

Probabilistic Models of the Visual Cortex Fall 2019

Alan Yuille

Part 1. Purpose of this course:

Towards a Unified Theory of Artificial and Natural Intelligence – for Vision

- The relationship between the study of AI and NI is fascinating and very timely. AI is booming, but many scientists think that taking AI to the level of human performance will require better understanding of NI
- Marr and Poggio conjectured (1978) that AI and NI could be studied jointly by distinguishing between three different levels of analysis (i) computational, (ii) algorithmic, (iii) hardware.
- AI and NI should be similar at the computational level, possibly at the algorithmic level, and not at the hardware level.
- The goals of unifying AI and NI include both scientific understanding of intelligence and the development of more advanced AI systems that can augment human intelligence, cooperate with humans, and in some situations replace humans.

Part 2. What is Vision?

- To extract information from the environment in order to take action. More specifically, to estimate the physical properties of the 3D world from light rays that reach our eyes (or cameras).
- These physical properties vary from coarse interpretation of an image (horse in a field) , to more detailed (hair on horse, is it sweaty, is the horse young or old, sick or healthy, what is it doing).
- Images are formed by light rays, geometry of objects, material properties of images – computer graphics.
- Vision can be subdivided – for ease of study – into many different tasks (object recognition, object detection, depth estimation), but these sub-divisions are “fictions”, and all the tasks need to be done together.
- Vision is really the full AI problem (Poggio). It starts with processing images but also involves language, reasoning, analogy, action, and almost all aspects of intelligence.

Part 2: What is Vision?

The more you look the more you see.

- Humans can extract a lot of information from a single image.
- “There is a fox in the garden” (coarse).
- “There is a young fox emerging from behind the base of a tree not far from the view point, it is heading right, stepping through short grass, and moving quickly. Its body fur is fluffy, reddish-brown, light in color, but with some variation. It has darker colored front legs and a dark patch above the mouth. Most of the body hairs flow from front to back.” (detailed).



Part 2. What is Vision?

The Full AI problem

- Understanding of objects, scenes, and events. Describing them in language.
- Reasoning about functions and roles of objects, goals and intentions of agents, predicting the outcomes of events

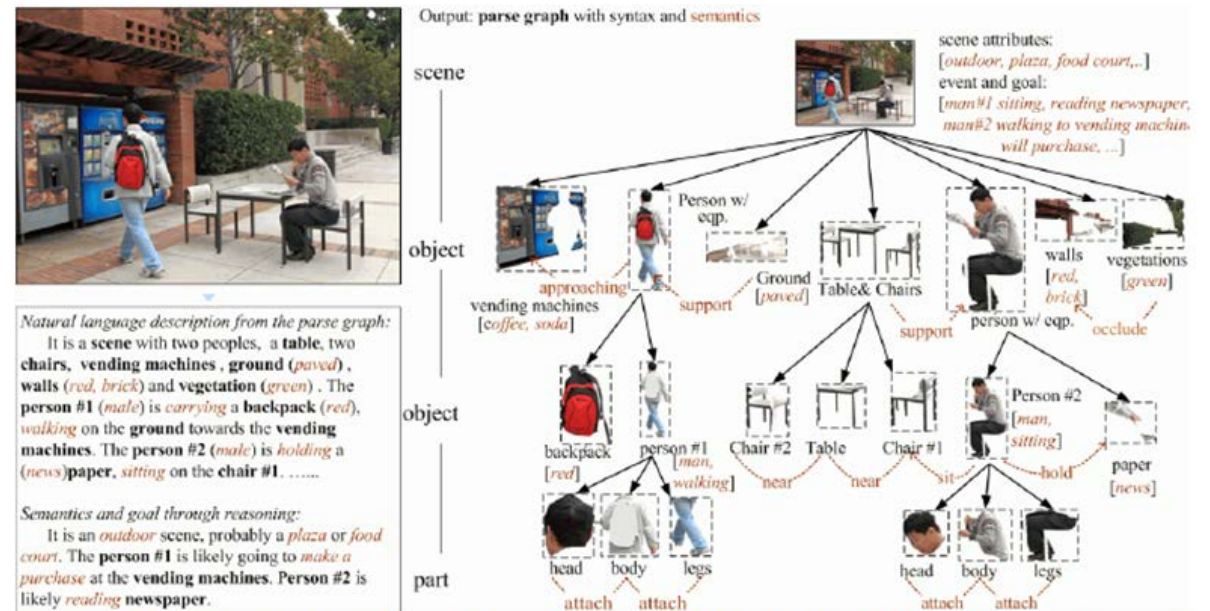


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph may be converted to a description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

Part 3. Why do Humans have Vision?

- Vision enables us to get information about the world from distances ranges from fractions of an inch to millions of light years. We use this information to take actions.
- It is our underappreciated superpower. (most animals have limited vision and rely on other senses, like smell).
- It enables us to get information at long distance – e.g., to see what is happening one hundred miles away and to walk/run/drive/fly there.
- It enables us to get precision information at short distances – e.g., to thread a needle, to detect if a plant is edible, to track animals using subtle cues (e.g., trodden grass under shaded trees).

Part 3. Why do Humans have Vision?

- Arguably “Vision” combined with our abilities to perform actions (walk/run long distances, use tools with our opposable thumbs, etc.) are the reasons why humans are the dominant animal on this planet.
- It allowed our ancestors to travel long distance to better environments, it enabled them to track and kill animals, it allowed them to grow and harvest food, it allowed them to live in caves and build habitats. (No other animals has our abilities to travel long distances quickly and to perform high precision tasks, but none of these would be possible without vision).
- More speculatively, the need to do vision and action required our ancestors to evolve sophisticated neural hardware, which could then be extended to enable us to perform advanced cognitive tasks (e.g., mathematics). Intelligence may result (evolve from) our ability to do vision and action.
- AI pioneers thought that Vision was easy and “peripheral”, but found it was much harder than “intelligent tasks” such as playing Chess.

Part 4. Why is Vision Hard? Complexity.

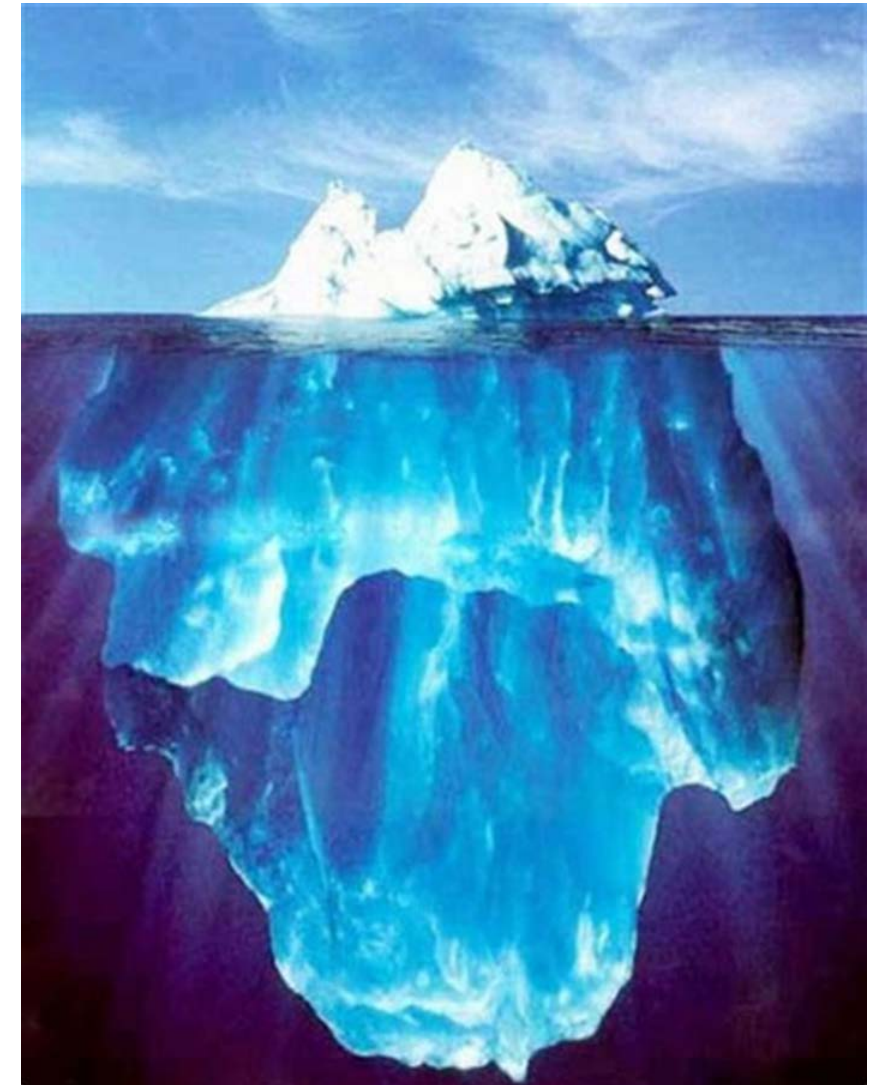
- Vision is extremely hard due to complexity and ambiguity.
- Complexity arises in several forms. The complexity of all images. The set of images is infinite. If we restrict each pixel value to take 256 possible values (as in a digital camera), then there are more 10×10 images than have been seen by all mankind over all history and pre-history. Humans see 10^9 each year.
- Complexity due to physical viewing conditions. For a single object – there are 13 viewing factors – and if we allow 1,000 values for each dimension, then we reach 10^{39} images for a single object!
- Complexity of scene compositions. A scene can be composed in a combinatorial number of ways – placing N possible objects into M possible positions – yielding M^N possible ways to build a scene (this ignores lighting, texture patterns, etc). This gets even worse if you consider changes in material patterns, lighting, viewpoint, occlusion.
- The complexity increases further for image sequences.

Part 4. Why is Vision Hard? Complexity.

- The set of images in any dataset are only an infinitesimal fraction of all images. Tip of the Iceberg.
- Image of a single object is a function of 13 parameters – camera pose (4), Lighting (4), material (1), scene (3).



Suppose we simply sample 10^9 possibilities of each parameter listed...



Part 4. Why is Vision Hard? Complexity.

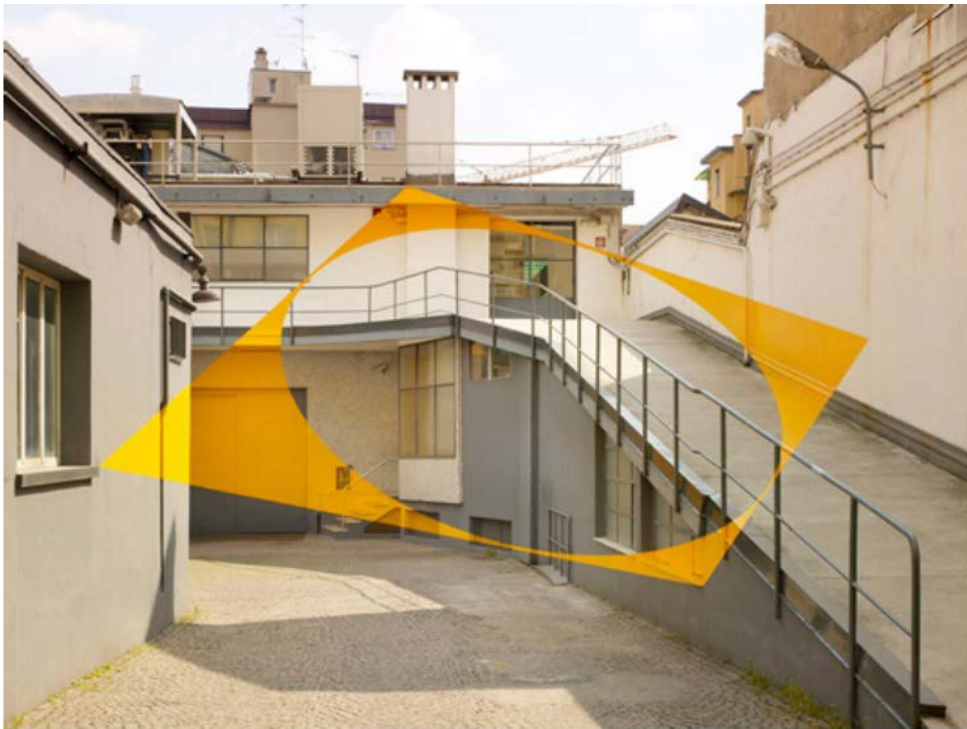
- This combinatorial complexity puts a challenge on machine learning methods (like deep networks).
- Machine learning assumes that we have training and testing datasets which are big enough to be representative of the underlying problem domain. Otherwise the methods will be biased to the datasets and will perform badly on rare events (those underrepresented in the datasets).
- But if the problem domain is combinatorially complex – then it is impossible to have training and testing datasets which are big enough.
- This gives new challenges -- How to train models, if your datasets are too small to be unrepresentative of the real world? How to test models and guarantee performance if you can only test on a tiny fraction of possible images?
- But the Human Visual system knows how to do this!

Part 4. Why is Vision Hard? Ambiguity.

- There are several types of ambiguity.
- Ambiguity in how images are generated from the 3D world:
Images are functions of the geometry and material properties of the objects (and the lighting). This can be ambiguous. Sometimes we can confuse material properties for geometry. And geometry for material properties (Demo later).
- Ambiguity without context – images are often locally ambiguous and need context to disambiguate them.

Part 4. Why is Vision Hard?

- Ambiguity – geometry, material properties, lighting. C. von der Malsburg.



Part 4. Why is Vision Hard

- Ambiguity – Toyota Video – C. von der Malsburg.



Part 4. Why is Vision Hard. The Local Ambiguity of Images

Airplane
Car
Boat
Sign
Building



Part 5. How can Humans do Vision?

- Humans and Primate devote an enormous amount of neural resources to do vision. Roughly 40-60% of the cortex (the seat of high level intelligence) is involved in vision.
- The human/primate visual system is extremely complex.
- The number of neurons is enormous (the number of trees in the Amazon rain forest) and the number of connections between neurons is even bigger (the number of leaves in the Amazon rainforest).
- Neurons are also very complex. They have many different types (but the majority are pyramidal cells) and they have a large diversity of morphology (shapes). They have complex dendritic structures, possible internal states (in the cell body), and maybe have mechanisms to change synaptic strength. They could be an order of magnitude (or more) complicated than artificial neurons.
- Similarly, neural circuits may be much more complex than artificial neural circuits.
- The human brain is by far the most complicated physical system that we know about.

Part 6. What do we know about Human/Primate Vision?

- We know an enormous amount. Many clever people have worked very hard and written an enormous number of papers and some very good books.
- But we do not know nearly enough!
- We cannot characterize human vision behaviorally – i.e. we do not know what visual tasks humans can perform in complex images (but we know a lot about what it can do in simple images, see discussion later). This makes it hard to design “visual Turing tests” (to test whether a computer can perform as well as a human).
- We do not know how the neural circuits in the human/primate visual system perform vision (although we have some findings).

Part 7. How can Vision be Studied?

- We can study vision as follows.
- Behavior – show humans/primate visual stimuli and find out what they perceive (verbal reporting, or clicking on computer keys).
- Non-invasively – we can use fMRI/EEG/MEG to record brain activity and see which brain areas are engaged in visual tasks and where representations are located.
- Neural Anatomy – we can study the structure of the human/primate visual system, decompose it into areas, obtain structure diagrams.
- Electrophysiology/Optical methods – we can measure the activities of single neurons (and even groups of neurons). We can even manipulate neurons to silence them, or cause them to be active (fire).

Part 7. How can Vision be Studied?

- Computational Modeling – we can develop models that predict human behavior (perception), neural activity (at the cell level, and at coarser regional levels).
- Reverse Engineering – we can develop AI Vision algorithms that can perform the types of visual tasks that humans/primates can do.
- The boundary between computational modeling and reverse-engineering is not easy to define. One distinction is that computational modeling starts by trying to explain specific experiments. While reverse-engineering starts by trying to solve a vision problem and then may relate it to experiments.
- Some models are bio-inspired (but do not attempt to relate their predictions directly to biology).
- Arguably all AI Vision algorithms can be thought of as reverse-engineering.

Part 7. How can vision be studied.

- Some types of vision research – arguably – can be described as “cottage industries”. They concentrate on trying to explain components of the visual system without bothering how these components could fit together into a complete system (that could perform all the tasks that human vision can do).
- Analogy – Newton thought of himself as picking up pebbles on a beach and noticing that some have nice properties, while the enormous ocean of truth stretched out before him.
- Cottage industry research is picking up pebbles.
- AI Vision – and reverse-engineering – is aimed at the “ocean of truth”.
- Real and Toy stimuli – a fault line between the studies of biological vision and AI vision, is that AI Vision researchers have to work on real world stimuli, while much biological vision research has been done on simplified stimuli.
- This is a problem because AI researchers have found that models that work on simplified stimuli often do not work on real stimuli without significant changes. When evaluating a claim in biological vision – it is very important to understand what data is what tested on. (Ignore most of the paper –read the “methods” section!).

Part 8. The Difficulties of Studying Human/Primate Vision

- The problem is complexity-squared !! There is the complexity of the images and scenes. There is also the complexity of the human/primary visual system.
- Simplifications must be made – to the visual tasks and to the types of stimuli.
- Simple stimuli have good properties and are controllable (needed for good experimental design). Unfortunately findings on simple stimuli often do not generalize to more realistic stimuli – but they can offer insight and inspiration as demonstrations.
- This differs from sciences like physics – where the fundamental laws and forces (e.g., Newton's laws of motion, gravity, electromagnetism, relativity, quantum mechanics) can be discovered and tested on simple stimuli and then extend naturally to much more complex situations.

Part 9. But Progress is Being Made

- Progress is being made and the future is promising (but not easy).
- Reverse-Engineering (AI-Vision) is making big progress (due to machine learning and big data), but the current versions are not nearly enough to model human vision.
- Computer graphics can synthesize very realistic images – showing that we understand the forward process (that we have to invert). C-G is increasingly being used to help reverse engineering and to obtain stimuli for behavioral, neural, and non-invasive experiments.
- There are many new experimental neuroscience techniques (multi-electrodes, optical methods, ability to switch neurons on and off).
- Crowd sourcing methods means that behavioral experiments can be run at enormous scales.
- But can we solve this in five years? The first man to claim he could develop a machine model for intelligence in five years was Leibnitz (350 years ago).

10. What are the key properties of human vision – those that distinguish it from AI Vision?

- Some aspects of Human visions worth copying for AI, but some are not.
- Marr & Poggio's three levels of analysis. Consider birds and airplanes. Wings are necessary for birds and airplanes. But airplanes do not need feathers.
- Brains versus Machines.
- The Brain uses 1 watt for computation (20 more watts to stay alive) while a computer with 4GPUS uses 1,500 watts.
- Real neurons get tired and need food (blood), artificial neurons do not.
- The brain is a product of evolution (a sequence of kludges?) but AI Vision is designed by humans (doesn't mean it is optimal).
- The brain has adapted to perform visual tasks in specific environments – we are good at reading facial expressions, but not at recognizing 100,000,000 faces (viewed front-on), or interpreting Computer Tomography images.

Part 11. Human Visual Failures.

- Lack of Attention– the Gorilla in the Room, Change Blindness, inability to realize an image is an impossible scene, the Gorilla in the CT image. (sensible “short-cut” strategies to avoid computations).
- Accidental Alignments – sensible assumption that works most of the time.
- After-effects – demo -- these may be due to “tired neurons” or might be a sensible strategy (e.g., for a system that has to repeatedly self-calibrate itself).
- Visual crowding in the periphery – know what the objects are, but get their order wrong (“short-cut” strategy so that only fovea is high resolution).
- Memory/resource limitations – inability to track more than five objects, to remember details of pictures.
- Other failure types – e.g., seeing motion in static images,

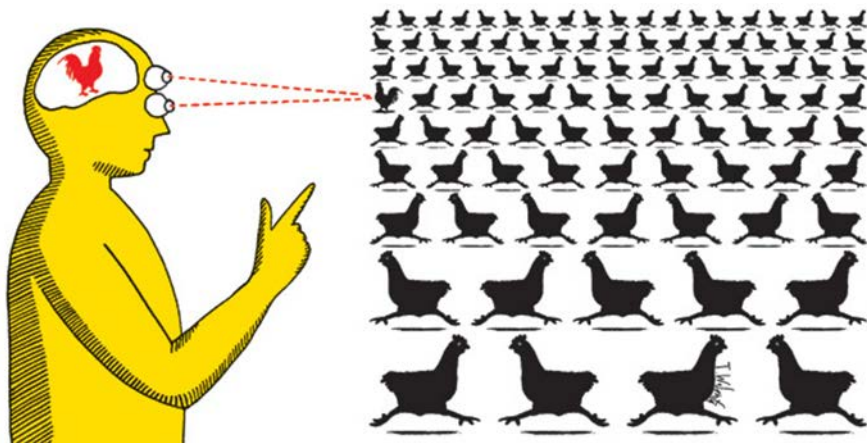
Part 11 Human Vision Failures

- Gorilla in the Room. We fail to see gorilla's if our visual attention is directed elsewhere. This may be due to a visual strategy that is very efficient (requires few computational resources) and correct most of the time. (Skilled illusionists perform tricks by diverting attention).



Part 11 Change Blindness

- We are bad at noticing differences between images (provided they have similar semantic content). We are bad at noticing changes outside our center of gaze.



Change Blindness (using flicker)
(from J. Kevin O'Regan -- <http://nivea.psych.univ-paris5.fr>)

Part 11 Human Vision Failure

- Accidental Alignment. This may result from a sensible visual strategy that is correct most of the time. (Far right – eye in the kitchen sink).



Part 11. Human Visual Failures

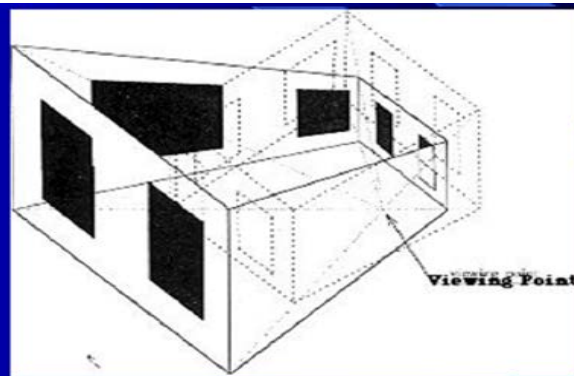
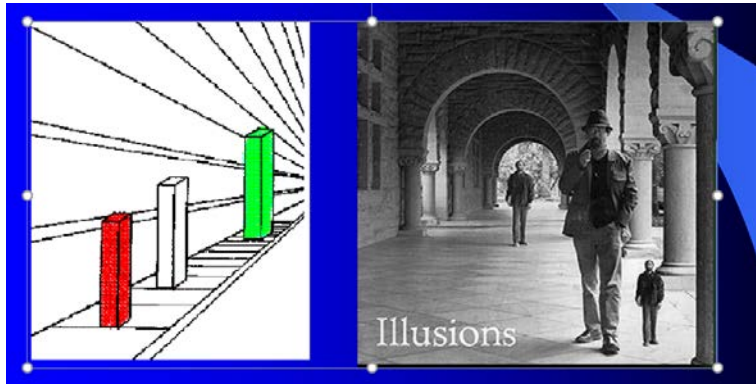
- Michael Bach's webpage gives a huge set of visual illusions with discussions and possible explanations. <http://www.michaelbach.de/ot/>
- Many of these can be explained as sensible visual strategies – e.g., in Bayesian terms (see Pavan later) -- which sometimes make mistakes.
- Consciousness, Visual Awareness, and Blindsight.
- Sometimes we see things but do not realize it – perhaps they are perceived but not stored in memory. E.g, we say we cannot see something but if asked to guess what it is, we will guess right most of the time. This relate to the problem of consciousness – fascinating scientifically, but unclear that it is relevant for AI-Vision.

Part 12: Key Aspects of Human Vision

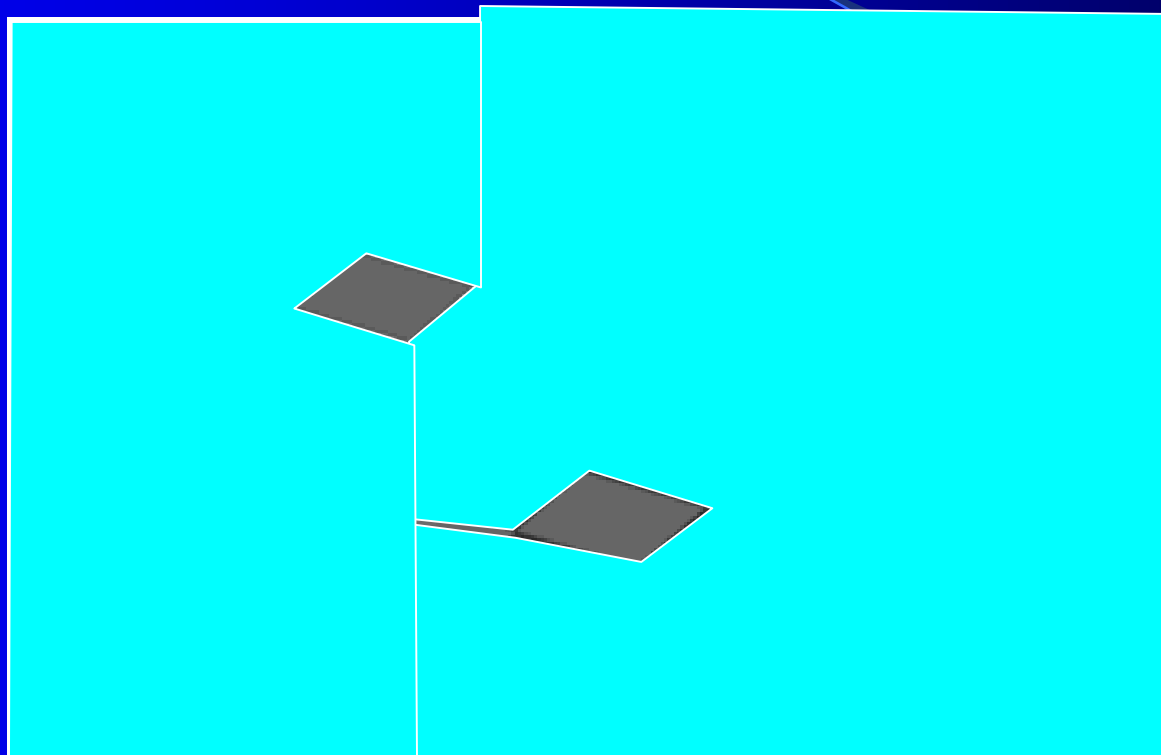
- We now discuss key aspects of human vision, which distinguish it from AI-Vision.
- Analysis by Synthesis: Humans have the ability to perform approximate and as-needed inverse computer graphics. This involves the ability to “parse the image” and explain every pixel (if required).
- This is only approximate – e.g., we may only estimate relative depth order instead of absolute depth. It is “as needed”, so it may be sufficient to construct the 3D world coarsely and ignore unimportant details (“the more you look the more you see”).
- This is a modern update to the classic ideas of “Vision as Inference” (Helmholz, Gregory). But inverse inference is hard. “Vision is not just a passive acceptance of stimuli, but an active process involving memory and other internal processes”.

Part 12: Key Aspects of Human Vision

- We see in 3D assuming normal 3D structure (and are fooled sometimes).



Which square is brighter?

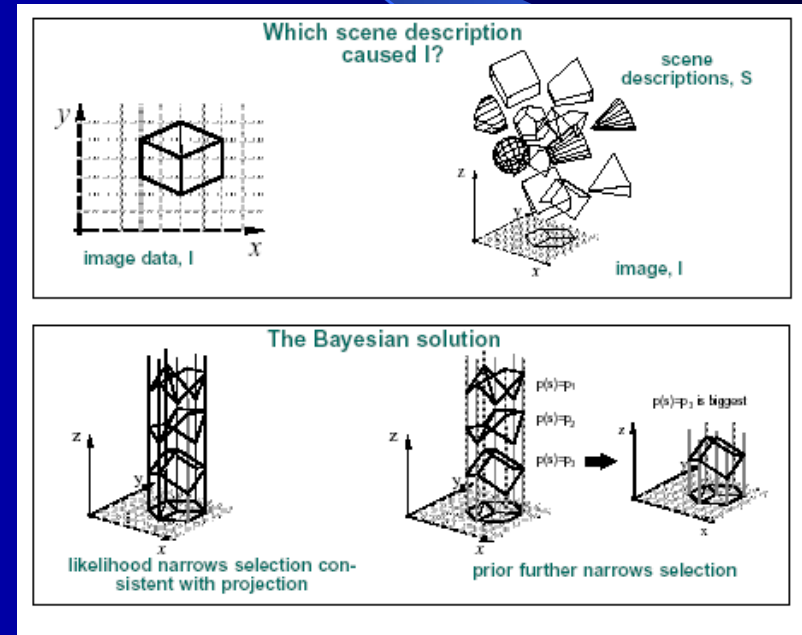


Part 12. Key Aspects of Vision

- Knowledge of the 3D world.
- The ability to do inverse-computer-graphics requires knowledge of the external 3D world.
- Pavan Sinha's examples of the Necker Cube – ambiguity is removed by using prior knowledge of what objects are most likely to be present in the world (and/or accidental viewpoint assumptions).
- Gibson's ecological constraints, Marr's natural constraints.
- Naïve physics (Tenenbaum).

Key Aspects of Vision: Bayesian Perspective

- There are an infinite number of ways that images can be formed.
- Why do we see a cube?
- The likelihood $P(I|S)$ rules out some interpretations S
- Prior $P(S)$ —cubes are more likely than other shapes consistent with the image.



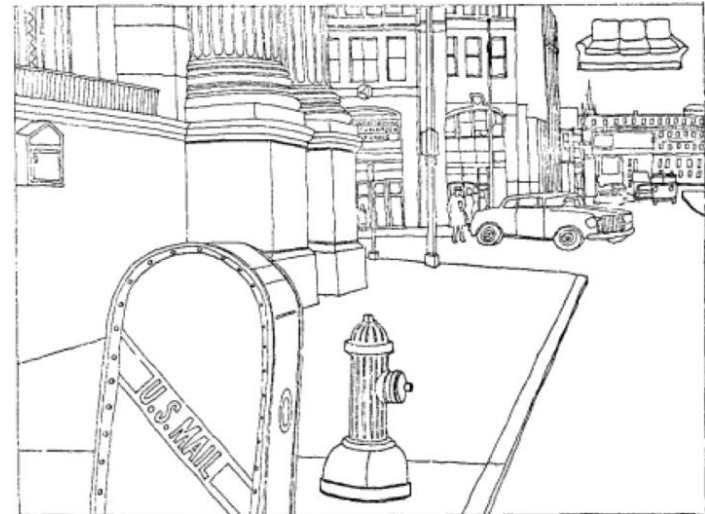
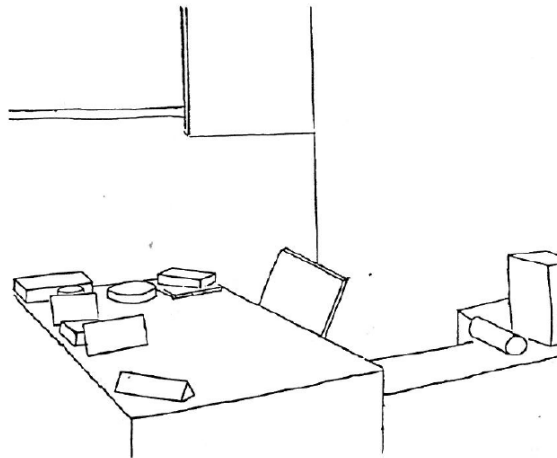
Part 12. Key Aspects of Vision

- How humans acquire knowledge: Development and Learning.
- Human vision brings to bear an enormous amount of knowledge acquired/learnt over a lifetime (unlike current AI-Vision systems).
- This knowledge is initially learnt, during development, by an orchestrated procedure where certain visual abilities are learnt first to enable the learning of more complex ones.
- This learning relies (at least initially) on exploiting image sequences, searching for causal structure, taking actions in the world, and exploiting other senses. “Learn like a child”.

Part 12: Key Aspect of Human Vision

- Humans have the ability to use Context. This results from our knowledge about the world. C, von der Malsburg.

Object recognition:
50% by context



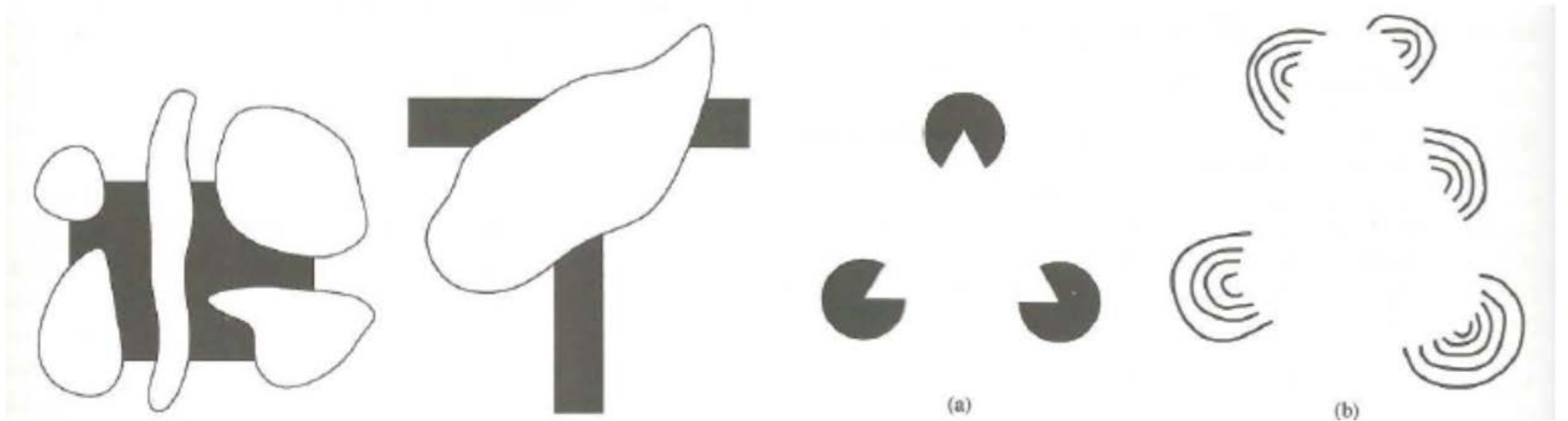
- Object recognition: 50% by context. Search guided by context.

Part 12. Key Aspects of Vision

- Perceptual Organization.
- Humans have also the ability to see patterns and to group basic elements into more complex structures. This was studied by Gestalt psychologists (e.g., Wertheimer, Kanisza).
- This can be illustrated by various grouping properties – accidental alignment, common fate, etc. (the phenomena are so strong – everybody gets the same perception) that demonstrations are sufficient.
- The ability to group patterns, of highly variable components, shows that human vision can deal with abstraction.
- Certain types of grouping (e.g., Kanisza) shows that human vision is aware of geometry and occlusion (independent of object knowledge)

Part 12. Key Aspects of Vision

- Kanisza



Part 12: Key Aspects of Human Vision

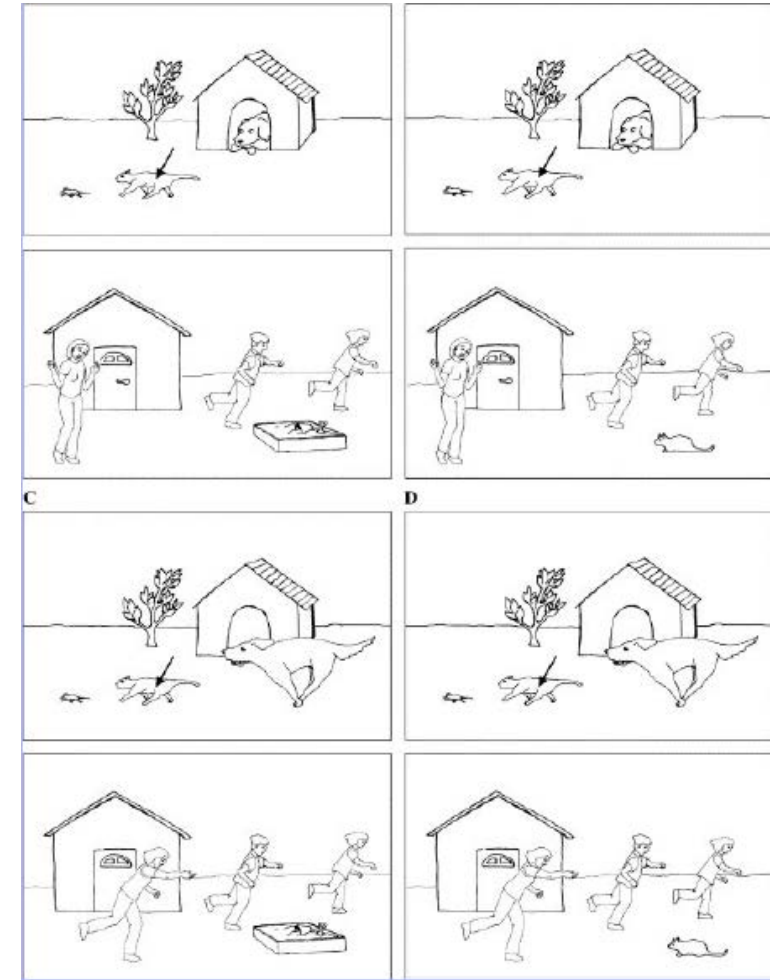
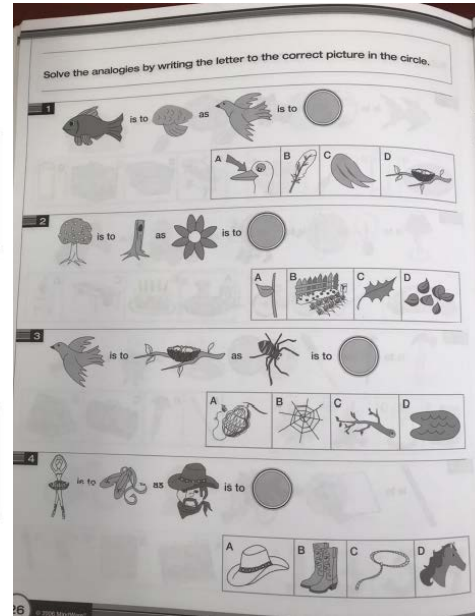
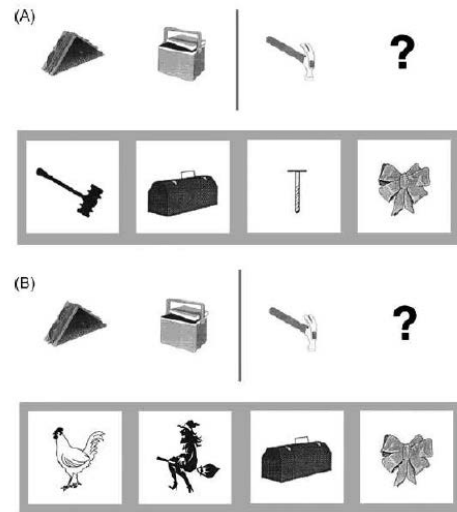
- Cues and Modularity.
- The study of human vision has identified visual cues which are sufficient for performing visual tasks in restricted (toy environments).
- E.g., Shape from shading, texture, contour, focus, and perspective.
- These cues are effective in simplified domains (toy worlds) although extending them to work in the complexity of real images is often extremely challenging.
- Many of these cues are now embedded in AI-Vision models, but many are not.

Part 12. Key Aspects of Vision

- Abstraction and Domain Transfer.
- Humans can understand an object from an image, from a drawing, from an highly abstract sketch. This is, in AI-vision terminology, an extreme form of domain transfer.
- Humans can factor shape and geometry – and recognize a blue tree, even if they have never seen one before.
- Humans can perform analogical reasoning. We can not only recognize visual similarity between objects, but also relationships (e.g., part-whole: paw to cat, hand to person), and functional relations (e.g., hammer is in toolbox, notebook is in backpack), and other relations (e.g., woman chases child is like cat chases mouse – but only in some ways!).

Part 12. Key Aspects of Vision

- Domain Transfer. Analogies.

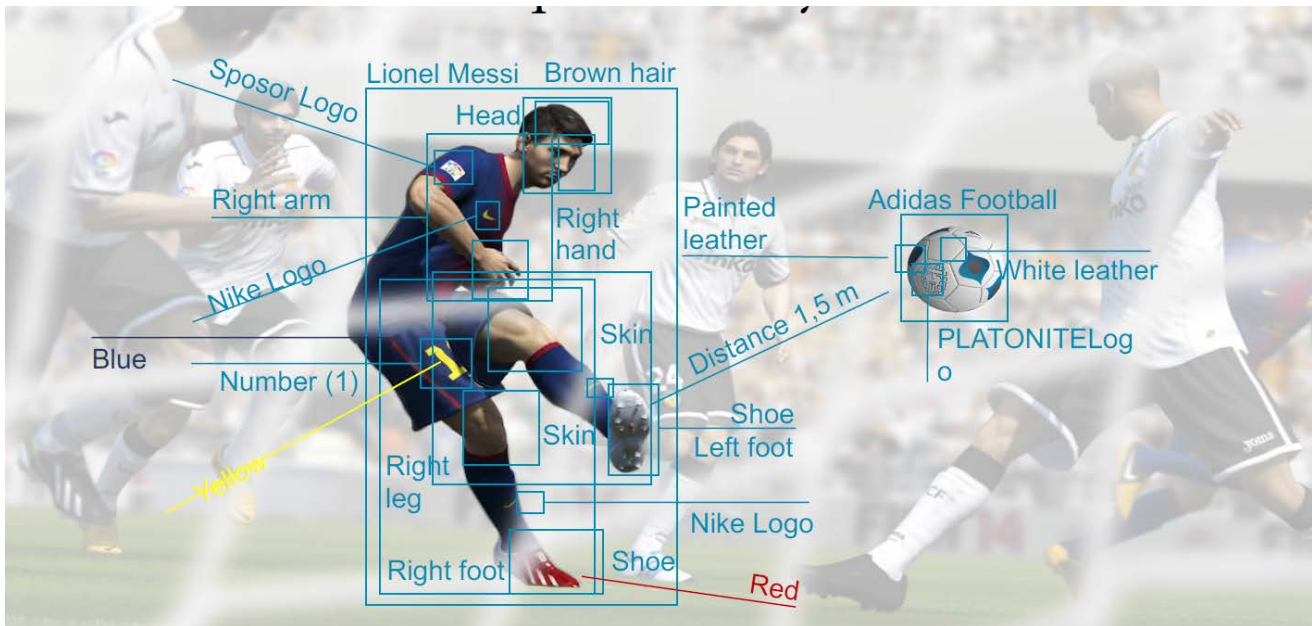


Part 12: Key Aspects of Vision

- Internal representations of parts. This helps make vision explainable.
- These parts can be detected without context. They can also be described by language. (Interestingly, humans may be fascinated by some types of visual stimuli – like fires and the flow of water – because we cannot describe them easily in words).
- Humans can explain why they have recognized an object – this is a car because I can see the wheels, the chassis, the doors, etc – and they satisfy the correct spatial relationships.
- These parts can also be abstract – i.e. we can recognize a fish even if it is constructed from bicycle parts.

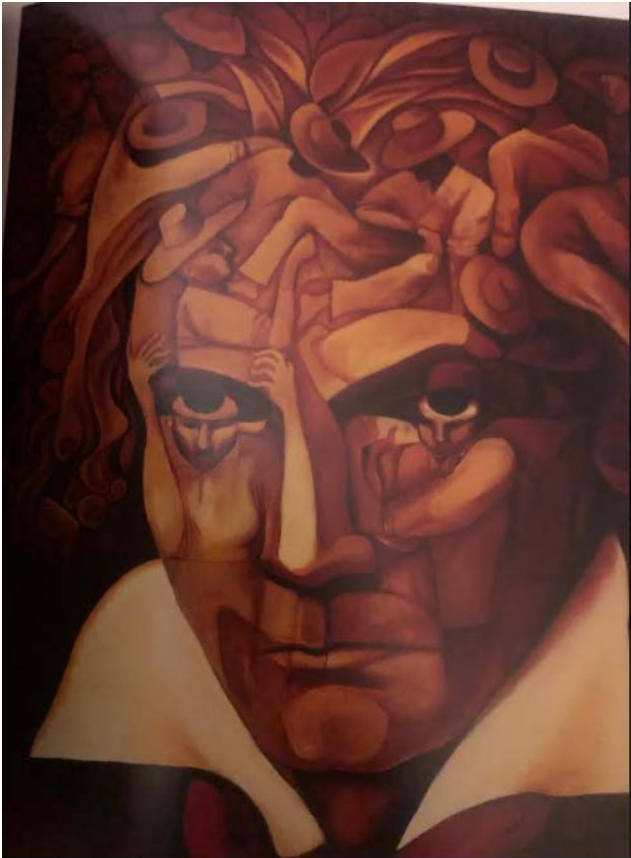
Part 12: Key Aspects of Human Vision

- Part Examples (from von der Malsburg)



Part 12: Key Aspects of Vision

- Parts. And Abstractions. Beethoven constructed from Humans.



Part 12: Key Aspects of Human Vision

- Humans have the ability to deal with the complexity of vision and to answer many different tasks using the same underlying representation (do we really know this?).
- Compositionality is an strategy for doing this. Objects are composed hierarchically in terms of basic elements. This gives the ability to construct an enormous number of objects from basic elements.
- Toy-model analysis (Y&M).

Part 13: Key Aspects of Human Visual Architecture

- The input to the visual system starts at the retina which captures the image and transmits it to the visual cortex (but the retina may also process it).
- The visual cortex is organized into visual areas. These start at visual areas V1 and V2, which are huge and contain roughly 70-80% of the neurons in the visual cortex. The ventral stream, where most object recognition is done, is organized hierarchically.
- The ventral stream (and the whole visual cortex) contains both feedforward neural pathways (up the hierarchy) and even more feedback neural pathways (down the hierarchy).

Part 13. Key Aspects of Human Visual Architecture.

- The hierarchy of the visual system has motivated several computational theories. We discuss some of the most influential.
- Marr's Theory. The visual system proceeds by constructing three representations: (I) The primal sketch, which captures the properties of the image, (II) The 2 1/2D sketch which captures the geometry of the scene, and (III) The 3D representation, which describes objects in terms of their 3D structure.
- Informally, it captures the notion that low-level neurons "know" about image properties, mid-level neurons "know" about geometry, and high-level neurons "know" about objects. Hence knowledge is distributed hierarchically – low- and mid-level vision have "generic" knowledge about the world, but do not have knowledge of specific objects.
- Marr's theory is feedforward – meaning that the representations are constructed bottom-up without top-down feedback.

Part 13: Key Aspects of Human Visual Architecture

- Fukushima and Neo-Fukushima.
- This is also hierarchical but does not involve 3D geometry. The processing is done feed-forward.
- It consists of a hierarchy with simple features at the lowest level and increasing more complex features as we ascend the hierarchy.
- An important property is that the features become increasingly specific as we ascend the hierarchy, but increasingly independent of position.
- Hmax (Riesenhuber and Poggio) is an influential theory of this class which has been related to properties of the ventral stream.

Part 13: Key Aspects of Human Visual Architecture

- Other theories include analysis-by-synthesis which combines bottom-up and top-down processing (Mumford 1991). Some instantiations of this are the Helmholtz machine and DDMCMC. It is also clear that top-down processing is important for video sequences when we want to predict motion (Rao and Ballard).
- Another metaphor is that the vision system is organized like a hierarchical organization like an Army (or Corporation) where knowledge is distributed hierarchically. The General has an executive summary and can contact the Colonels for more details, who can contact the Major and so on down to the Privates. Here information flows up and down the hierarchy.
- More generally, most (according to Ed) neuroscientists think there must be both feedforward and feedback processing.

Part 13: Key Aspects of the Human Visual Architecture

- A computational perspective is that the architecture is developed to solve the problem – object specific knowledge is the most powerful but is harder to apply because it requires recognizing the object. Instead it may be better to first perform “generic” processing (which exploits properties common to all objects) in order to obtain representations, like the 2 1/2D sketch, which can then be matched to objects.
- A related perspective, is the visual cortex must address the problem of representing an enormous number of objects, learning new objects from few examples, and rapidly processing images to determine which object is present. Toy-world models (Y&M) give one example.

Part 14. This course.

- The course aims at describing visual phenomena and the types of computational models that are used to describe them.
- This involves techniques like linear/non-linear filtering, Bayesian decision theory, markov random fields, neural network models, geometry, and radiosity.
- The hope is to introduce students to this fascinating area on the boundary between Artificial and Natural Intelligence.
- And to motivate researchers to take AI-Vision to the next level by developing models that can match human visual abilities.