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

Initialize

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 homework**
- **Mathematica Demonstrations**
- **Mathematica Demonstrations Illusions**
- **Extrastriate cortex--overview**
- **Scenes from images, scene-based modeling of images**

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 (current estimates are about 30 visual areas in the macaque). These areas have been identified through anatomical, 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 their inputs

Architecture: 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-V8).

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.

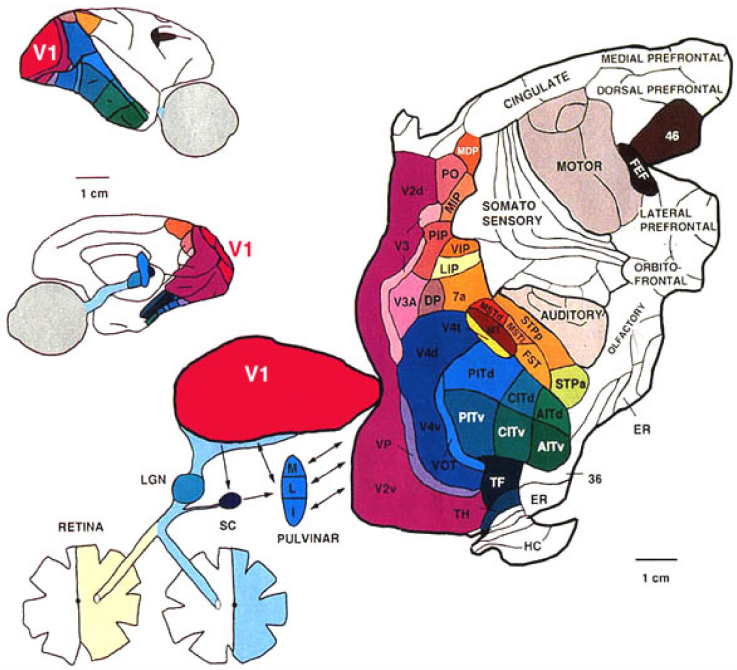
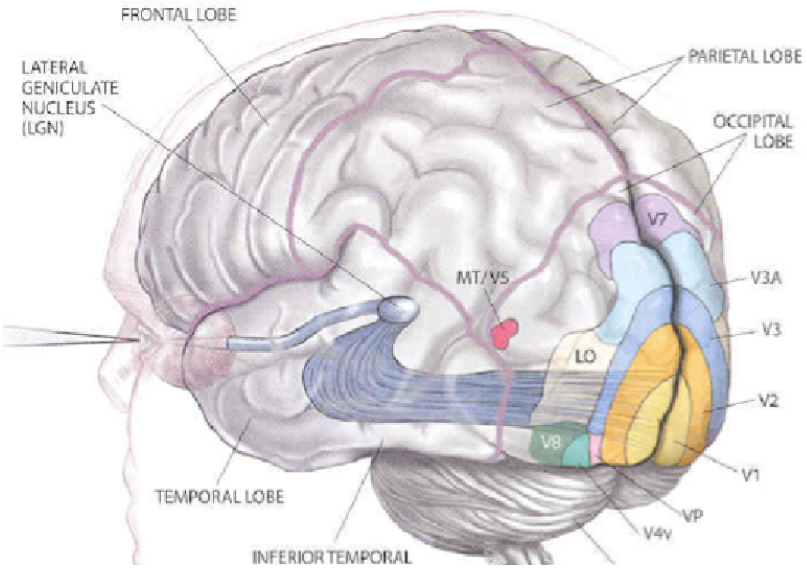


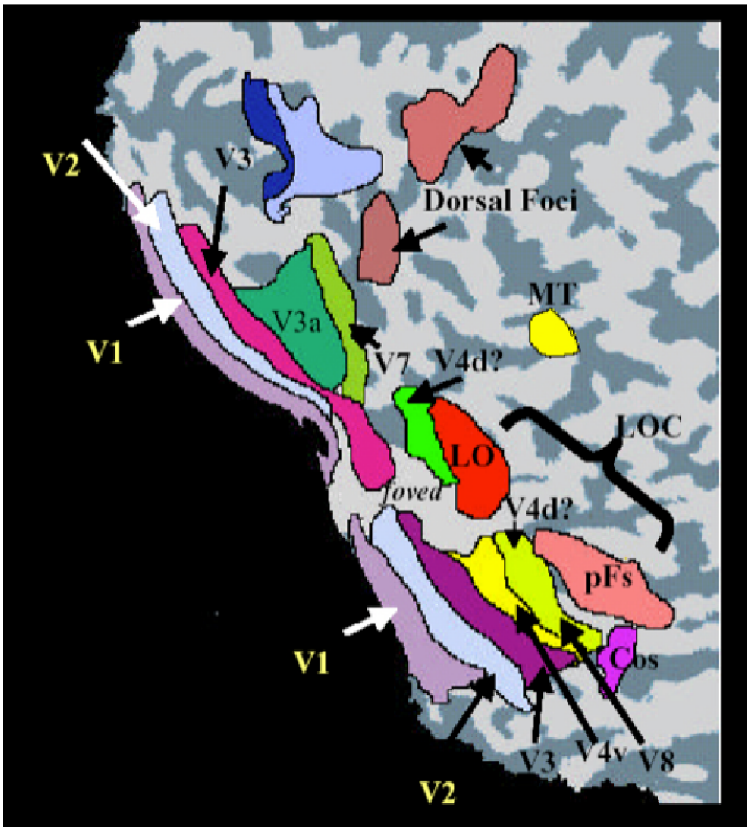
Figure 19. Much of V1 is located in the calcarine sulci and its relationship to other brain areas is best shown by unfolding the brain and showing it flattened open. The visually responsive areas of the macaque monkey are shown in color. From Van Essen et al. (1992).

From Van Essen et al. 1992

■ Human visual areas



From Scientific American,



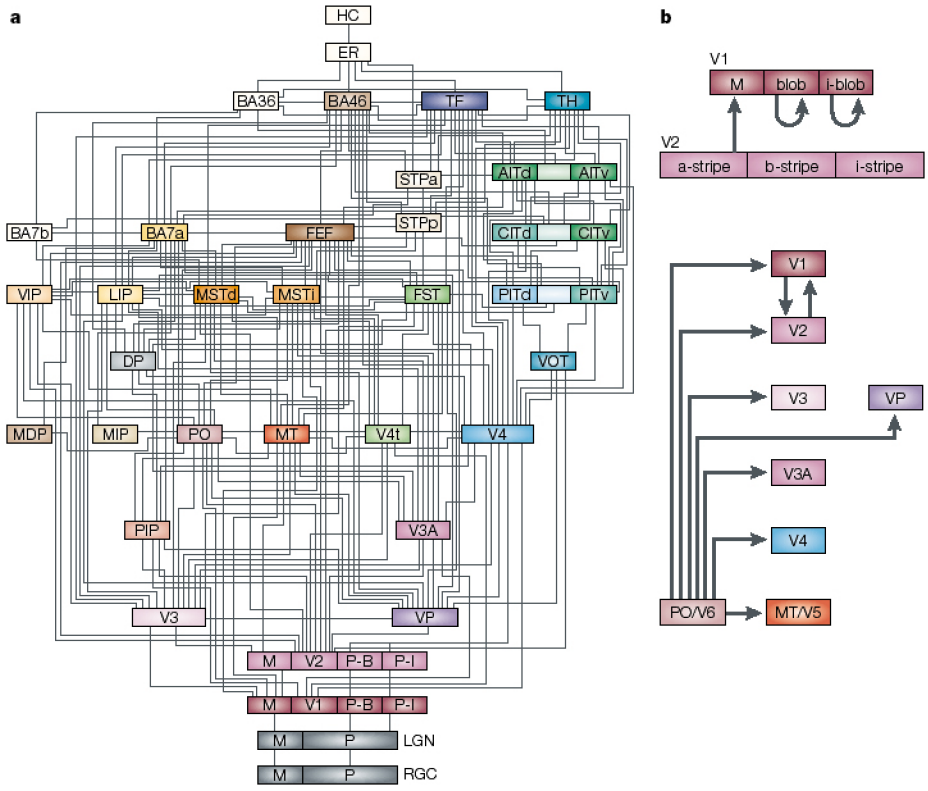
■ Visual hierarchy

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 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.



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 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 cause, is what is being brought together.

A few problems are:

stereovision

motion disambiguation

color constancy

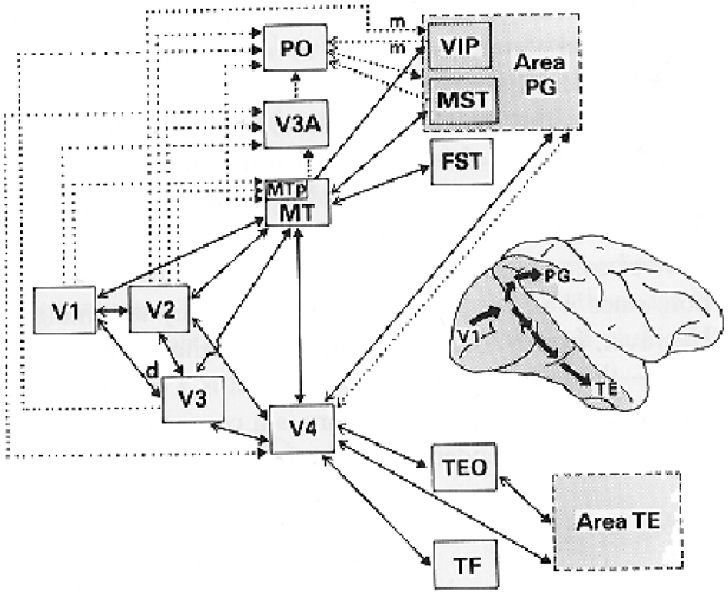
object contours & regions

Some indication of possible functional distinctions are illustrated below for smaller scale pathways.

■ 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.



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-viewer 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 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

■ Smaller scale pathways

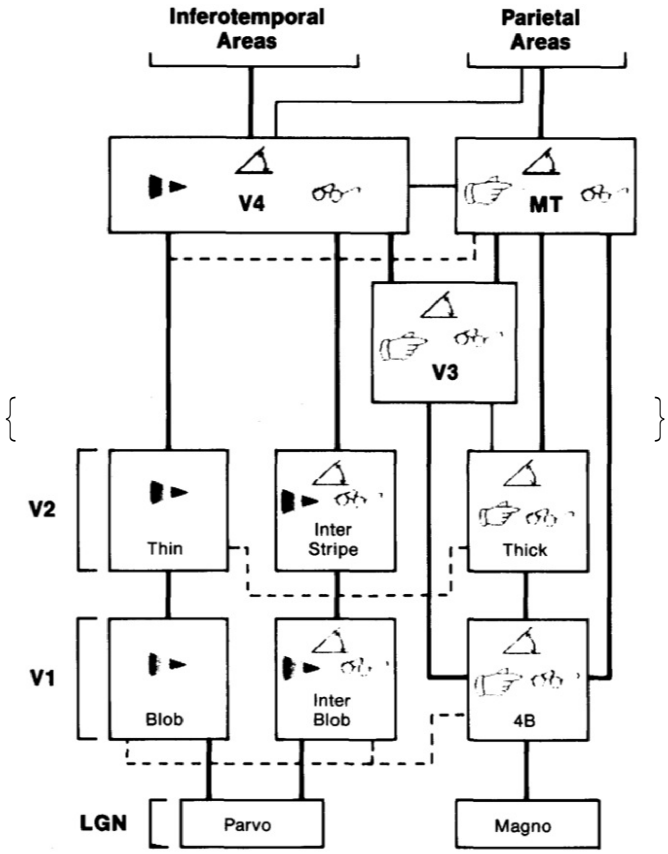


Fig. 3. Schematic diagram of anatomical connections and neuronal selectivities of early visual areas in the macaque monkey. LGN = lateral geniculate nucleus (parvocellular and magnocellular divisions). Divisions of Vl and V2: blob = cytochrome oxidase blob regions; interblob = cytochrome oxidase--poor 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)

■ General Extra-striate Functions

The fascinating discovery 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 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 actions.

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

Low-level--local measurements, simple grouping procedures

Intermediate-level--surfaces and surface-like representations, more global grouping processes, objects,..

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

Scene information from images

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 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. 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 dimensions 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

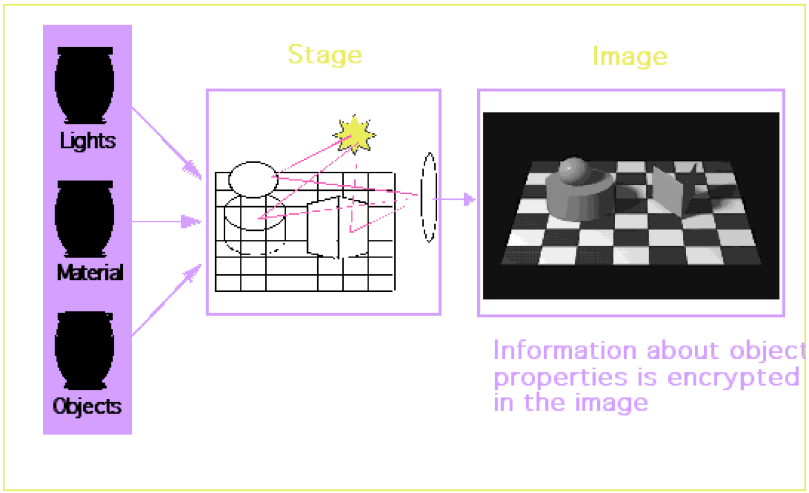
The image filtering approach can be thought of as primarily "image-based". The advantage of image-based modeling is that it is "closer to the input". Features are indexed spatially which makes sense given topographic representations. But as we begin to think about representing object properties, it may 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):

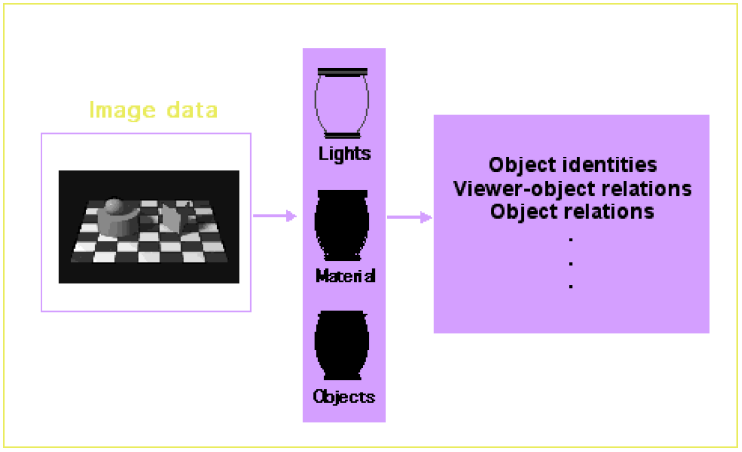
$$\mathbf{L(x,y) = f(R(x,y), N(x,y), V, E)}$$

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? 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.

Lecture 6 provided a preview of how Bayesian decision theory could be used to develop a framework for estimating scene properties from images.

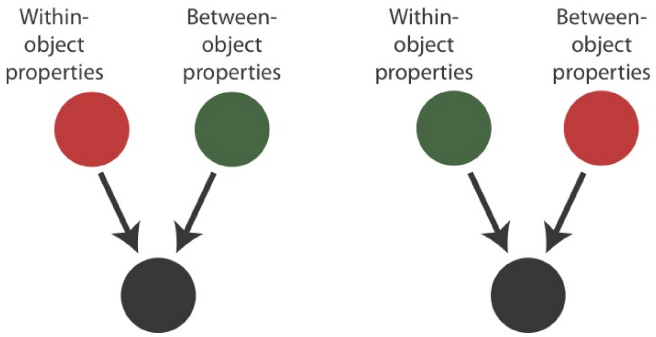
In particular, the ideal observer can be modeled as a Bayesian estimator of scene parameters, given image data. E.g. the MAP observer would pick the most probable scene given a fixed set of image measurements based on the posterior probability

p(scene | image measurements)

This formulation 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.

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.



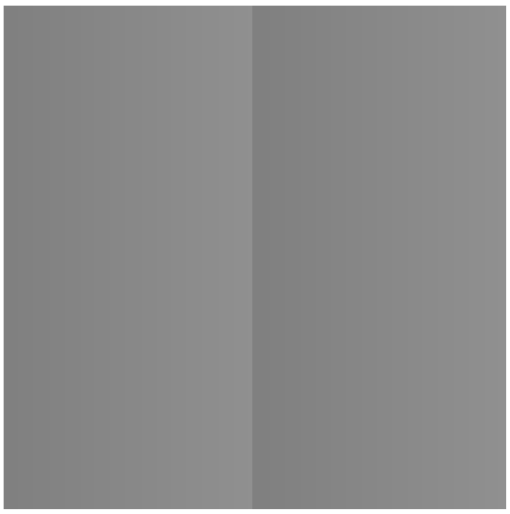
How would the computational problems of size estimation differ for the tasks of grasping an object vs.

recognizing an object?

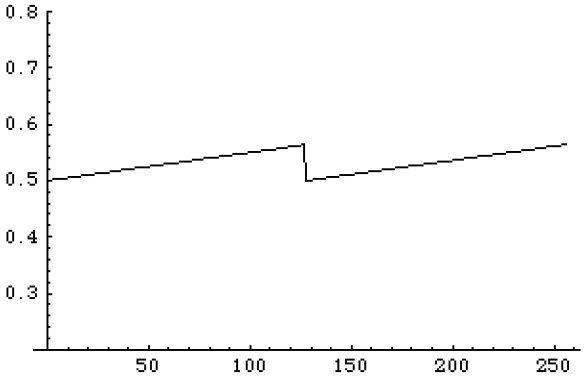
How does the inverse optics or graphics view differ from efficient coding?

Two cylinders lightness illusion 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:



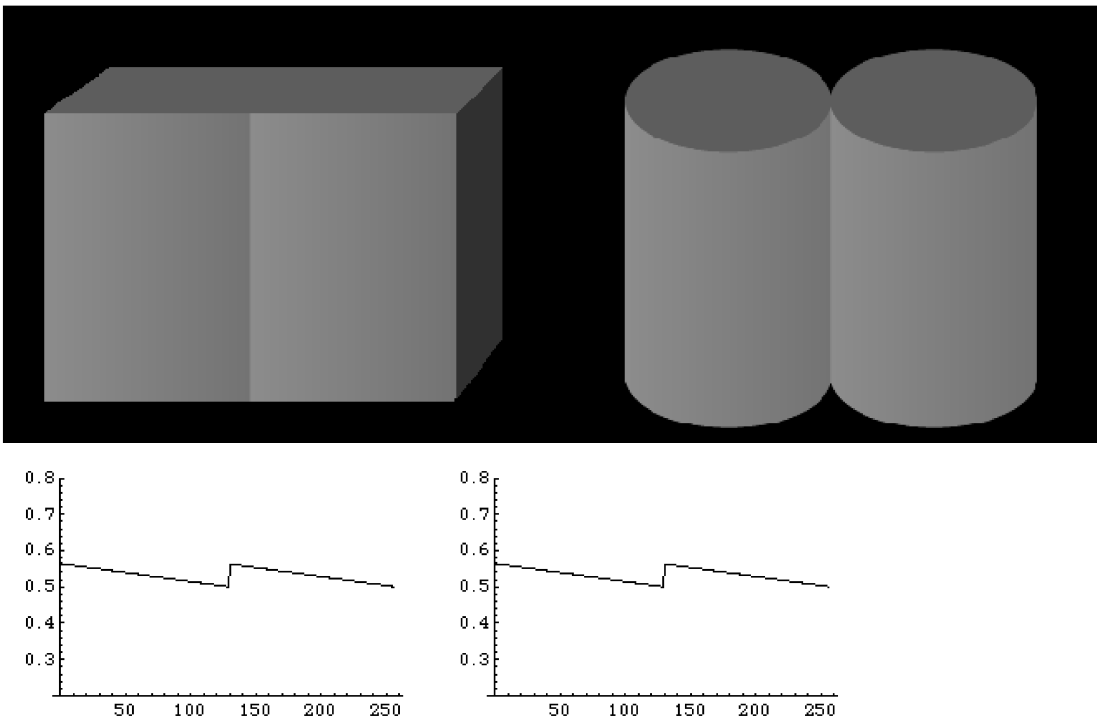
The two ramps are identical...tho' not too surprising in that that is how we constructed the picture. How can we explain this illusion based on what we've learned so far?

We saw that 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

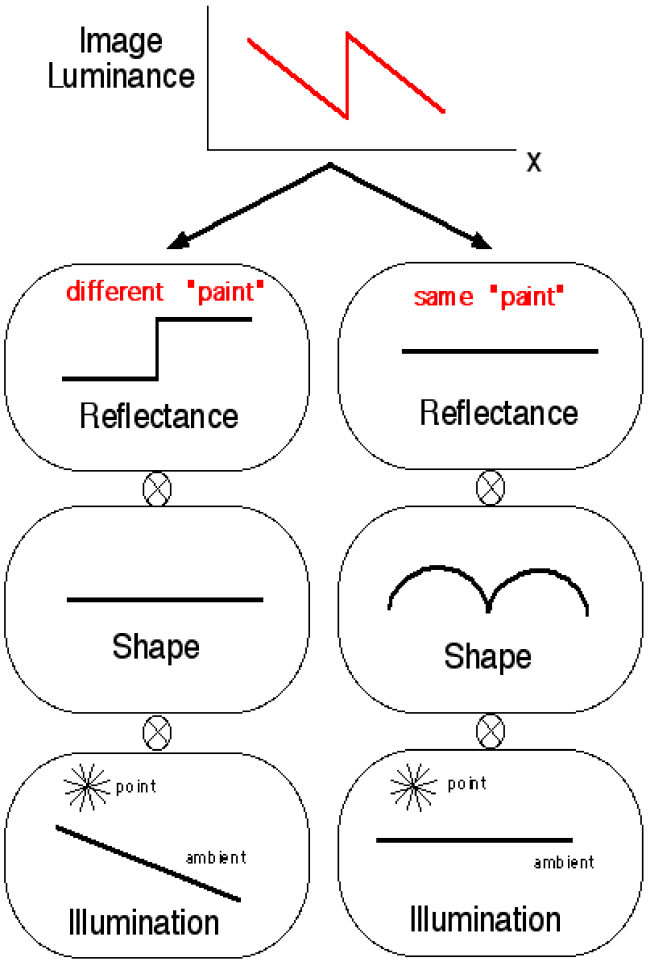
Relatively speaking, this is computationally straightforward.

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 metaphor & "the two-cylinders & no-happening"



In comparison with the image-based models of lightness perception developed for illusions such as the Land-McCann, inverse optics computations are hard.

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, 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.e-du/~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