Vision as Bayesian Inference Spring 2021

Alan Yuille

The study of vision.

- In 2011 Vision was a very complicated and confused research field. A student asked me “what papers should I read in computer vision? There are so many and they are so different?”
- Vision researchers tried to make a list of twenty techniques all computer vision researchers should know. They gave up after the list rose to four hundred.
- In 2013 Vision researchers “discovered” deep neural networks. This led to huge progress in all visual tasks when evaluated on annotated datasets.
- Does this mean “vision is solved”? Can vision be reduced to a supervised machine learning problem?

Vision as supervised Machine Learning

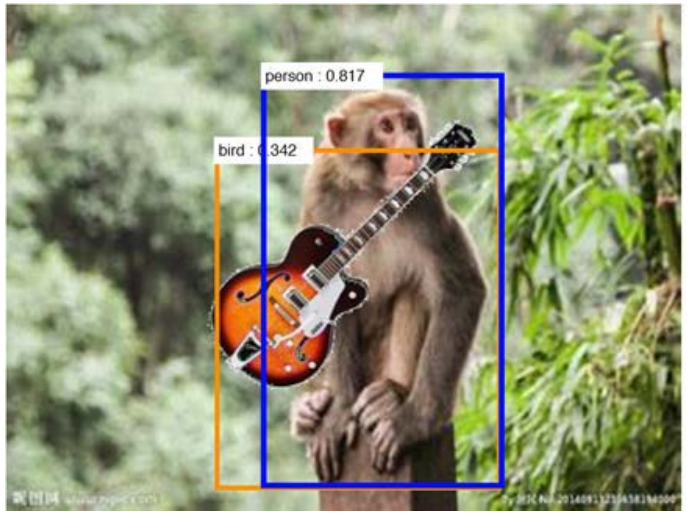
- Three basic ingredients.
- (1) An annotated dataset which consists of training, testing, and validation set. Each with input and desired output.
- (2) A set of machine learning algorithms, e.g., deep networks.
- (3) A large set of computers with GPUs to train and test the deep networks.
- Train the algorithm on the training set, fine-tune on the validation set, and test on the testing set.
- Do good results on the testing set mean that the problem is “solved”?

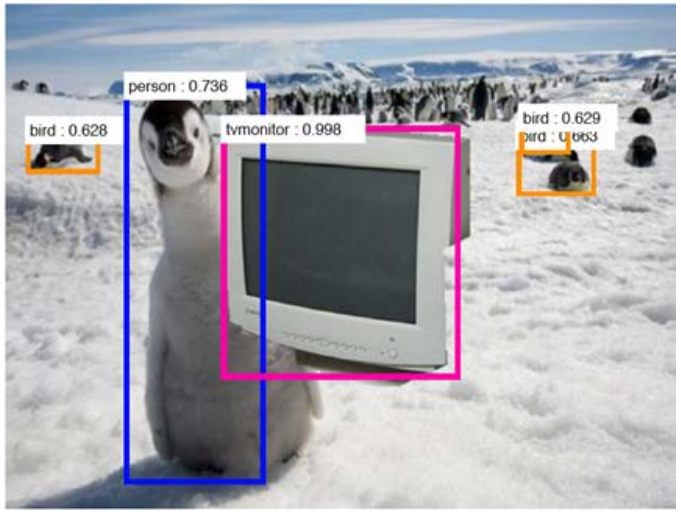
My answer is no!

- “The proof that vision is solved in that the ImageNet object detection challenge has been retired. Deep networks can recognize objects better than humans” (U.C. Berkeley Professor. Shanghai. 2018).
- Not so fast:
- (1) Human make mistakes on ImageNet due to linguistic problems and lack of expertise for some object categories.
- (2) More seriously, it is possible to make small modifications to images in ImageNet and reduce performance of Deep Networks to almost zero.
- Basic Reason: the set of images is infinite, so performance of algorithms on finite sized datasets does not guarantee good performance for all images.
- Moreover, annotation is only possible for a limited set of vision tasks.

Deep Networks are sensitive to occluders and context..

- Giving the penguin a TV turns the penguin into a human.
- Giving the monkey a guitar turns the monkey into a human and the guitar into a bird.





- See also “The elephant in the room” A. Rosenfeld et al. Arvix. 2018)

Deep Networks have made great progress

- There has been great process in computer vision due to machine learning methods, like deep networks, used with annotated datasets.
- They are sufficient to solve a restricted class of vision problems. Computer vision is becoming a billion, or trillion, dollar industry.
- But the really interesting vision problems remain unsolved, but not unsolvable. So it is an exciting time.

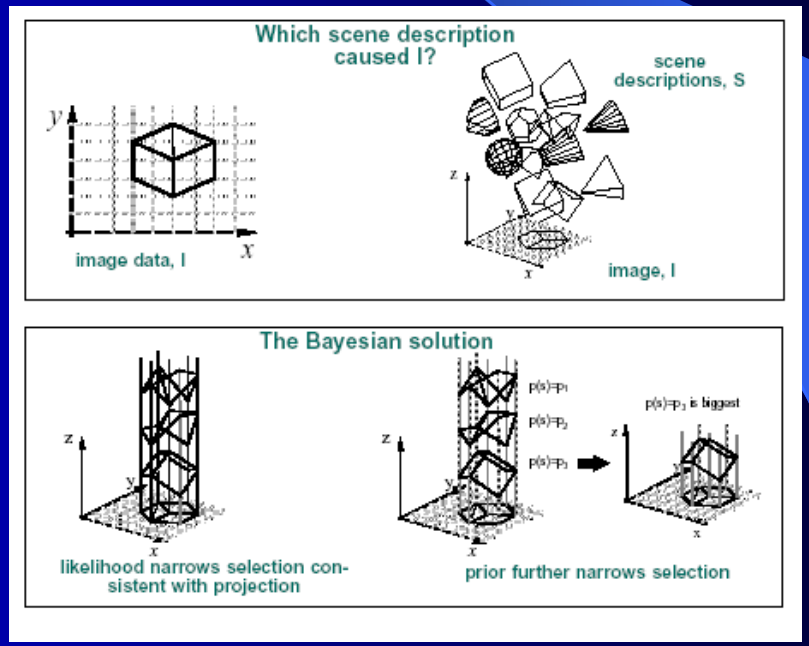
This course: vision as Bayesian inference

- This title is partly misleading.
- My belief is that vision will only be solved when it is formulated as Bayesian inference.
- Analysis by synthesis – vision as inverse computer graphics.

- But a complete theory of Bayesian inference is lacking. So the course is incomplete. So the course might be better described as 20 (or more) techniques that every computer researcher should know. (Hence the course will touch on a very large range of mathematical and computational methods. But one course is not enough to teach them all).

Vision as Bayesian Inference: Analysis by Synthesis

- 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 interpetations S
- Prior P(S)– cubes are more likely than other shapes consistent with the image.



Computer Vision and Human Vision

- The goal of computer vision is to emulate and then surpass human vision (human vision is limited by physiological and cognitive resources).
- To develop algorithms that can pass a visual Turing test. Algorithms that can look at images and get the same information that a normally sighted human observer can.
- Computer vision is nowhere near this ability.

Part 1. 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. It starts with processing images but also involves language, reasoning, analogy, action, and almost all aspects of intelligence. “Born to see”. “Vision is human’s underappreciated superpower”.

Part 1: 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 1. 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

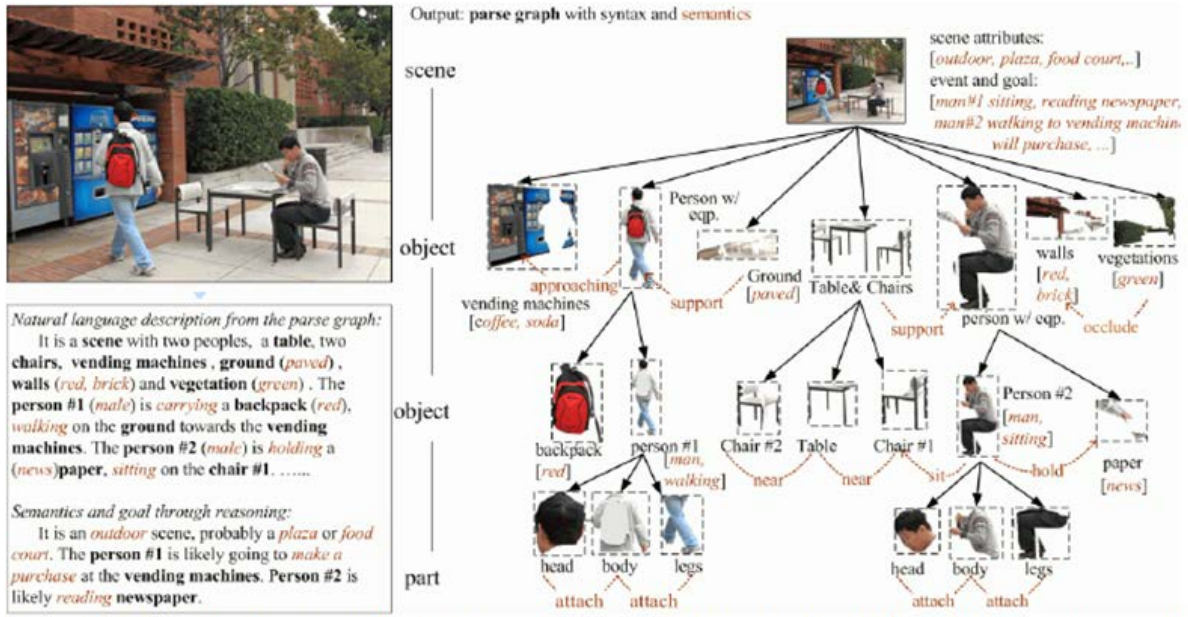


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description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

Part 2. Why is Vision Hard? Complexity.

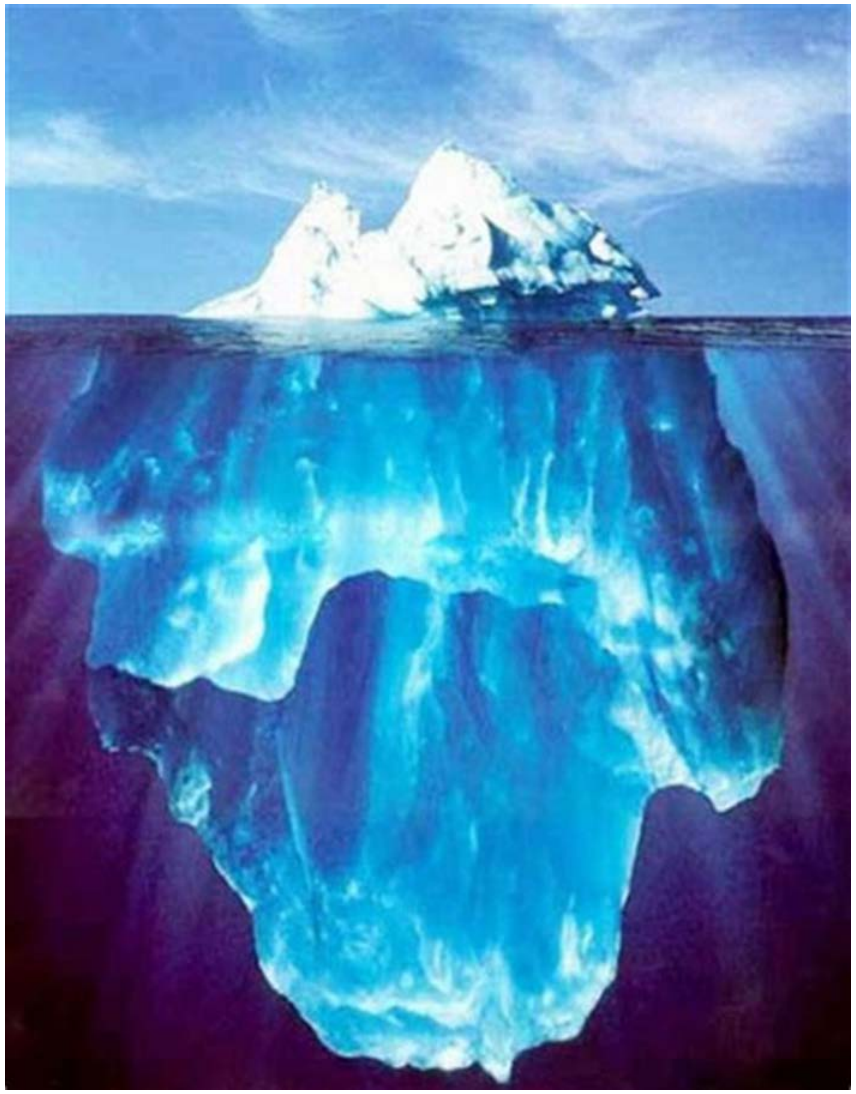
- Vision is extremely hard due to complexity and ambiguity.
- Complexity arises in several forms. The complexity of all images. The set of images is infinite. If we restrict each pixel value to take 256 possible values (as in a digital camera), then there are more 10x10 images than have been seen by 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 2. 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 10^3 possibilities of each parameter listed...



Part 2. 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?
- The Human Visual system knows how to do this.

Part 2. 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.

- Ambiguity without context – images are often locally ambiguous and need context to disambiguate them.

Part 2. Why is Vision Hard? Ambiguity.

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

Felice Varini

Part 2. Why is Vision Hard? Ambiguity.

- Toyota Video – C. von der Malsburg.

Part 2. Why is Vision Hard.
The Local Ambiguity of Images

Airplane
Car
Boat
Sign
Building

3. What are the key properties of human vision – 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.

3. Types of Human Visual Failures.

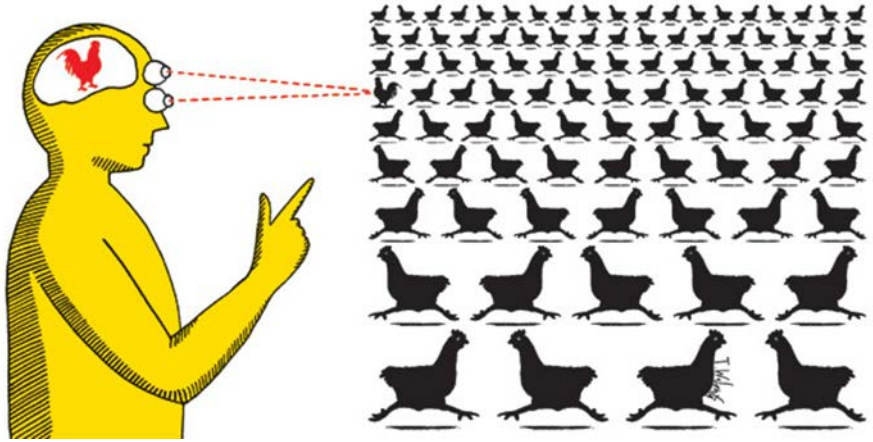
- *Lack of Attention*– the Gorilla in the Room, Change Blindness, inability to realize an image is an impossible scene, the Gorilla in the CT image. (sensible "short-cut" strategies to avoid computations).
- *Accidental Alignments* – but these are arguably sensible assumptions that works most of the time (and most would disappear if the viewer moves).
- *Lack of global consistency* – Escher staircase.
- *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, after-effects, crowding in periphery.

3. Human Vision Failures: Attention

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

3. Human Vision Failures: 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.



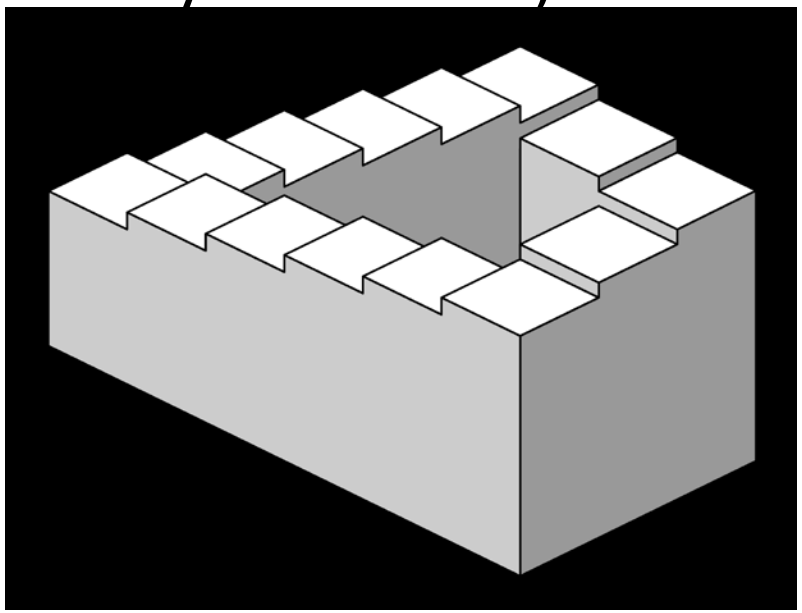
Change Blindness (using flicker)
(from J. Kevin O'Regan -- http://nivea.psycho.univ-paris5.fr)

3. Human Vision Failures: Accidental Alignment

- Accidental Alignment. These may result from a sensible visual strategy that is correct most of the time and most would disappear if the observer could move. (Far right – eye in the kitchen sink). These give some evidence for inverse computer graphics (see later).

3. Human Vision Failures: Consistency

- Failures of consistency. Without careful attention we may fail to notice the inconsistency. Perhaps this a sensible efficient (i.e. lazy) strategy (the 3D world is mostly consistent).



4. Strengths of Human Vision: Knowledge of 3D World

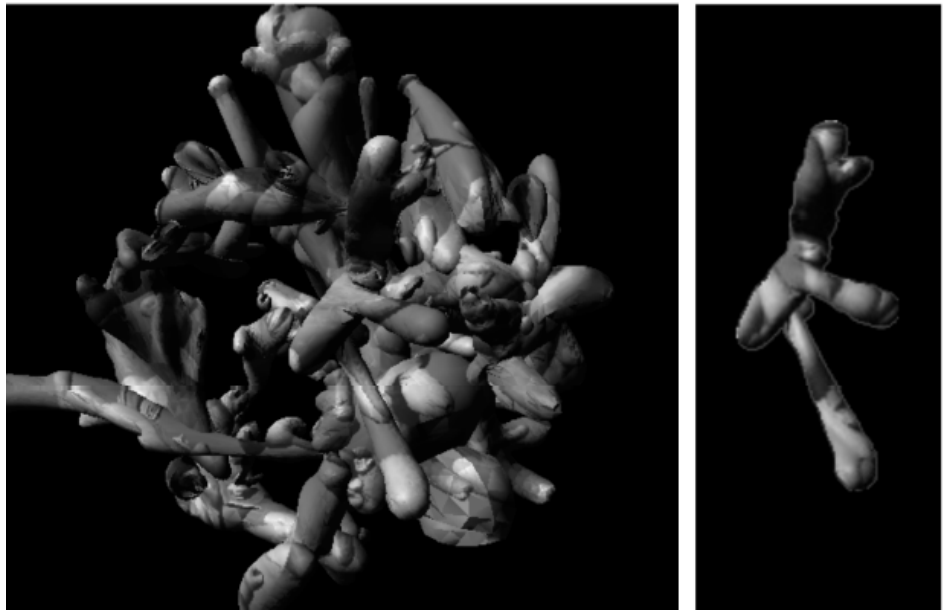
- Knowledge of the 3D world.
- The ability to do inverse-computer-graphics requires knowledge of the external 3D world.
- The Necker Cube.– ambiguity is removed by using prior knowledge of what objects are most likely to be present in the world (cube are more likely than the other objects which are consistent with the image, and/or accidental viewpoint assumptions – the cube would look similar to the image if the view changed slightly, not so for other objects.).
- Gibson's ecological constraints, Marr's natural constraints. Properties of typical visual environment and the 3D physical world.
- Naïve physics (Tenenbaum). (But often not real physics! The fundamental laws of science are rarely intuitive and it tools scientists many years to discover them). Also maybe only for tasks we are familiar with (e.g., catching a ball, balancing a stack of books).

4. Strengths of Human Vision: Knowledge of the 3D World

- How humans acquire this 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 leant, during development (Spelke, Kellman), 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. Theory of mind (Gopnik)
- This accumulated knowledge allows humans to learn from remarkably few images.

4. Strengths of Human Vision: Knowledge of the 3D world.

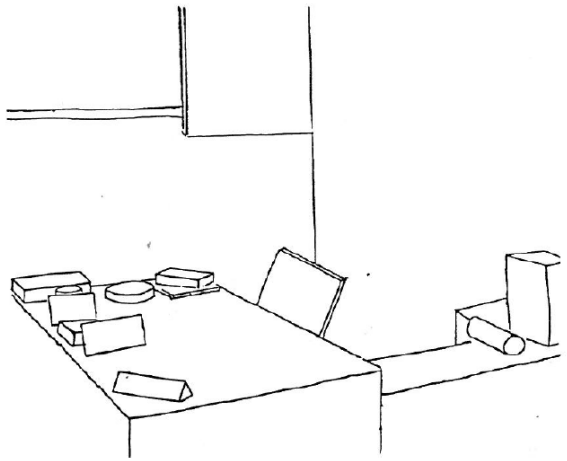
- Kersten's Digital Embryos for detecting camouflaged objects (Kersten ran the "null condition" which he thought would be impossible, but some people could do it). Tenenbaum's Tufas: humans can rapidly learn what a "tufa" is from few examples, and organize these (unfamiliar) objects hierarchy (consistent for different people).

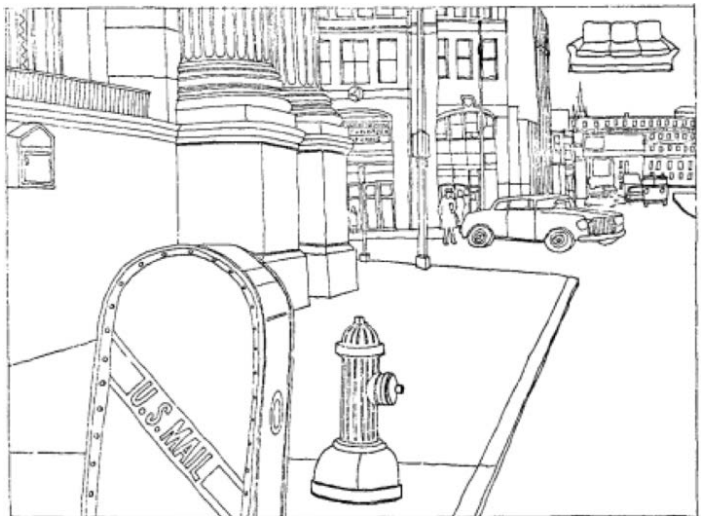


4. Strengths of Human Vision: Context

- Humans have the ability to use Context (for impoverished stimuli). This results from our knowledge about the world. C, von der Malsburg. Of course, for more realistic stimuli context may not be necessary.

Object recognition:
50% by context





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

4. Strengths of Human Vision: Gestalt 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)

4. Strengths of Human Vision: Gestalt Organization

- Gestalt Organization (left). Dalmation dog (right).

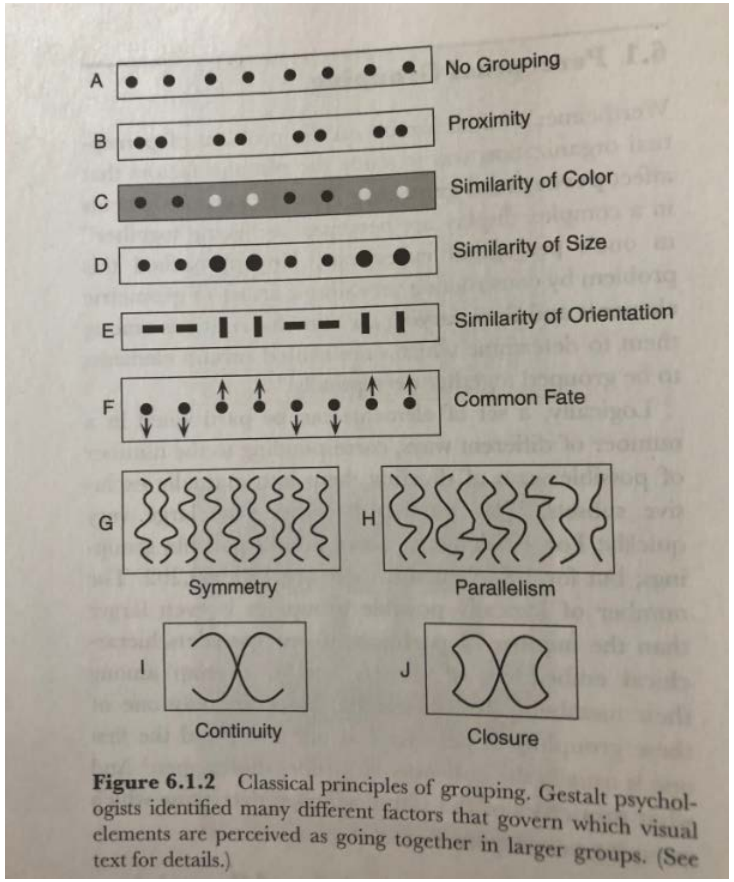
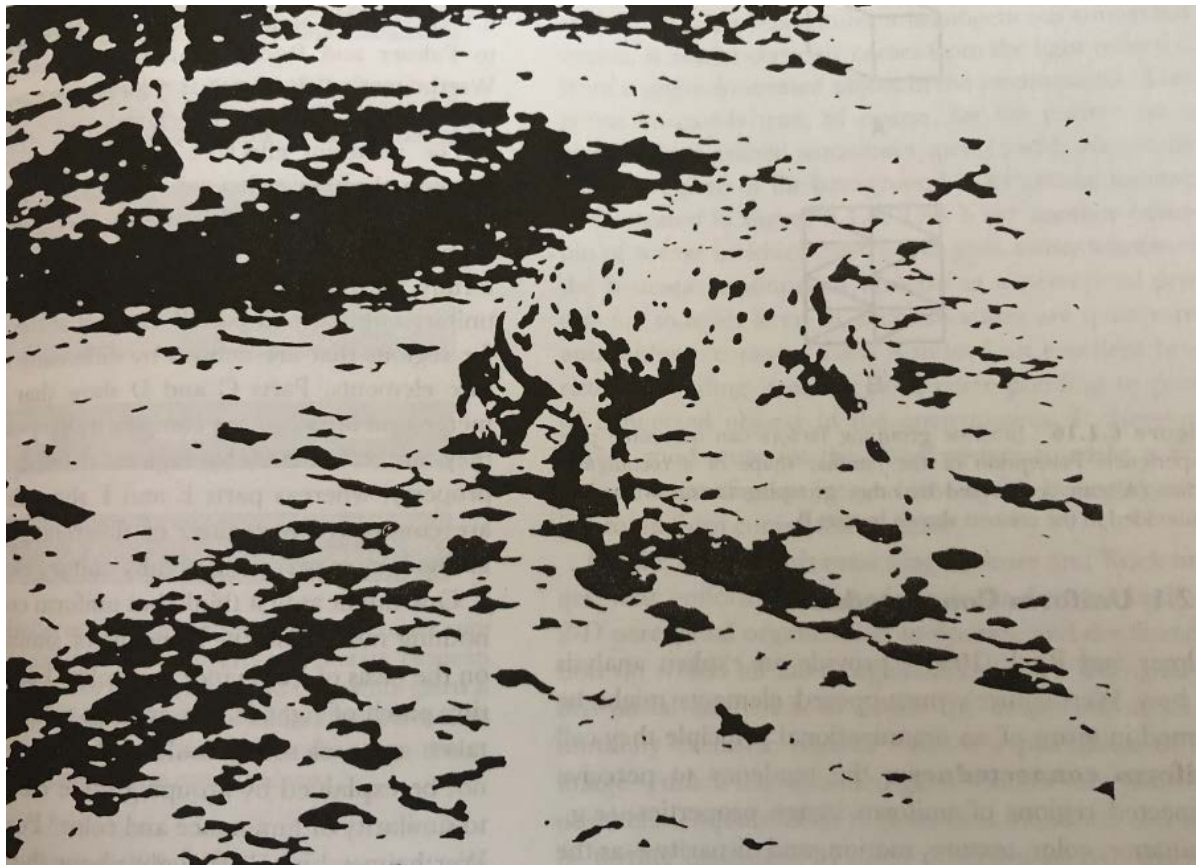
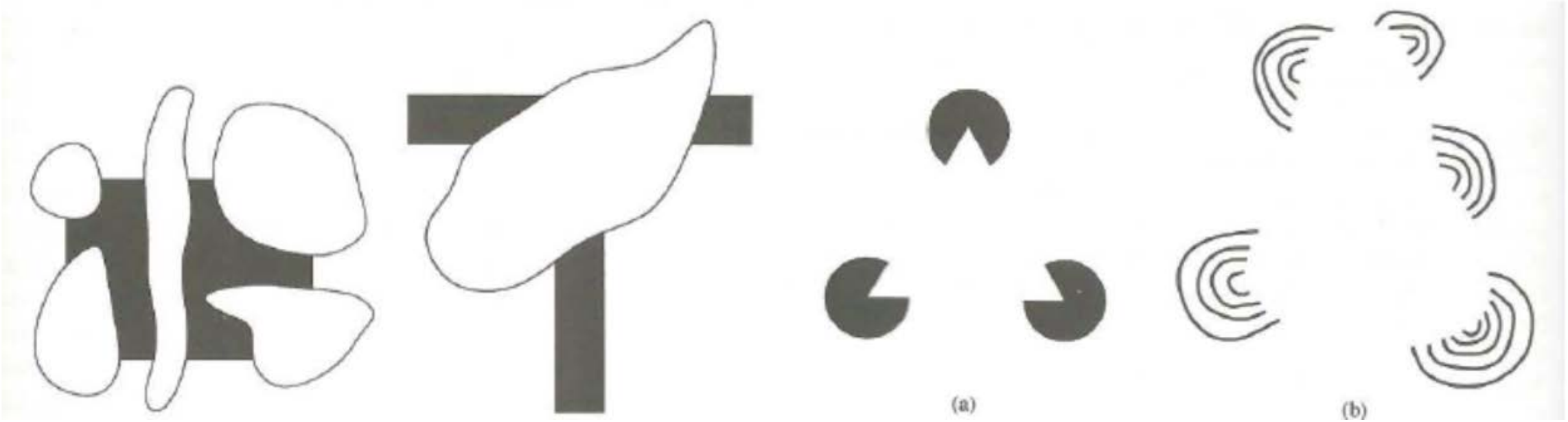


Figure 6.1.2 Classical principles of grouping. Gestalt psychologists identified many different factors that govern which visual elements are perceived as going together in larger groups. (See text for details.)



4. Strengths of Human Vision: Gestalt Organization

- Kanisza. All humans (almost all?) perceive similar/identical interpretations (foreground objects, in white, occluding background objects).



4. Strengths of Human Vision: Visual Cues

- 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. For simplified stimuli, the cues are sufficient (modular) but it complex stimuli they are tangled together.
- Many of these cues are now embedded in AI-Vision models (and were helpful for motivating AI-Vision algorithms in the 1980's), but many are not.

4. Strengths of Human 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 (domain transform is seen as a challenge for Deep Learning).
- Humans can factor shape and geometry – and recognize a blue tree, even if they have never seen one before. Humans can also reason about occlusion and complex foreground-background relationships (see Kanisza figure earlier).

- 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!). This requires abstraction and domain transfer.

4. Strengths of Human Vision: Abstraction and Domain Transfer.

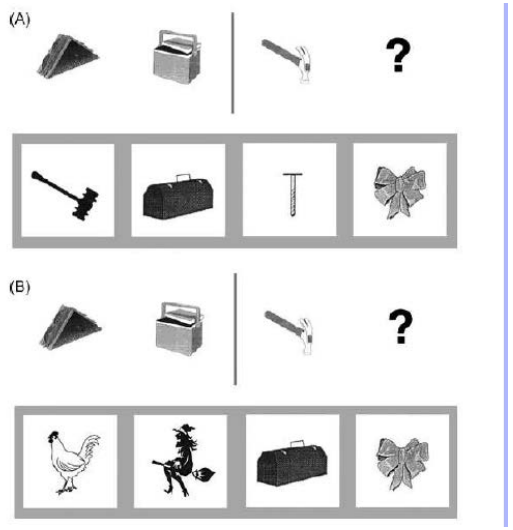
- Face Examples (but front-on faces may be easy). Easy for a human to realize that these are all faces – despite huge differences in the images.

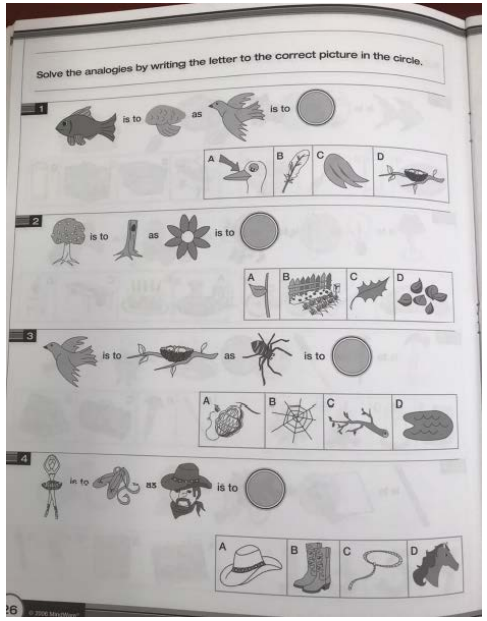
4. Strengths of Human Vision: Abstraction and Domain Transfer.

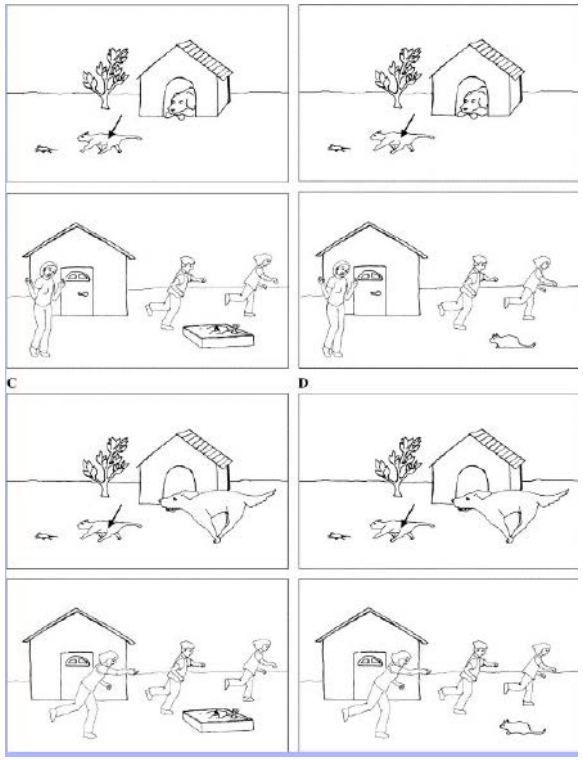
- Text Fonts. Humans can easily read text/digits in new fonts.
- Real Humans and Point Light Sources. Point-light-sources are very different from real humans, but NI-vision can perceive human motion from point-light-source stimuli.

4. Strengths of Human Vision: Analogies

- Some are visual. Others are semantic (e.g. part-f), while others are functional.





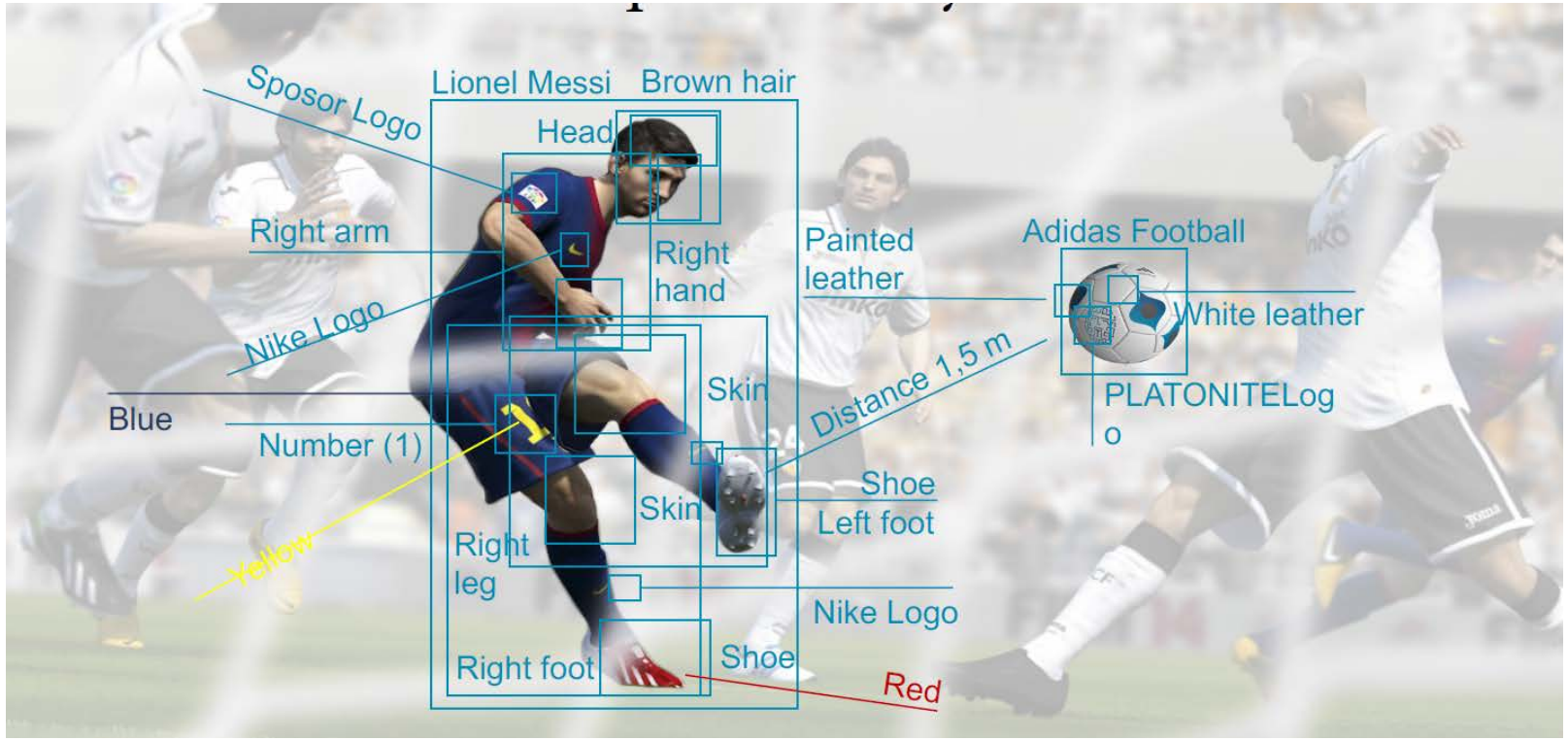


4. Strengths of Human Vision: Internal Representations of Parts.

- Internal representations of parts. This seems necessary to deal with abstractions and some domain transfer. It also helps make vision explainable.
- These parts can be detected without context (not always, but often). 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.

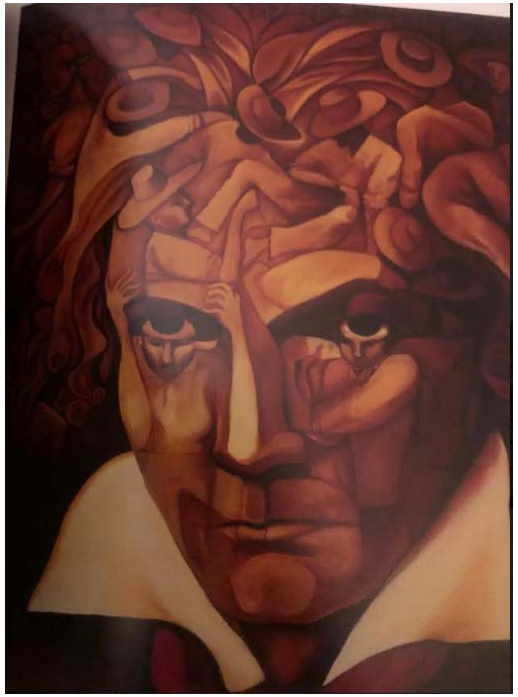
4. Strengths of Human Vision: Internal Representations of Parts.

- Part Examples (from von der Malsburg). Detecting and describing the footballer in terms of his parts is necessary for understanding his actions.



4. Strengths of Human Vision: Internal Representations of Parts.

- Parts and Abstractions.
- Beethoven constructed from Humans. A fish made from bicycle parts.



Part 5. This course.

- The goal of the course is to formulate vision as Bayesian inference. But, in practice, it falls short and describes the 20+ techniques every computer vision researcher should know.
- This involves techniques like linear/non-linear filtering, Bayesian decision theory, markov random fields, neural network models, geometry, radiosity, adabost, support vector machines, deep networks, and more.
- The study of human vision gives challenges to computer vision algorithms. We will not solve vision until we have overcome them.