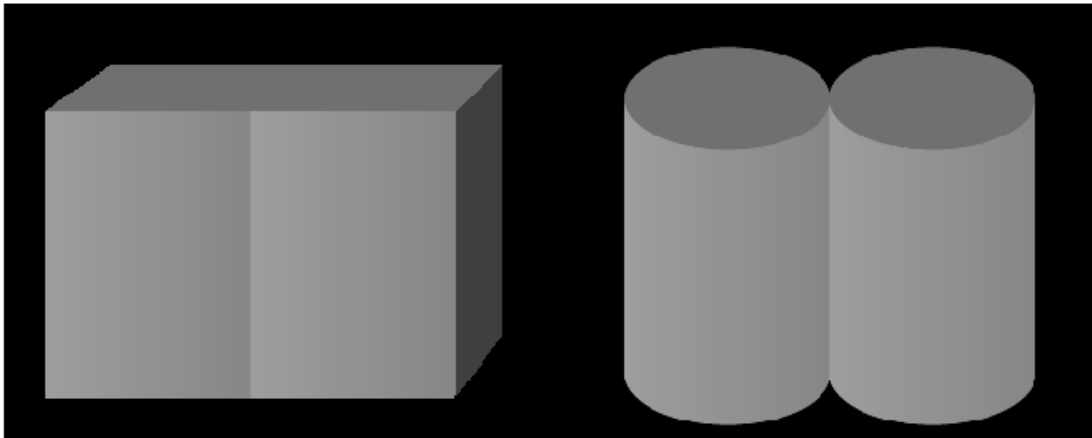
But the perceived lightness contrast for the slabs is significantly stronger than it is for the two cylinders. A spatial convolution/derivative model would predict the same for both. The spatial convolution operation won't work as an explanation! So what will?

■ **The inverse graphics metaphore & "the two-cylinders & no-happening"**
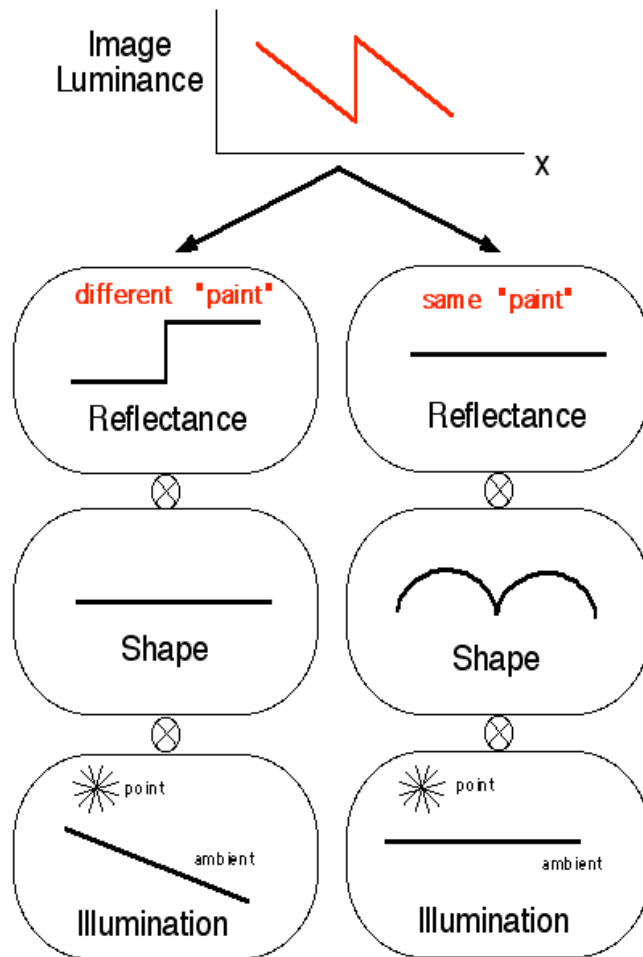
Low-level ("early"), intermediate-level (middle), and high-level vision.

Low-level--local measurements, simple grouping procedures

Intermediate-level--surfaces and surface-like representations, objects,..

High-level--functional tasks, object recognition, reaching,...

Appendices

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.

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, restora-tion 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