Probabilistic Models of the Cortex: Stat 271.

Alan L. Yuille.

UCLA.

Goals of the Course

- To give an introduction to the state of the art computational models of the mammalian visual cortex.
- To describe the relevant evidence from anatomy, electrophysiology, imaging (fMRI), and psychophysics.

Course Materials

- Latexed Notes (D. Kersten and A.Yuille).
- Papers – reading list.
- Handouts.

- Grading: Homework Assignments and Course projects.

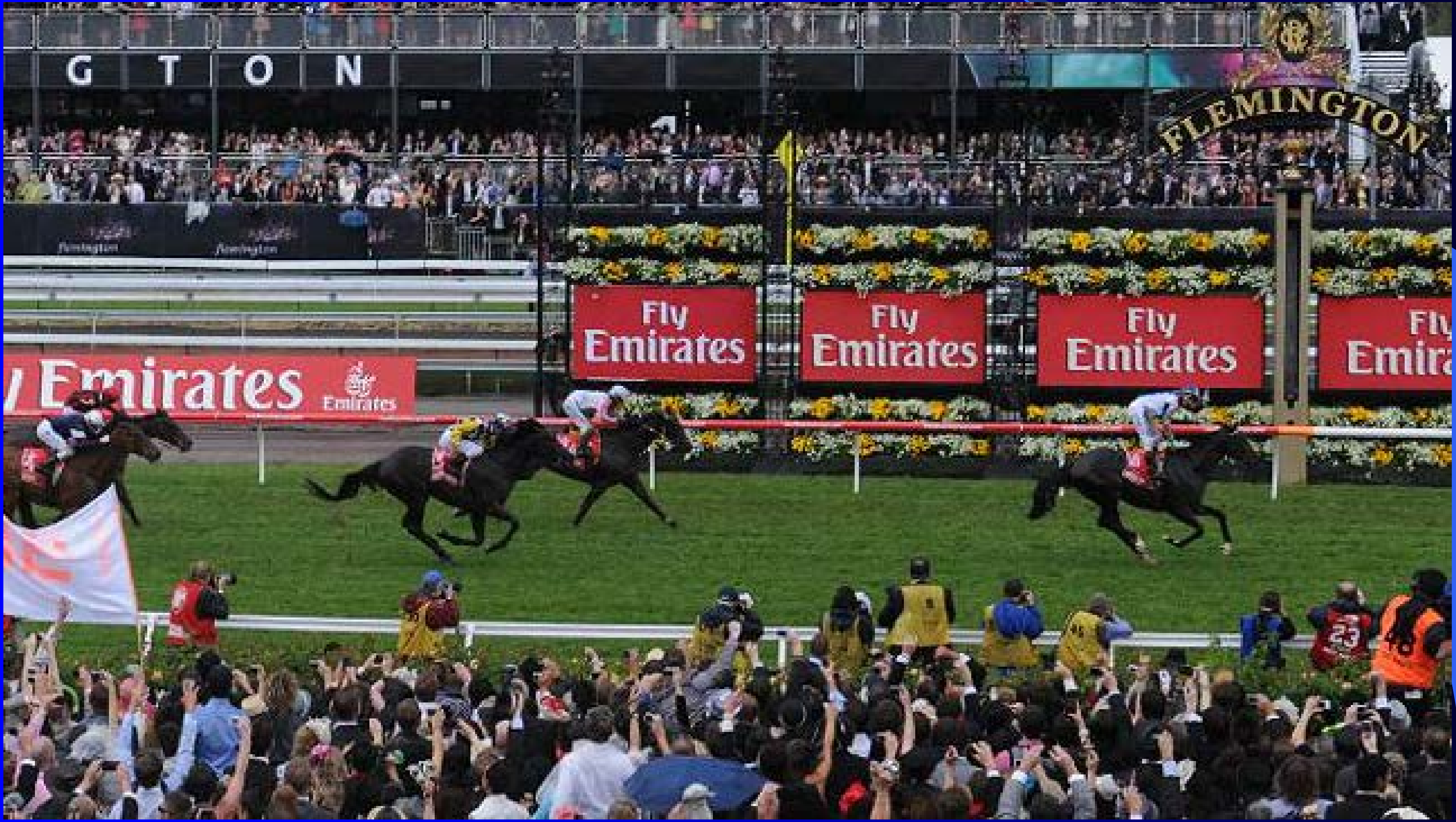
Main Points of this Talk

- (I) Vision is very hard. Humans are vision experts.
- (II) The key problems: complexity, ambiguity, and invariance.
- (III) Visual models and hierarchies.
- (IV) The visual system is extremely complex and only partly understood.

The Purpose of Vision.

- “To Know What is Where by Looking”. Aristotle. (384-322 BC).
- Information Processing: receive a signal by light rays and decode its information to understand the scene.
- *Vision appears deceptively simple, but this is highly misleading.*

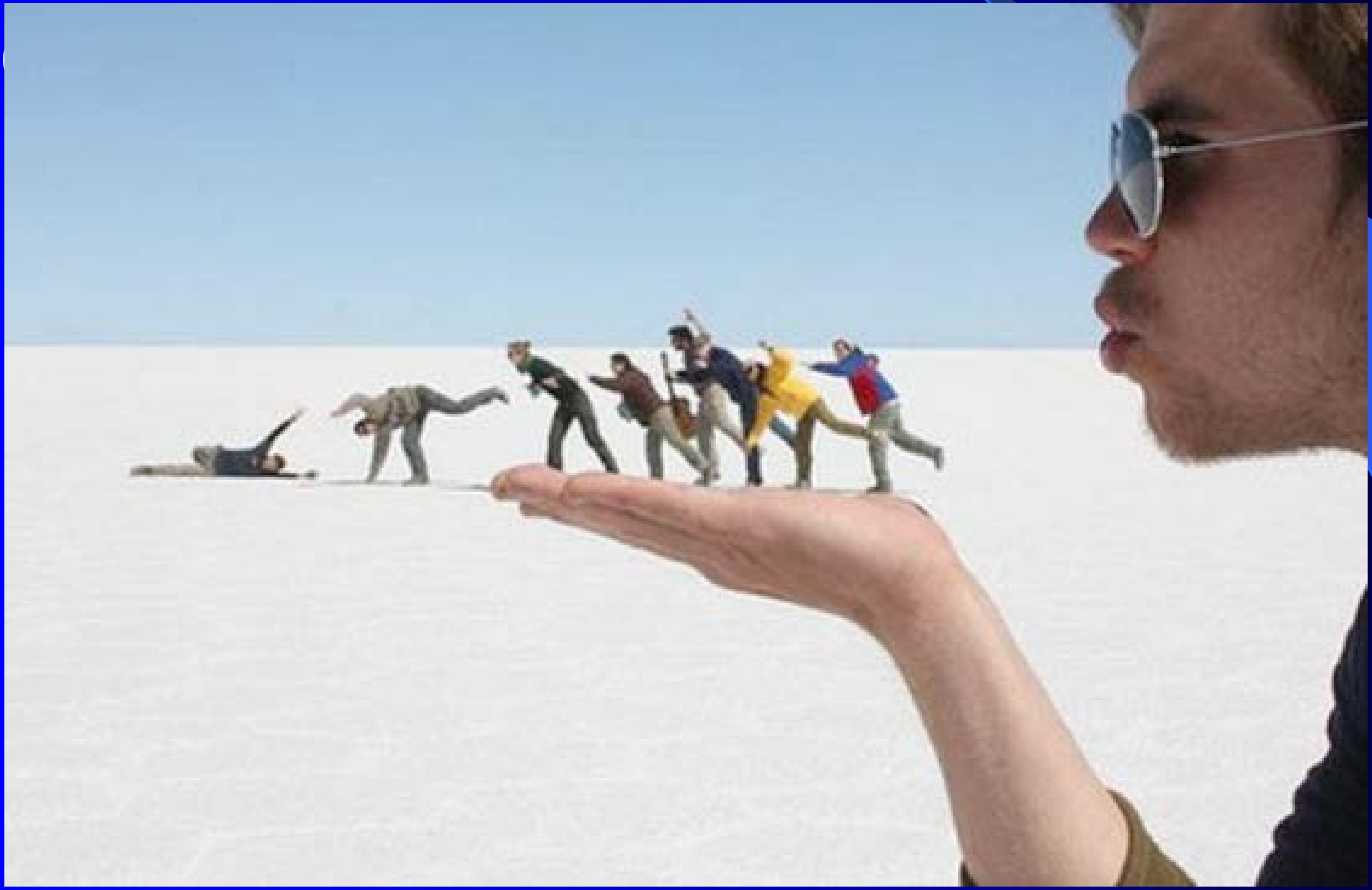
Humans can understand complex images.



But sometimes we make mistakes.

- Perception is not reality.

We are fooled by accidental alignments

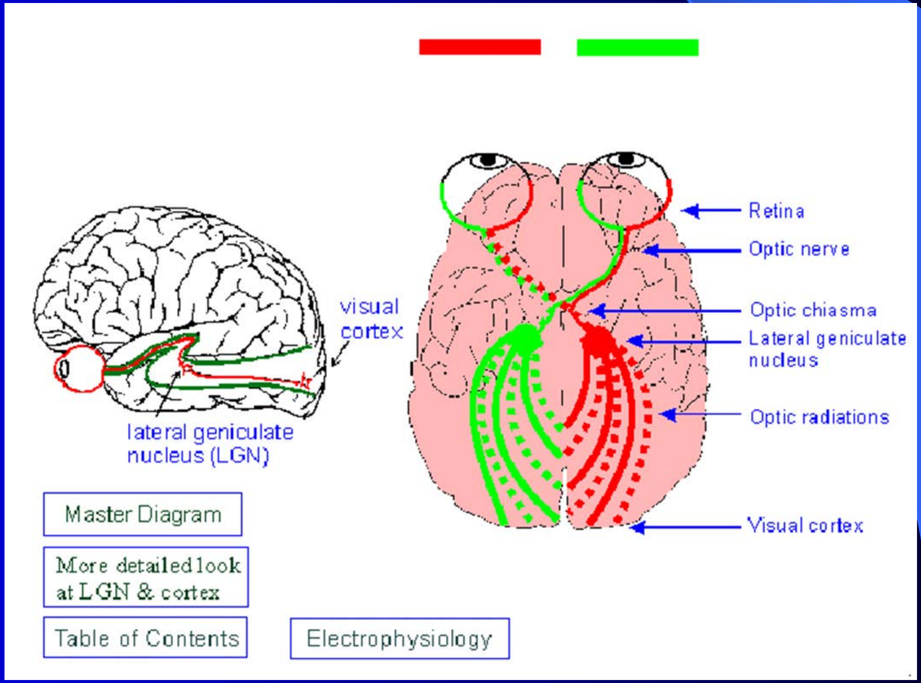


Why are humans good at vision?

- Humans appear understand images effortlessly. But this is only because of the enormous amount of our brains that we devote to this task.
- *It is estimated that 40-50% of neurons in the cortex are involved in doing vision.*
- Humans are very visual. We get much more information from our eyes than other animals.

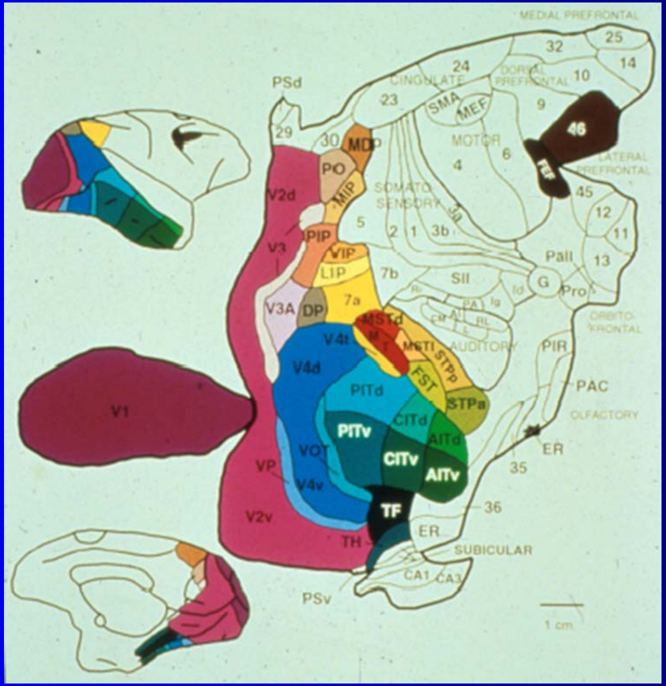
Vision: The retina and optic nerve.

- Images are captured at the retina.
- They are transmitted to the visual cortex by the optic nerve.



The Visual Cortex

- Vision/perception is performed in the visual cortex. This is organized hierarchically.
- The visual cortex is roughly 40% of the entire cortex.

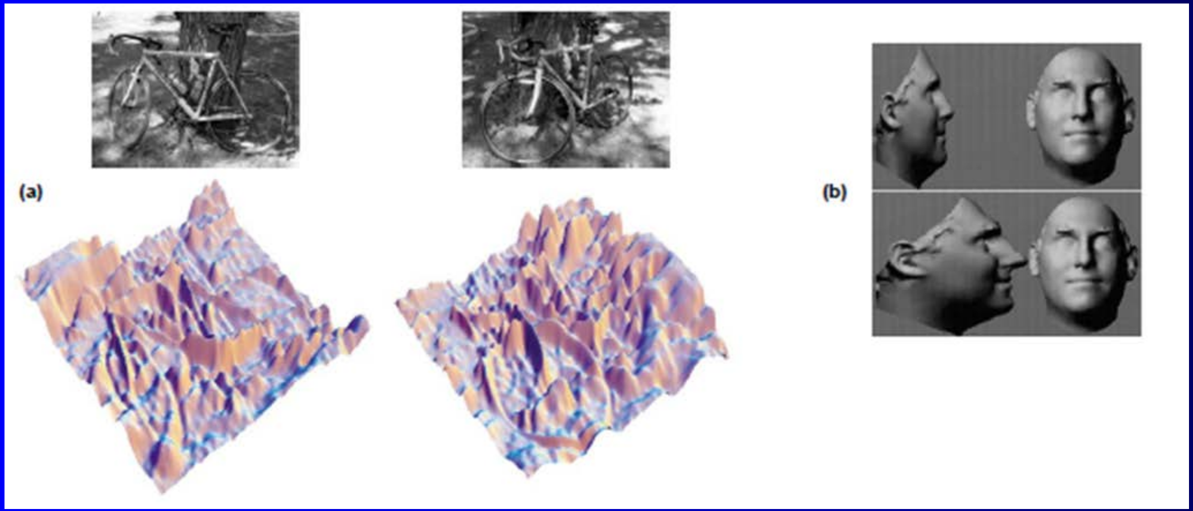


A Brief History of Artificial Intelligence and Vision

- Initially AI researchers thought vision was easy. “Solve vision in a summer”. (1966).
- *But the difficulty of vision rapidly became clear as researchers tried to get computer programs to interpret images. Nothing worked.*
- Researchers started realizing that vision was much harder than “intelligence tasks” like playing Chess.

Why is Vision Hard?

- Look at the raw input displayed as a set of numbers which plot the intensity as a function of position (bottom left).
- The images are very complex. They are of the same bike and tree. But they look very different.



Complexity of Images and Visual Scenes

- The set of all images is practically infinite.
- *Only a tiny fraction of all possible images have been seen by humans.*

- The number of visual scenes is also enormous. There are 30,000 different types of objects. They can be arranged into 1,000 scene categories.

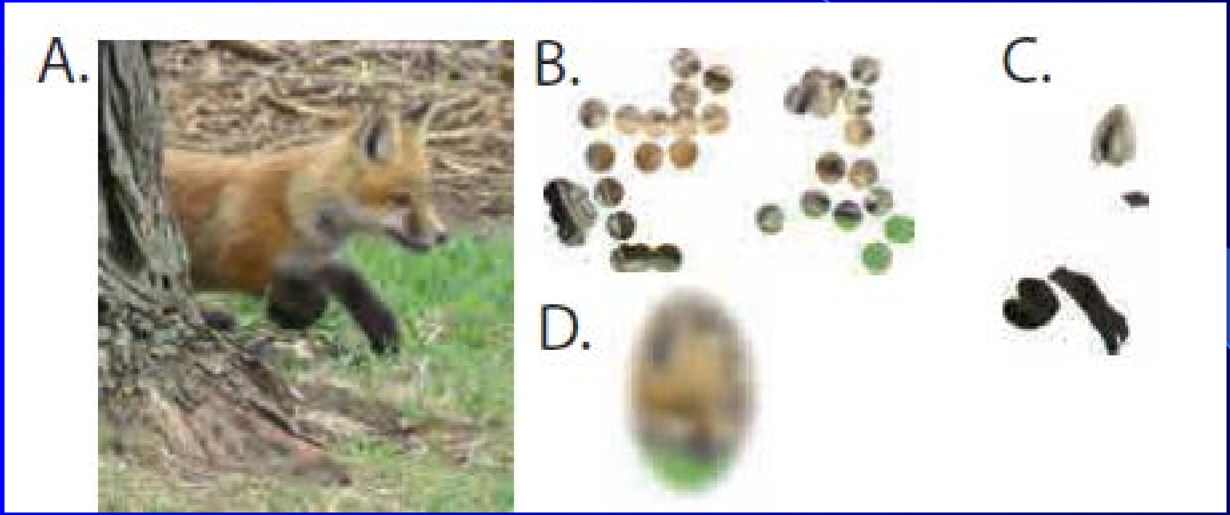
Vision as Inverse Inference

- Images are generated by highly complex processes – light bounces off objects and is captured by the retina/camera.
- This process is studied in Computer Graphics. It models objects, light sources, and how they interact.
- *Vision must invert this image formation process to estimate the "causal factors" – objects, lighting, and so on.*

Local regions are ambiguous

Airplane
Car
Boat
Sign
Building

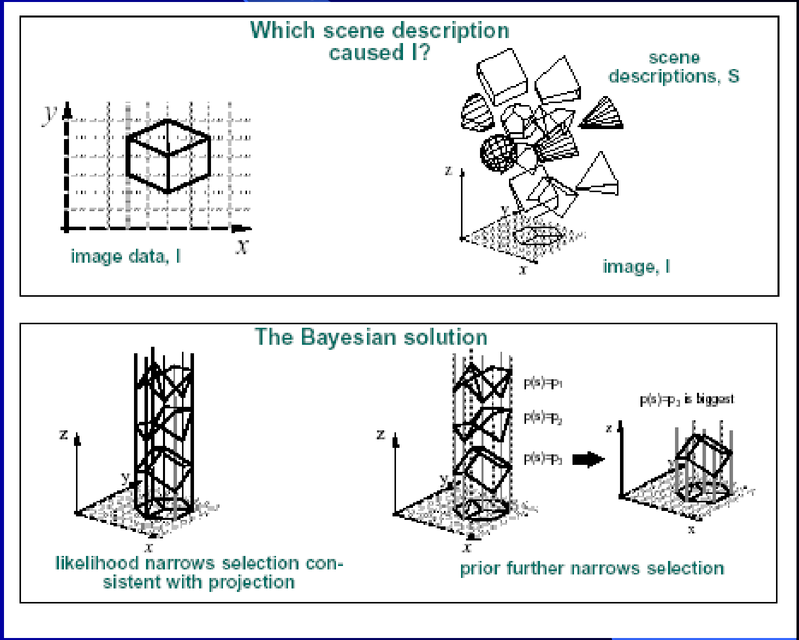
Humans can estimate rich descriptions.



- Humans can easily detect the fox, the tree trunk, the grass and the background twigs.
- And can also estimate the shape of the fox’s legs and head, its type fur, what it is doing, is it old or young, is this winter or summer?
- But the local regions are highly ambiguous.

Inverse Problems are hard

- There are an infinite number of ways that images can be formed.
- Why do we see a cube?
- Prior – cubes are more likely than other shapes consistent with the image.


Which scene description caused I?
scene descriptions, S
y
x
image data, I
z
y
x
image, I
The Bayesian solution
z
y
x
likelihood narrows selection con-sistent with projection
p(s)=p1
p(s)=p2
p(s)=p3
p(s)=p3 is biggest
z
y
x
prior further narrows selection

Bayesian Decision Theory

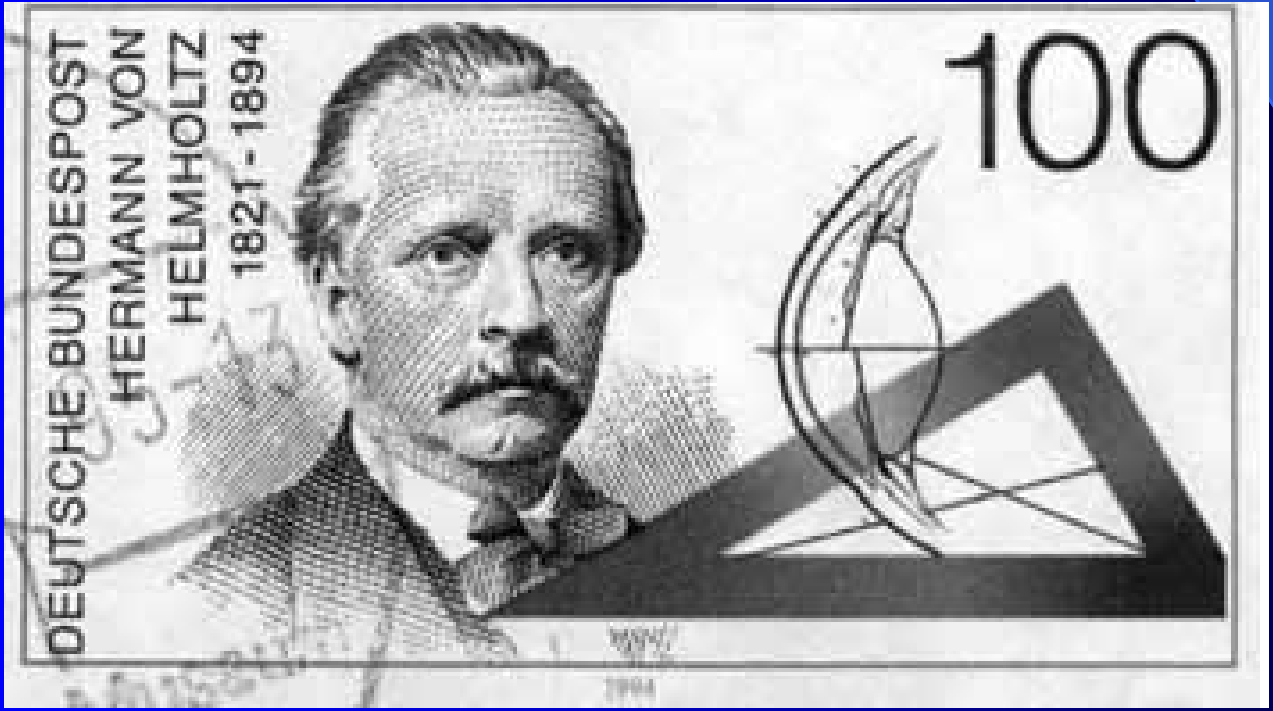
- Bayes’ Theorem gives a procedure to solve inverse inference problems.
- It states that we can infer the state S of the world from the observed image I by using prior knowledge.
- $P(S|I) = P(I|S)P(S)/P(I)$.
- Rev. T. Bayes. 1702-1761

Bayes and Machine Learning

- Bayes gives a conceptual framework for visual perception.
- But it is far from being sufficient. Many techniques have been adapted from CS, Engineering, Mathematics, and Statistics.
- *In particular, machine learning methods have become increasingly effective.*

Vision as Inference

- Helmholtz. 1821-1894.
- “Perception as Unconscious Inference”.

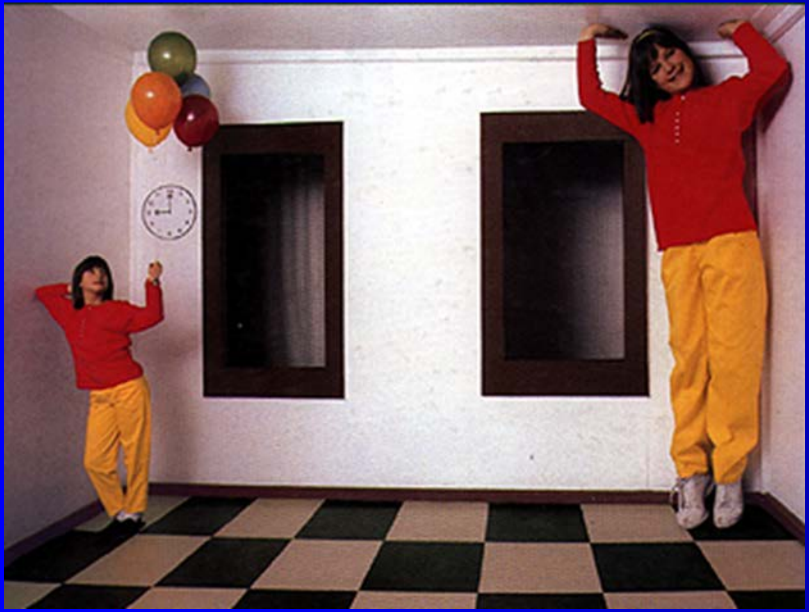


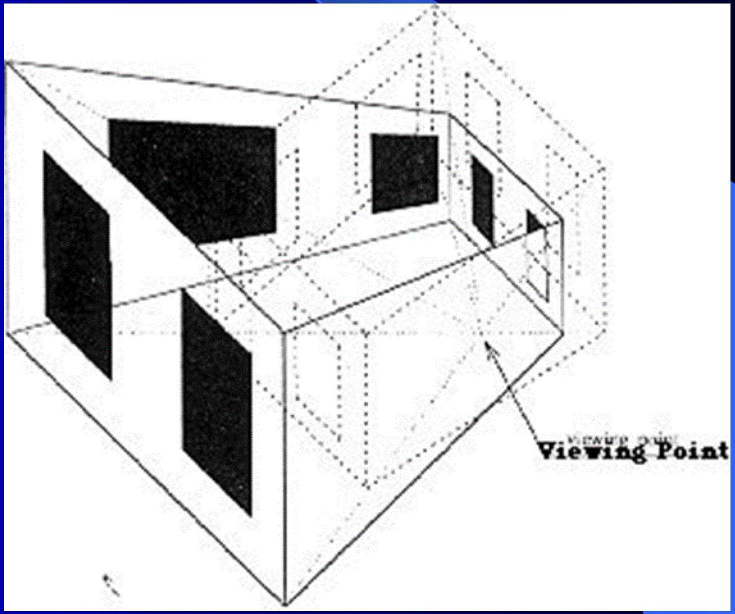
But human vision is not perfect

- Human perception is often incorrect.
- Visual illusions suggest that perception is often a reconstruction, or even a controlled hallucination.

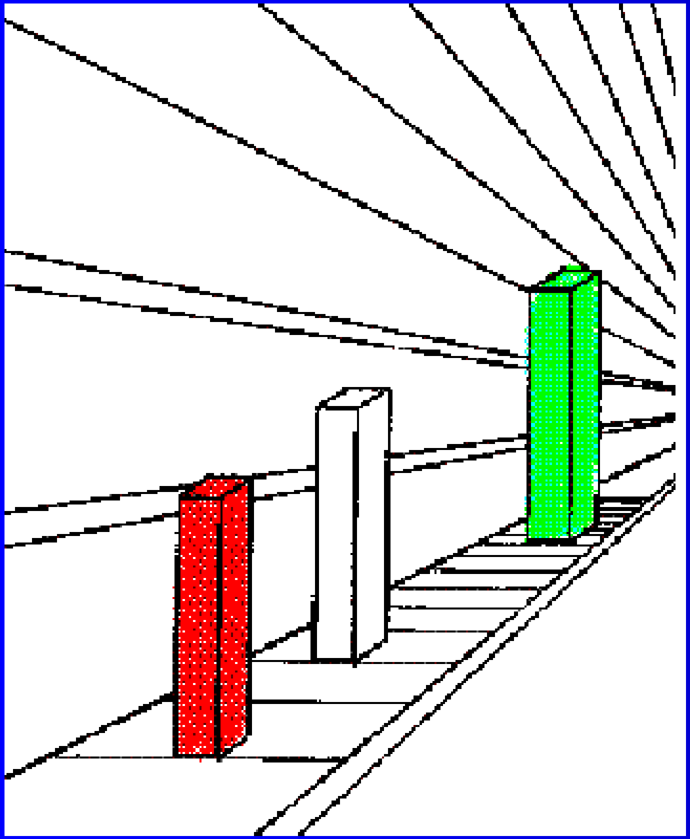
Ames room

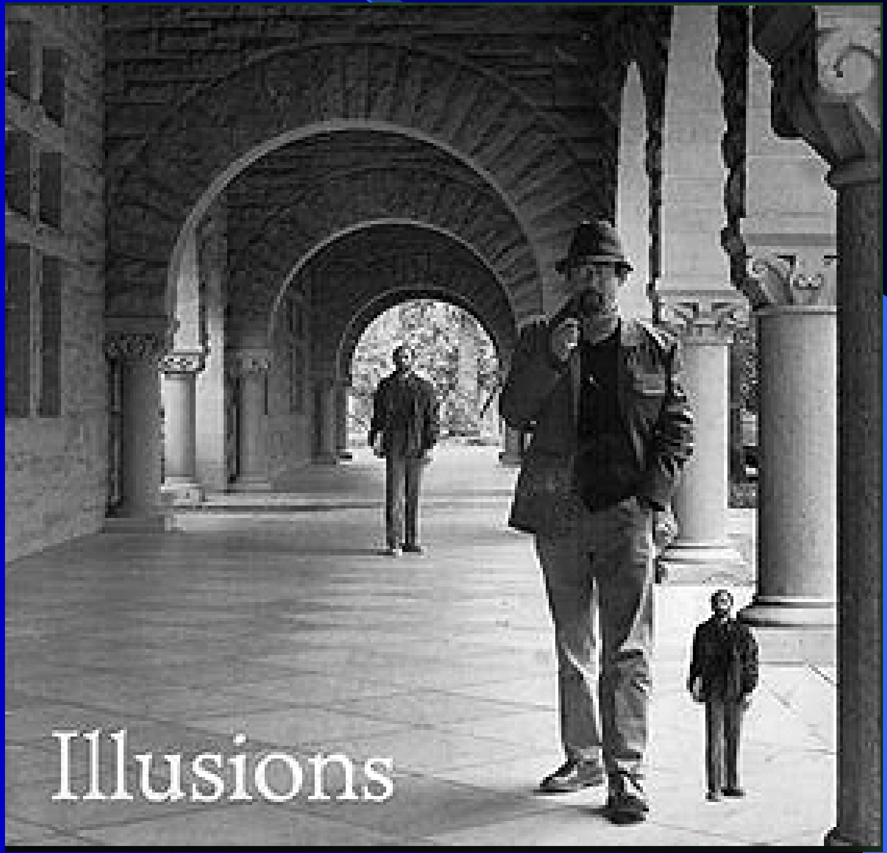
- Which girl is bigger? Trick question.



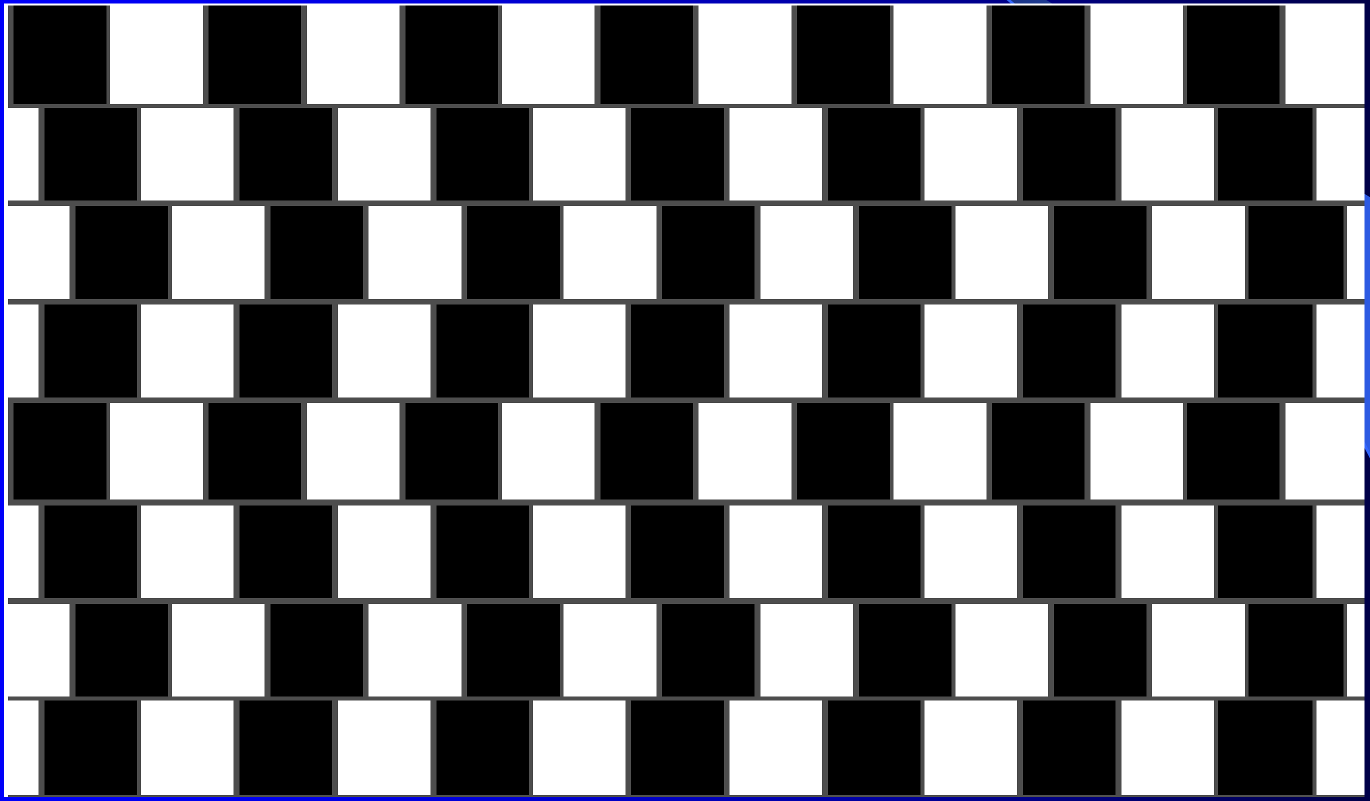


Perspective Cues

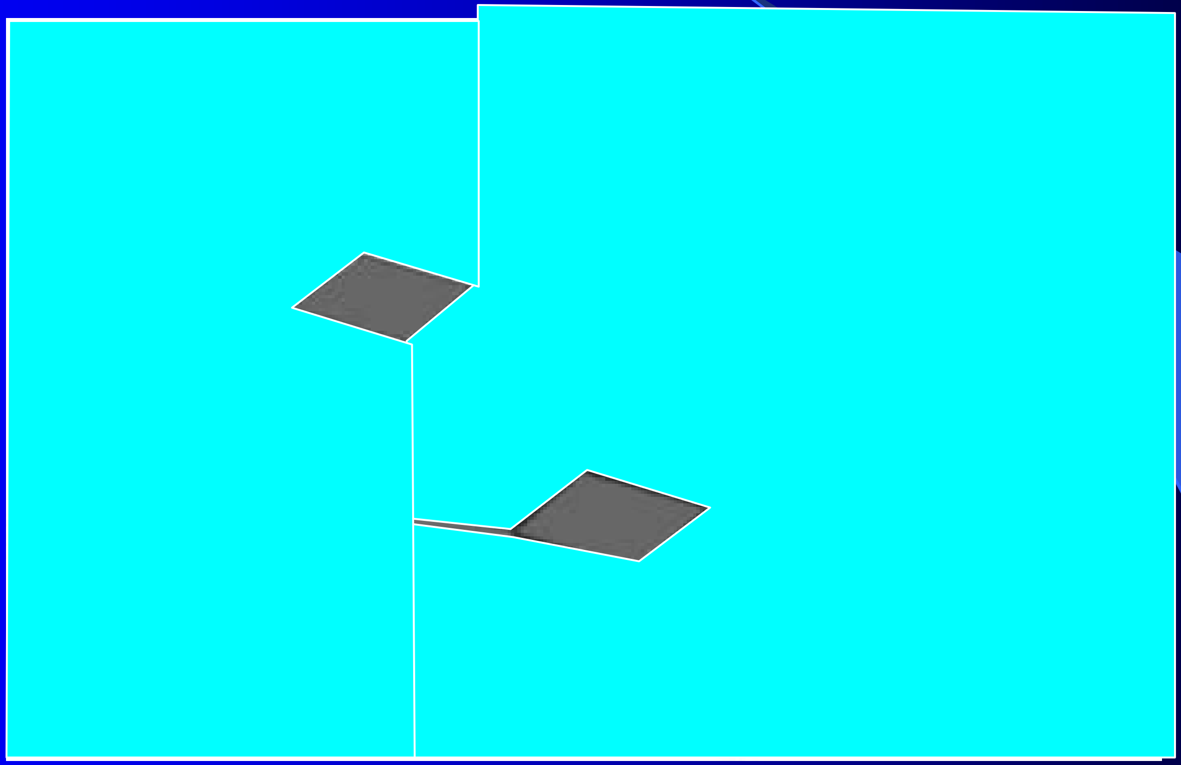




Are these lines horizontal?



Which square is brighter?

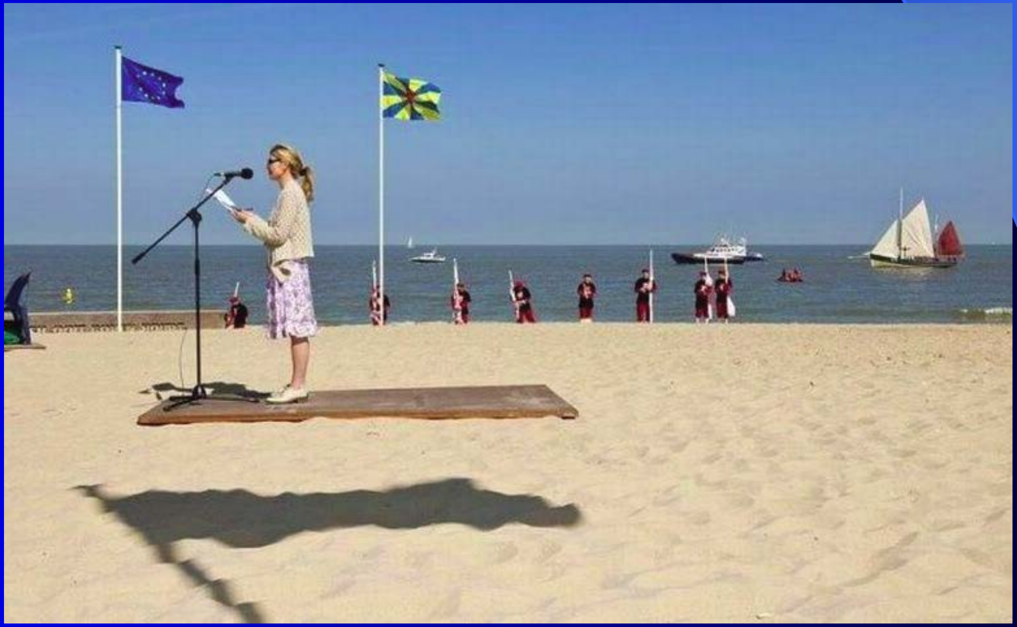


Visual Illusions

- The perception of brightness of a surface,
- or the length of a line *depends on context*.
- Not on basic measurements like:
- the no. of photons that reach the eye
- or the length of line in the image.
- Humans perceive images by making assumptions about the structure of the scene. Illusions arise when these are wrong.

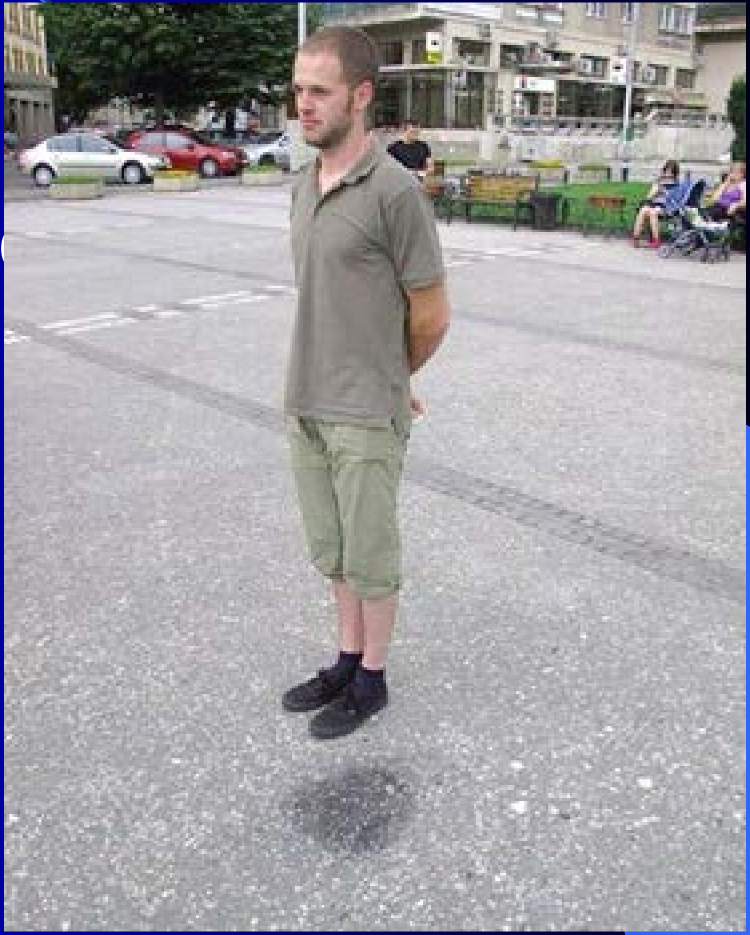
Flying Carpets?

- Can people fly? You think that the shadow is cast by the rug the woman is standing on.
- But instead it is cast by a flag outside the picture.



Levitation?

- Accidental alignment.
- The wet spot on the groun
- is miss-interpreted as
- a shadow.
- This shadow is usually
- at the contact point of
- the human and ground.



Theories of Vision.

- Feedforward Theories of the Ventral Stream (Hmax).
- Marr’s Theory of Vision
- Representations and Tasks/Bottom-Up and Top-Down
- .

Feedforward Models

- Fukushima 1980. Riesenhuber and Poggio 1999.
- Multiple stages of processing which build progressively more complex/abstract representations.
- This mimics knowledge of the Ventral Stream – V1,V2, V3 , V4, IT.
- Hierarchical organization – where V1,…V4 all have a complet map of viewed space.

Hmax: Poggio et al.

- Invariant representations of objects are built up rhough a hierarchical feedforward algorithm.
- Each stage is composed of separate populations of neurons that perform feature extraction and spatial pooling.

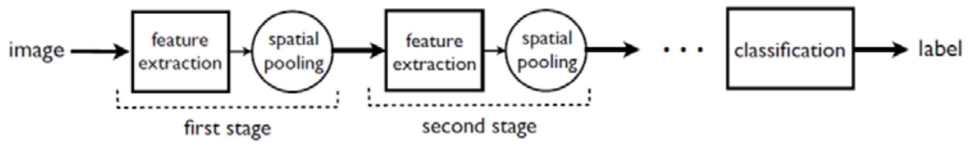


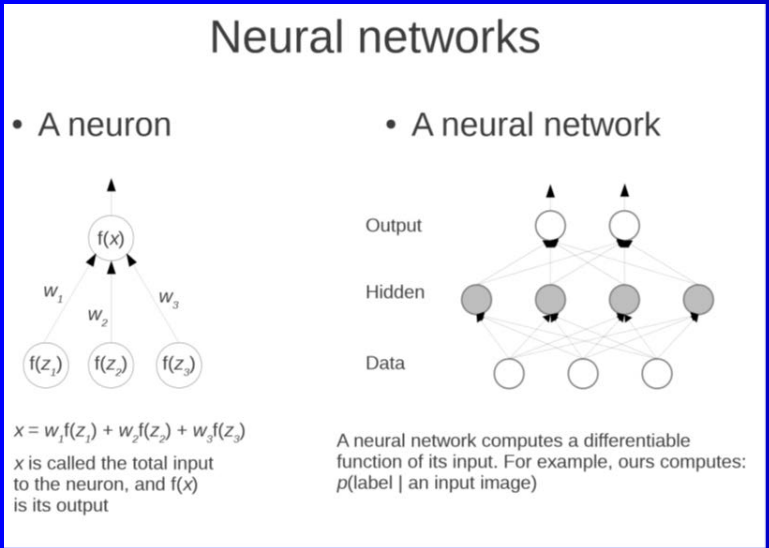
Figure 1: Hierarchical feedforward processing model of visual cortex. Each stage contains separate populations of neurons for feature extraction and pooling, resulting in object selective responses at the top stage that are invariant to variations in pose.

Invariance

- Basic Idea – each successive stage learns progressively more complex features of the input which are built on the features extracted in the previous stage.
- Pooling over spatial positions gives progressively more tolerance to variations in the position of features, culminating in object-selective responses which are invariant to variations in pose of an object.

Bio-Inspired Models

- Deep Belief Networks (Krihevsky et al. 2013)
- Image classification -- ImageNet



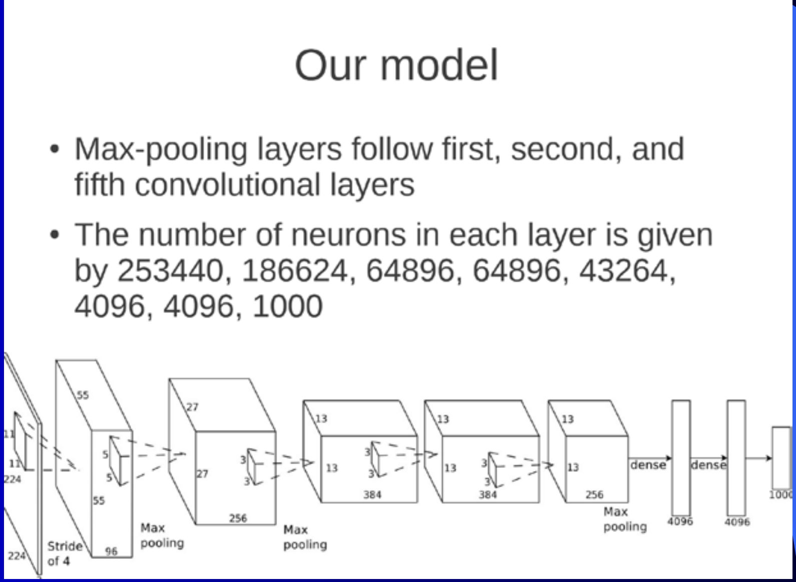
Neural networks

- A neuron
- A neural network

$x = w_1 f(z_1) + w_2 f(z_2) + w_3 f(z_3)$

x is called the total input to the neuron, and f(x) is its output

A neural network computes a differentiable function of its input. For example, ours computes: *p*(label | an input image)



Our model

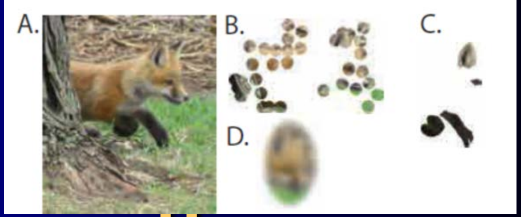
- Max-pooling layers follow first, second, and fifth convolutional layers
- The number of neurons in each layer is given by 253440, 186624, 64896, 64896, 43264, 4096, 4096, 1000

Visual Tasks

- These classes of models do capture some of the properties of the visual cortex.
- Deep belief networks are also very effective at some classification tasks (e.g., ImageNet).
- But the class of problems that these models solve – benchmark tasks – define the problem of perception too narrowly.

Tasks: (B. Olshausen)

. So, what are these tasks? What do animals use their senses for? Answering these questions is a research problem in its own right. One thing we can say with certainty is that visual systems did not start out processing HD resolution images, and auditory systems did not start out with well-formed cochleas providing time-frequency analysis of sound. Rather, sensory systems began with crude, coarse-grained sensors attached to organisms moving about in the world. Visual systems for example began with simple light detectors situated in the epithelium. Remarkably, over a relatively short period of time (estimated to be 500,000 years) they evolved into the wide variety of sophisticated eye designs we see today (Nilsson & Pelger 1994). What was the fitness function driving this process? Presumably it was the ability to plan useful actions and predict their outcomes in complex, 3D environments. For this purpose, performance at tasks such as navigation or judging scene layout is crucially important. From an evolutionary perspective, the problem of 'recognition' - especially when distilled down to one of classification - may not be as fundamental it seems introspectively to us humans.



Tasks: Scene Understanding

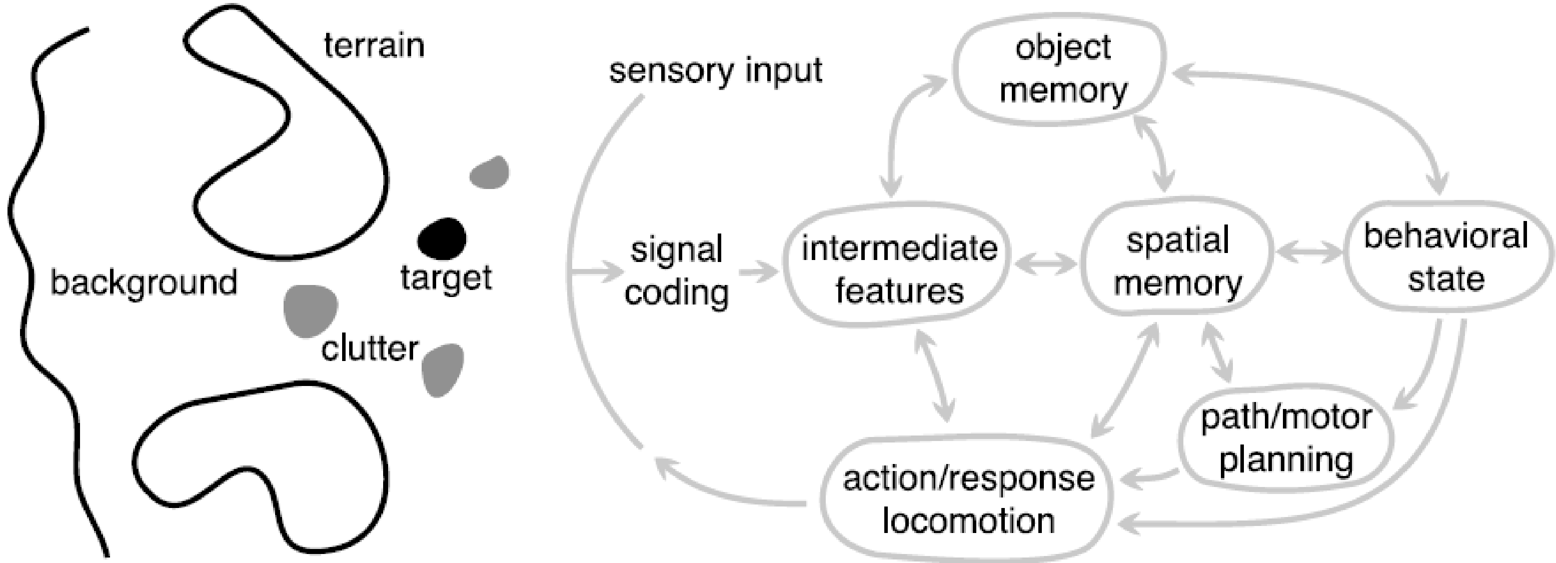


Figure 2: Components of scene analysis. The scene itself contains not just a single target object, but other objects, terrain, and background, all of which may be important for behavior. The neural structures enabling scene analysis contain multiple levels of representation and analysis. The level of "intermediate features" is where inferential processes come into play. (From Lewicki, Olshausen, Surlykke & Moss, 2013)

Marr’s Theory of Vision (1982)

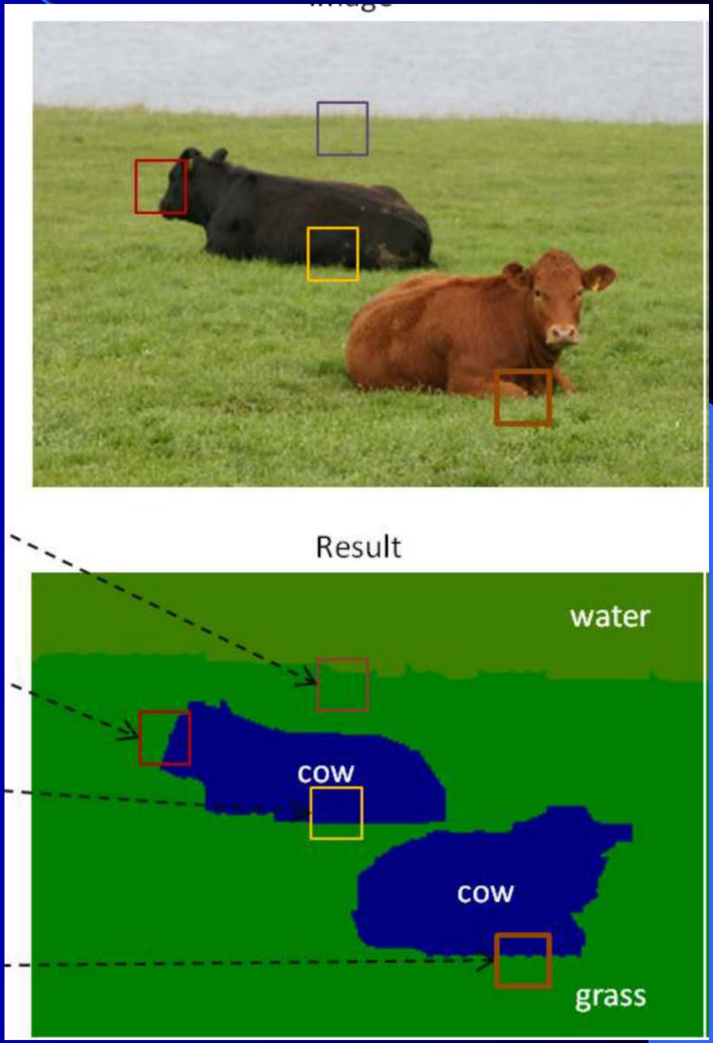
- Marr proposed that vision is broken down into constructing a sequence of representations.
- (I) A primal sketch – represents features and tokens from the image.
- (II) A 2.5-D sketch that makes explicit aspects of depth and surface structure.
- (III) An object-centered 3D model representation of objects.

Marr’s Theory and Low-, Mid-, and High-Level Vision.

- Vision can be broken down into low-, mid-, and high-level vision (very roughly).
- Low-level vision – local image operations which have limited knowledge of the world.
- Mid-level vision – non-local operations which know about surfaces and geometry.
- High-level vision – operations which know about objects and scene structures.

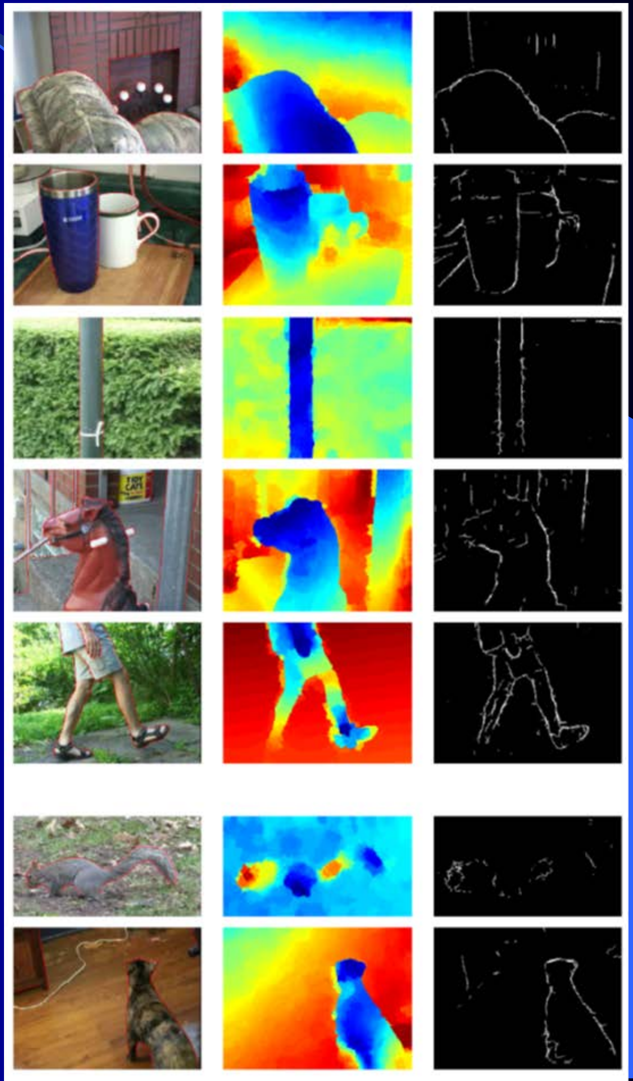
Low-level vision

- Image processing.
- Filtering, denoising,
- enhancement.
- Edge detection.
- Image segmentation.
- Right: ideal segmentation
- (followed by labeling).



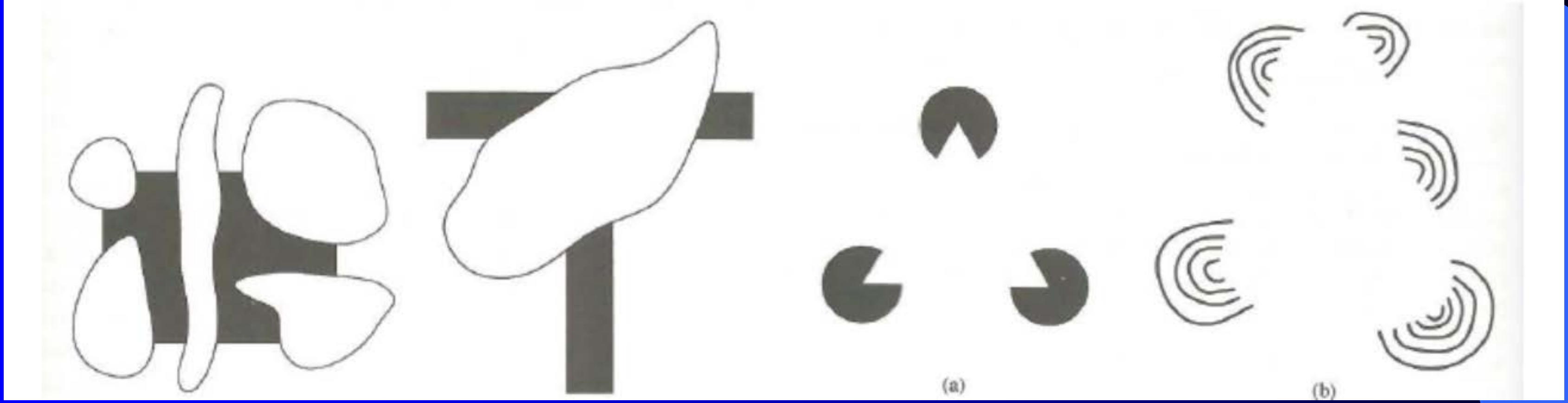
Mid-Level Vision: Depth

- Estimation of 3D surfaces:
- E.g., binocular stereo,
- structure from motion,
- Figure: Images, Depth,
- Segmentation.
- (Blue close, red far).



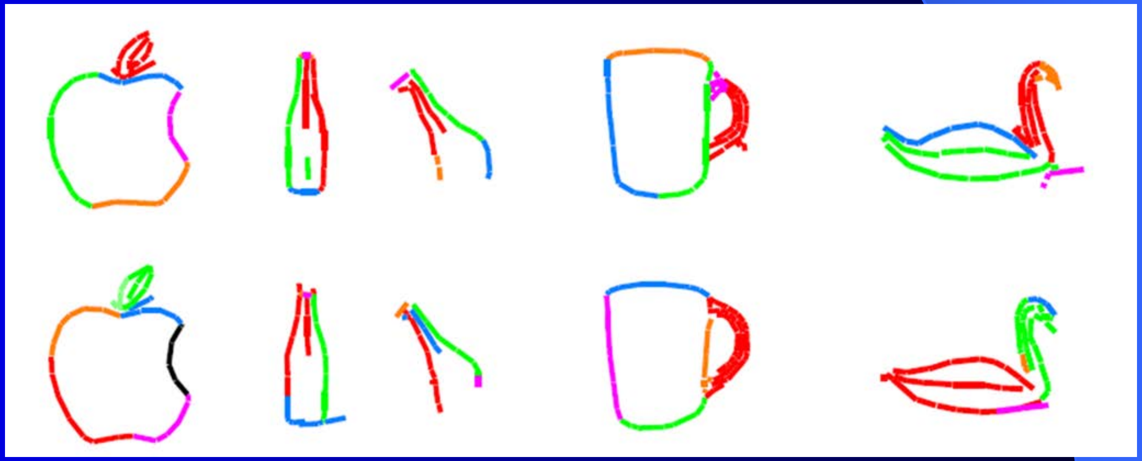
Mid-Level Vision: Grouping

- Kanisza. Humans have a strong tendency to group image structures as surfaces.



High-Level Vision

- Object detection and Scene Understanding.
- Example: detect objects and object parts.



Critique of Marr and Feedforward

- There seem to be limits to what feedforward methods can achieve (unless they propagate many hypotheses).
- Low-level vision is often very ambiguous, particularly to a population of neurons in V1.
- Context is needed – probably supplied by neurons higher up the hierarchy.

Local Ambiguity

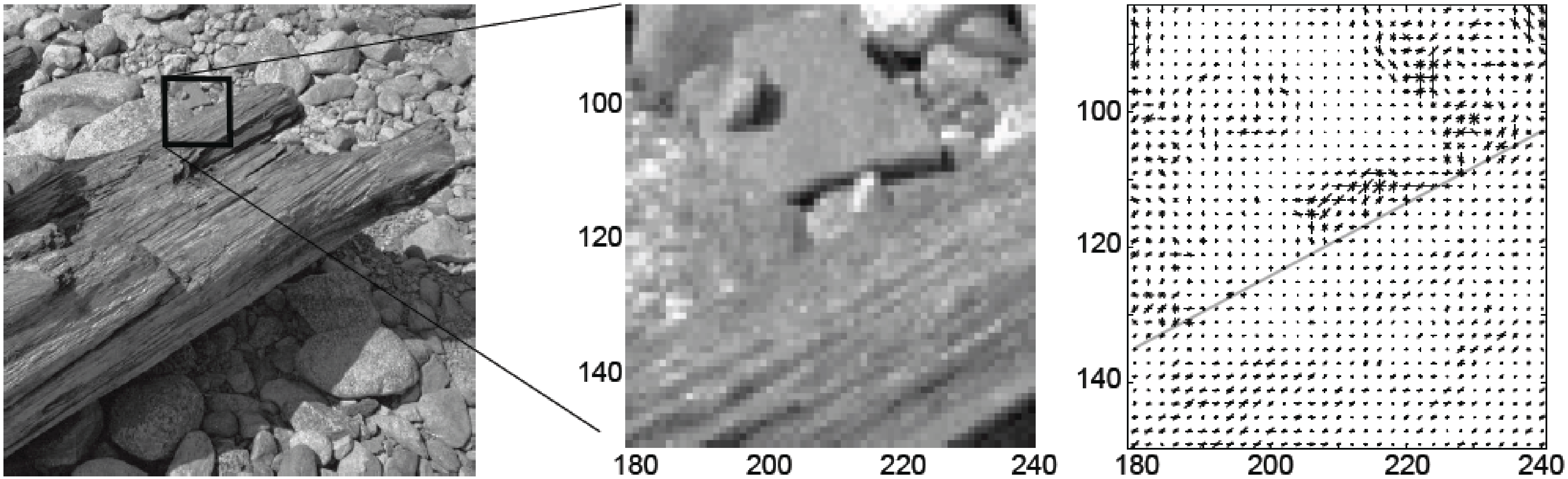


Figure 4. The outlined region around the boundary of the log (left panel) is shown expanded in the middle panel. The right panel shows how a hypothetical array of model V1 neurons (Gabor filters at four different orientations) would respond to the image subregion shown at left. The length of each line segment indicates the magnitude of response of a neuron whose receptive is situated at that position and orientation. An array of such neurons provides only weak or ambiguous cues about the presence of object boundaries in natural scenes.

Intrinsic Images

- How to interpret the pattern of light in an image in terms of surfaces and reflection?
- Humans have this ability – psychophysics.
- Recall the checker-board pattern, the cylinder, and the shadow (earlier slide).
- This requires complex processing to disentangle all the factors that have caused the images. Hard to do bottom-up.

Generative Models.

- Mumford (1991) argued that the – feedback and feedforward connections – suggests that the visual system uses generative models.
- Lee and Mumford (2003) proposed a more detailed hierarchical framework.

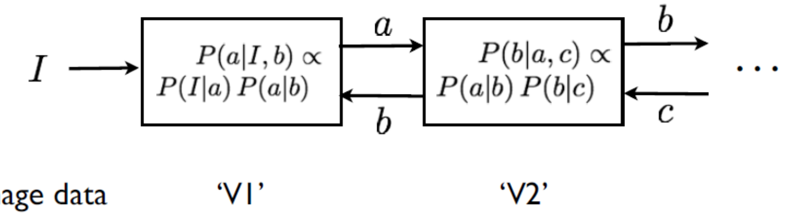
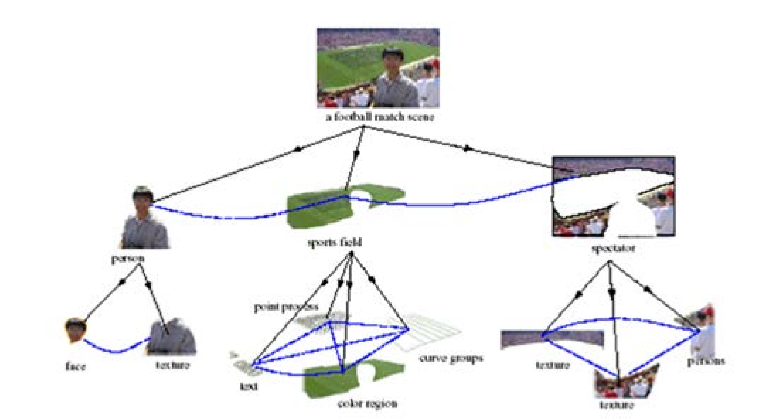

I
$P(a|I,b) \propto P(I|a)\,P(a|b)$
a
b
$P(b|a,c) \propto P(a|b)\,P(b|c)$
b
c
$\dots$
image data
'V1'
'V2'

Figure 7. Hierarchical Bayesian inference. The variables represented at each level are inferred from a combination of bottom-up and top-down inputs. Bottom-up inputs enter into the likelihood, while top-down inputs enter into the prior. The two are combined to form the un-normalized posterior, which guides the inference of variables at each level.

Realistic Generative Models.

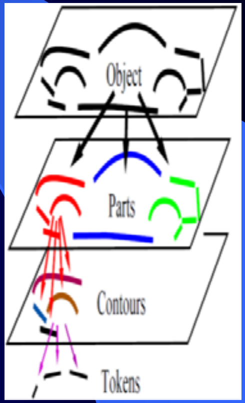
- Bottom-up proposals are validated by top-down generative models.

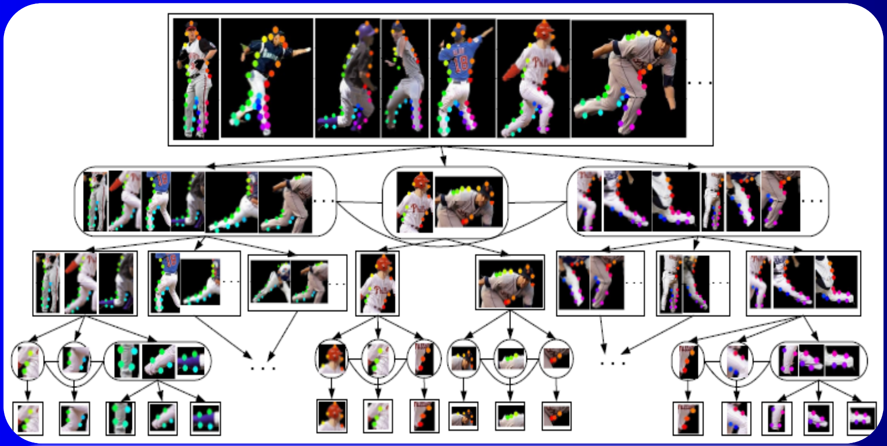


- From Z. Tu, X. Chen, A.Yuille, and SC Zhu.

Hierarchical Compositional Models.

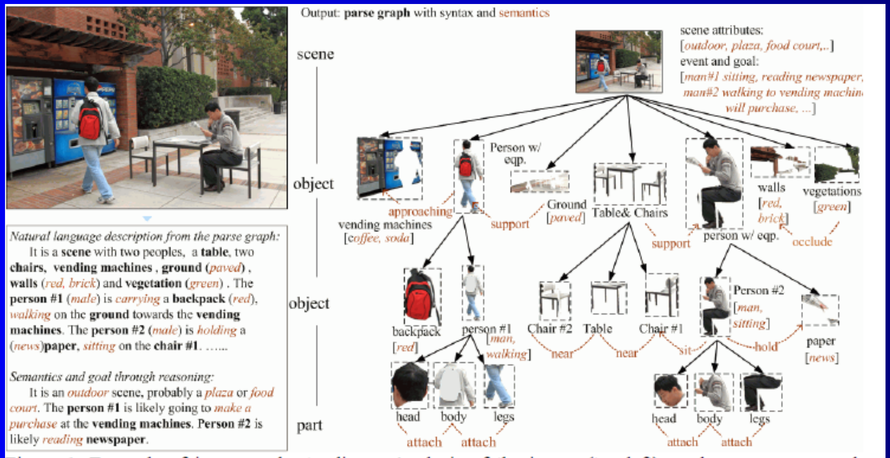
- Compositional models represent objects and scenes in terms of compositions of elementary shared parts.
- This offers a possible solution to the complexity problem of vision.





High-Level Vision

- Understanding of objects, scenes, and events.
- Reasoning about functions and roles of objects, goals and intentions of agents, predicting the outcomes of events.
- The full Artificial Intelligence problem.



Natural language description from the parse graph:

It is a **scene** with two peoples, a **table**, two **chairs**, **vending machines** , **ground** (*paved*) , **walls** (*red, brick*) and **vegetation** (*green*) . The **person #1** (*male*) is *carrying* a **backpack** (*red*), *walking* on the **ground** towards the **vending machines**. The **person #2** (*male*) is *holding* a (*news*)**paper**, *sitting* on the **chair #1**. …...

Semantics and goal through reasoning:

It is an *outdoor* scene, probably a *plaza* or *food court*. The **person #1** is likely going to *make a purchase* at the **vending machines**. **Person #2** is likely *reading* **newspaper**.

Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description in natural language (bottom-left).

Summary: (I)

- A lot is known about mammalian visual systems – but much more remains to be known.
- Current theories of the visual system capture only a very limited amout of human abilities.
- These abilities include the enormous variety of visual tasks that we can perform.

Summary II:

- Human ability to perform all these visual tasks suggests that we can construct rich representations of the input stimuli – from image properties, to geometry of objects and scene structure.
- It appears that these representations are hierarchical – from local and low-level (V1) to global and high-level (IT).

Representations and Bottom-Up and Top-Down Processing

- The enormous range of visual tasks that human perform suggests that the visual system is able to construct rich representations – of geometry and scene structure.
- It seems very hard to construct these representations with sophisticated

Conclusion III

- Constructing these representations is very difficult, because of the complexity and ambiguity of images.
- This suggests that there must be sophisticated feedforward and feedback processing between different representations.
- Increasing experiments on the brain suggest that feedback/top-down is important.

Conclusion IV

- Much knowledge of the visual system is at the level of details – these big-picture theories of vision are hard to test.
- We will start looking at the details in the next lecture -- and return to some of the big picture issues later in the course.
- Finally, realize the visual system is very complicated and only partly understood.

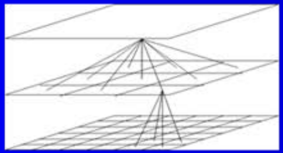
Some readings

- B. Olshausen. Perception as an Inference Problem.
- D.K. Kersten and A.L. Yuille. Inferential Models of the Visual Cortical Hierarchy.
- Both to appear in The Cognitive Neurosciences, 5th edition (MIT Press).
- Sensation and Perception. Eds. J.J. DiCarlo and R.I. Wilson.

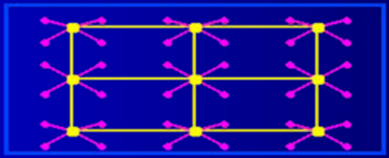
Models for Object Detection

- Deformable Part Models (DPMs).
- Objects are represented in terms of parts.
- DPMs are trained on benchmarked datsets.

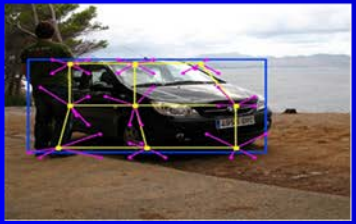
Parent-Child spatial constraints



Parts: blue (1), yellow (9), purple (36)



Deformations of Car



Deformations of Horse