Vision as Bayesian Inference
Spring 2025

Alan Yuille

Bloonberg Distinguished Professor

Depts. Computer Science and Cognitive Science

What is Vision?

- Humans can extract an enormous amount of information from this image. *We can perform many visual tasks* -- recognize all these objects, estimate their 3D pose and other properties, their positions in the 3D world.
- *We can combine this with common sense knowledge about* the physical/functional/social properties of objects, agents, and their interactions.
- We can answer questions like *“what accidents are waiting to happen in this picture?”*

A Brief History of Vision

- When I started researching on Vision in 1982 the subject was in its infancy. The subject was only studied at a few universities worldwide, almost all in the US and Canada.
- There were hardly any courses. There were very few techniques that researchers needed to know.
- Much work was on simple synthetic stimuli because only a limited number of images could be stored on computers – and the internet was over ten years in the future. Special purpose hardware was designed so that convolution could be performed in a few seconds.

Taming the Wilderness

- Computer Vision (CV) researchers found it was extremely hard to design algorithms that worked well on natural images.
- Images were very complicated and not well understood. Researchers tried developing algorithms on simple synthetic stimuli. But found that this algorithms rarely worked on real images.
- But gradually, over time, CV researchers made progress. They divided vision up into different visual tasks and studied them separately. They borrowed and adapted mathematical and computational techniques from other disciplines.
- There were conceptual big picture theories and attempts to put CV on a sound theoretical basis, but most of these were judged to be impractical,

Computer Vision developed a large toolbox.

- Perspective Geometry. Geometry of Surfaces. Linear Algebra.
- Sparsity. Radiosity. Linear filtering.
- Non-linear filtering. K-means. K nearest neighbors. Stochastic Grammars. Probability Distributions on Graphs. Markov Random Fields.
- Free Energy methods. Mean Field Theory. Belief Propagation.
- Max-Flow MinCut. Dynamic Programming. The EM algorithm.
- Stochastic Processes. Support Vector Machines. AdaBoost, Reinforcement Learning. and much more.
- An attempt to organize “twenty techniques that all CV researchers should know” stopped when the list got to over four hundred.

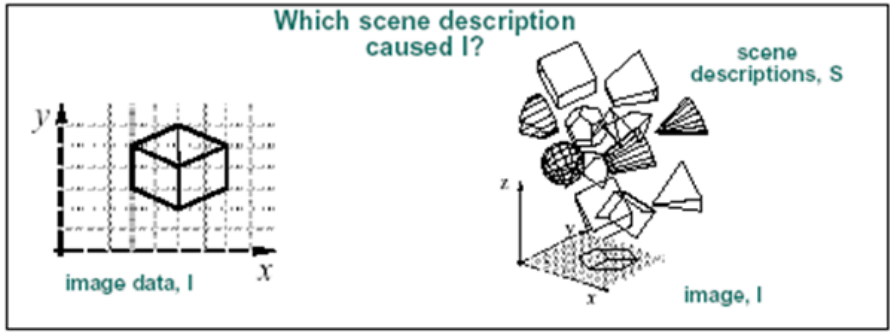
Computer Vision and Bayesian Inference

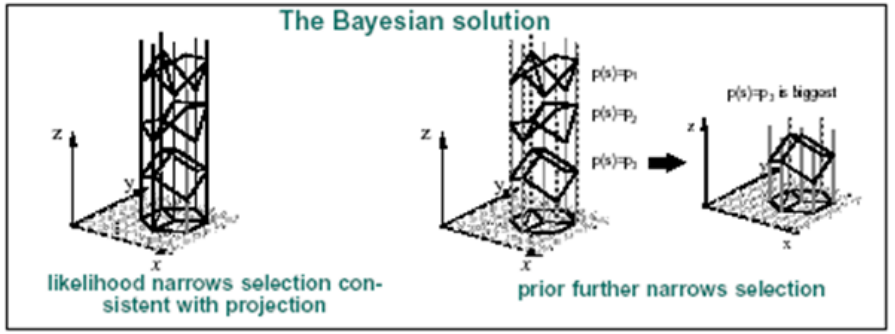
- For myself, I was drawn to the concept of vision as Bayesian Inference and Analysis by Synthesis.
- This requires modeling the types of patterns that happen in images and how they are generated by properties of the 3D world.
- This includes three core ideas: (I) Causal models of how patterns are generated instead of simple statistical models of patterns. (II) modeling the underlying 3D/4D world. (III) Structured compositional representations.
- Also, the ability to learn new models from small amounts of data, building on prior knowledge about the world, the ability of algorithms to perform many tasks in a consistent manner, and to generalize to novel stimuli.

Analysis by Synthesis and Bayes.

- P(I|S) – models how the image is generated from the 3D world.
- P(S) – prior knowledge of the world.
- P(S|I) – the posterior distribution.
- P(S|I) = P(I|S) P(S)/P(I) – Bayes Rule

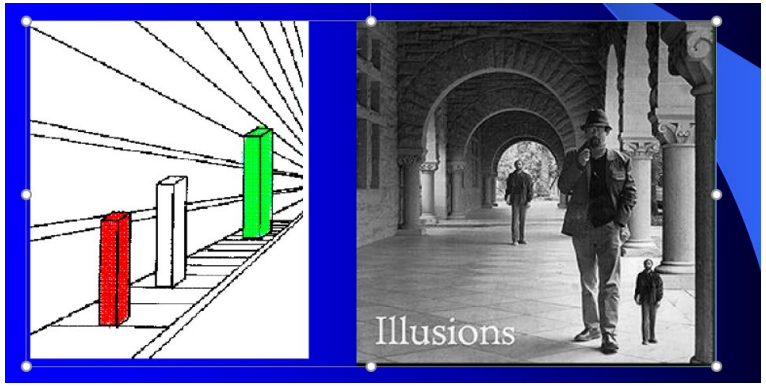
- Why do we see a cube? (P. Sinha).
- There are many ways this image could be formed.
- The likelihood P(I|S) rules out some interpetations S
- Prior P(S)– cubes are more likely than other shapes consistent with the image.

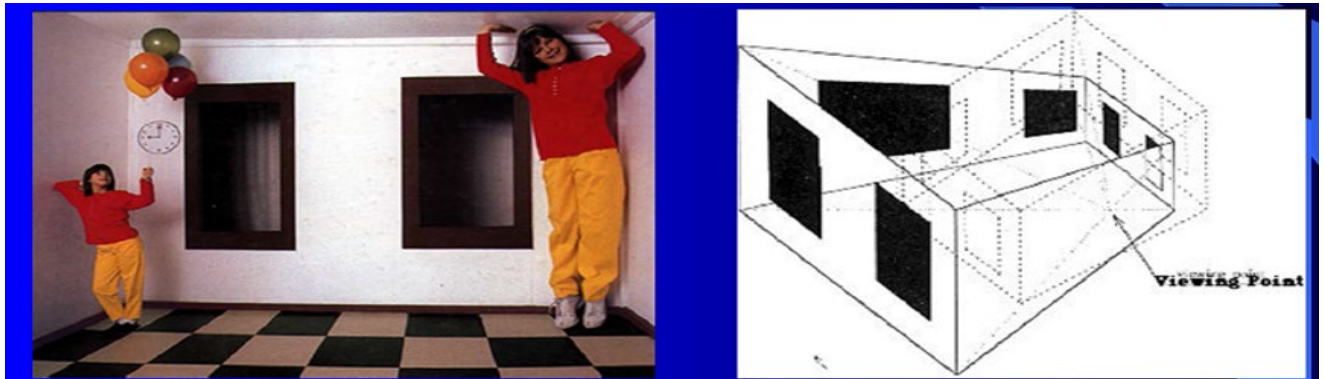




Illustrations that Humans use Bayes

- The world is 3D. We interpret images assuming normal 3D structure (ground planes, shadows, and 3D structure), but can be fooled. *Easy to demonstrate but hard to prove.*





Human Perception, Cognitive Science, and Bayes

- Michael Bach has a website of a large number of visual illusions and qualitative explanations about how they can be explained using Bayesian Ideas. https://michaelbach.de/ot/
- Cognitive Scientists show that Bayesian models can account for a large range of cognitive phenomena.
- “Bayesian Models of Cognition: Reverse Engineering the Mind”. Ed.s T. Griffith, N. Chater, J. Tenenbaum. MIT Press. 2024.
- Their work concentrates of non-vision cognition. But they also emphasize the same three principles: (i) casuality instead of statistics, (ii) world models, (iii) structured compositional representations.

Back to Computer Vision

- Bayesian techniques were fairly common until the rise of big datasets. Their conceptual advantages were acknowledged.
- But as the machine learning paradigm – training and testing algorithms on balanced annotated datasets – the value of Bayes declined.
- In this paradigm, only the posterior distribution P(S|I) is important. There is no need for P(I|S) and P(S). Bayesian methods could still be used but they were generally slower and less effective.
- This was particularly true with the rise of deep networks and large datasets like ImageNet and CoCo.

Convolutional Neural Networks (CNNs).

- *The provocation:* AI researchers were skeptical about the potential of Deep Learning (Frontiers of Computer Vision NSF workshop 2012). “What will it take to convince the Computer Vision community that Deep Networks are the future?” Yann LeCun was at the workshop. So was Fei Fei Li, who was creating ImageNet.
- *The Three Ingredients:*
- (1) *Very Large Dataset (Millions of Images) with Annotation*. Divided into Training Set, Validation Set, and Test Set.
- (2) *Deep Network Regression Model P(C|I;W) as a* ***differentiable*** *function of parameters W. This enables efficient learning* by Stochastic Gradient Descent (Backpropagation). Specifically Convolutional Neural Networks.
- (3) Computers. *Graphics Processing Units (GPUs).* Required skilled Cuda programmers soon replaced by Python packages.
- CNNs needed big datasets (ImageNet had 1,000,000 images, earlier datasets were orders of magnitude smaller).

The Rise of Large Benchmarked Datasets

- *The success of CCNs on ImageNet led to an explosion of large datasets.*
- There are 50-1000 benchmarked datasets (depending on how you count). Evaluating is done by average case performance with the ML paradigm (testing data is similar to training data). Competition is fierce and worldwide. Publishing requires outperforming state of the art (SOTA) on several benchmarked datasets.
- *Each is for a specific visual task. E.g., (1) ImageNet for object classification, (2) CoCo for semantic segmentation, (3) Datasets for estimating human 3D pose, (4) Vision Question and Answer datasets, (5) Action recognition datasets.*
- Only a small number of visual tasks are represented (annotation is hard). Humans can perform many visual tasks in a unified manner, but AI mainly concentrates on a single task.

Vision as Machine Learning

- Three basic ingredients. The train-test paradigm.
- (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”?

Why Deep Networks always win?

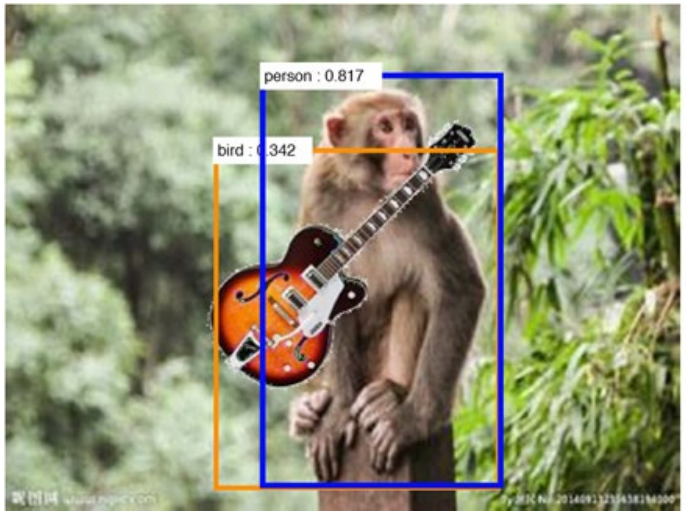
- **There are many AI techniques. But Deep Nets started always winning on benchmarks. Why?**
- (1) They have very simple and efficient learning algorithms --stochastic gradient descent – enabling end-to-end training. They can be implemented very efficiently on GPUs and are very fast. Alternative algorithms are slower and much harder to train.
- (2) Provided there is enough data, these visual tasks can be thought of as robust memorization. They take short cuts which exploit the biases of the datasets. *There perform badly on corner cases.*
- (3) They are well suited for to standard ML paradigm where AI algorithms are trained and tested on data which are random samples from the same underlying source. (There is mathematics -- Probably Approximately Correct -- which clarifies the basic assumptions). *There are less good at generalizing to data that differs significantly from the training data.*

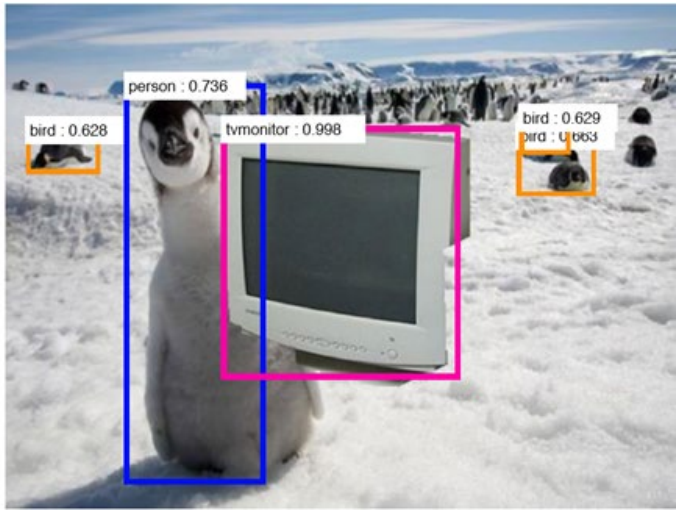
But is this Sufficient?

- “The proof that vision is solved is that the ImageNet object detection challenge has been retired. Deep networks scan 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) The Deep Networks exploit the biases of the dataset. It is possible to make small modifications to images in ImageNet and reduce performance of Deep Networks to almost zero.
- (3) The original studies comparing humans to deep networks were flawed. More careful studies show that humans and deep networks both perform very well on the easy cases. But human also perform well (but slower) on the harder cases, while Deep Networks start failing badly.

Deep Networks are non-robust and 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.

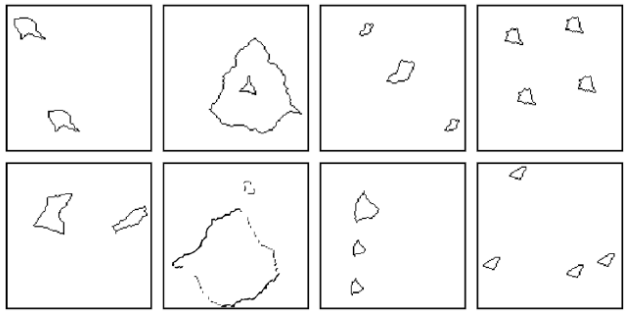




- See also “The elephant in the room” A. Rosenfeld et al. Arvix. 2018.

What are Deep Networks bad at? Spatial Structure.

- Visual Task: *assign an image to one of two categories determined by the spatial arrangement of constituent parts. (D. Geman. T. Serre).*
- CNNs are not good at this task. Even with huge training data.



- Arguably an easy tasks for a stochastic grammar – e.g., L. Zhu et al. 2008.
- But this is not a large benchmarked dataset so my students won't work it.
- There are many tasks – Kanizsa triangles – where non-neural networks work much better than neural networks.

What are the problems?

- ***(1) Domain Transfer:*** AI trained on data from one source often does not transfer and perform well on data from another source (e.g., change of weather, counting crowds in New York of in Beijing).
- ***(2) Lack of Robustness, Fairness, and Interpretability:*** AI algorithms have unexplainable failure modes and lack fairness. Like (1), this reflects AI's weakness at generalizing outside the training domain (relates to the infinite space of images).
- ***(3) Limited Visual Tasks:*** AI only performs a small number of visual tasks because it is hard to create annotated datasets for some tasks, particularly those involving 3D. AI researchers have good datasets for the 3D depth, shape, and pose of humans – so there are good AI algorithms for estimating 3D properties of humans, but not for most other object categories.
- (4) ***Specificity of AI:*** AI researchers use different neural architectures for different visual tasks. No unified neural architecture which can address all visual tasks.

Mega Data: the Empire Strikes Back

- The second generation of Deep Networks is much richer. The models are trained on mega data. Vision models are coupled to large language models (LMMs). This enables them to have access to the commonsense knowledge about the world learnt by LMMs and embodied in GPTs.
- Has this changed my belief in Bayes?
- No. These VLMs are both extremely smart and sometimes very stupid.
- If anything this second generation of Deep Networks makes it practical to construct Bayesian models which are capable of dealing with the complexity of real world stimuli. Specifying models like P(I|S) is extremely challenging and impossible without big data,

Illustration: The Genex Project

- This project enables AI to imagine an explore a virtual world initialized on a single input image.
- https://engineering.jhu.edu/news/a-generated-world-of-pure-imagination/

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 (what color is the hair on the horse? is the horse 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 – in short, 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.
- “Language is high level vision”

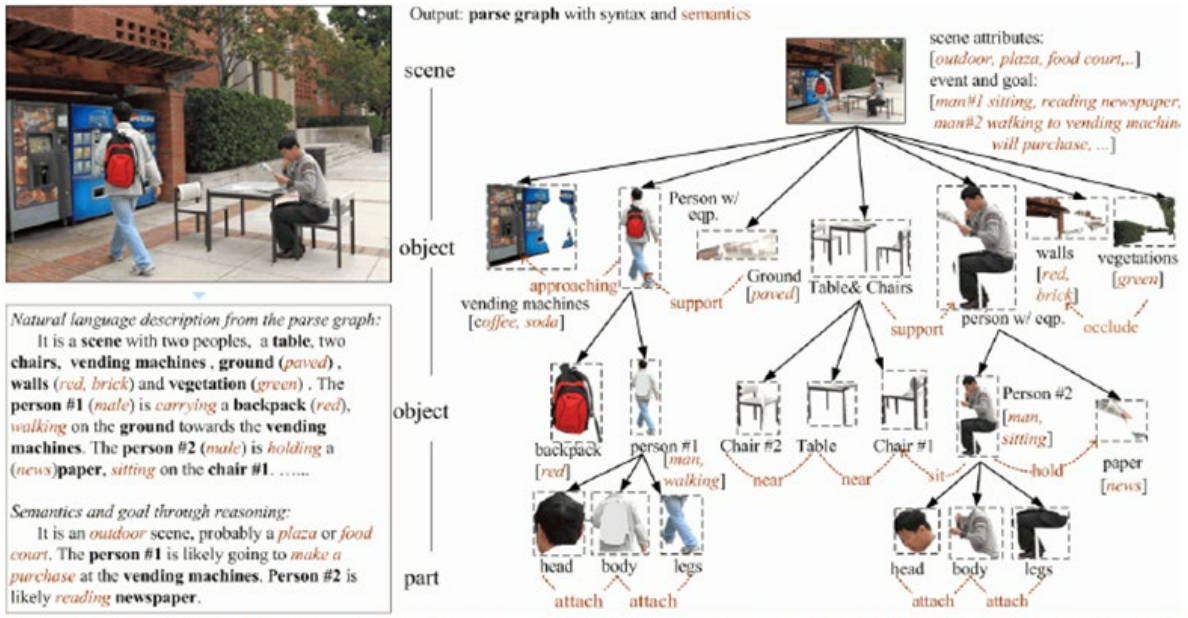


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description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

Part 2: Why is Vision Difficult?

- **(1) There are infinitely many images.**
- *The space of images is infinite. The few we see during our lifetime, or those seen during the lifetime of our species, are at best a drop in the ocean.*
- *Mathematically speaking, there are many different types of infinities and some are much bigger than others. Arguably, images are a "big" type of infinity.*
- **(2) There are very many visual tasks, but we can only evaluate a few of them.**
- *Assuming that purpose of vision is to extract information about the world, we can taxonomize this into performing* ***visual tasks****.* E.g., detect an object, classify an object, segment and object, estimate the material properties of an object. AI researchers evaluate performance on benchmarked datasets which are annotated for specific visual tasks.
- *I don't know an exhaustive list of visual tasks that humans can perform. But we can surely make a list of visual tasks that are most vital for our species to survive in the real world.* (I once worked on AI to assist the blind and visually impaired – this gave some insight).
- ***So we can only study a few visual tasks on a tiny subset of possible images.***

Part 2. Why is Vision Hard? Complexity.

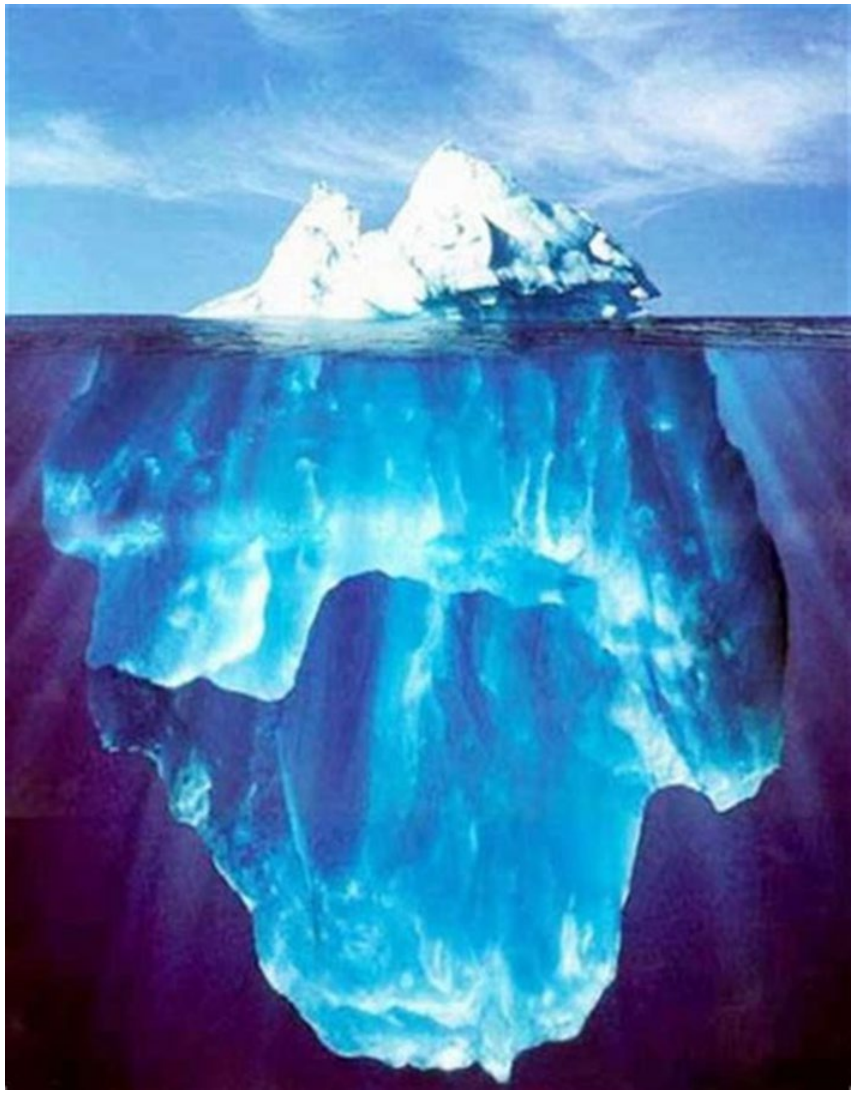
- Vision is extremely hard due to complexity and ambiguity.
- Complexity arises in several forms. The complexity of all images. The set of images is infinite. If we restrict each pixel value to take 256 possible values (as in a digital camera), then there are more 10x10 images than have been seen by all mankind over all history and pre-history. Humans see roughly 10^9 each year (assuming 30 frames per second).
- 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. The tip of an enormous 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 (most of the time).

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

Bayes Models
Knowledge of 3D World

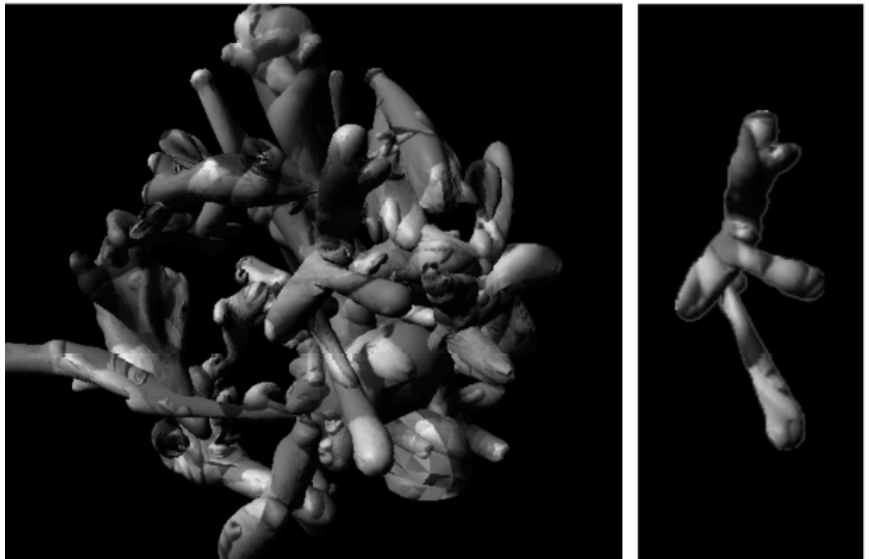
- Knowledge of the 3D world.
- The ability to do inverse-computer-graphics requires knowledge of the external 3D world.
- The Necker Cube.– ambiguity is removed by using prior knowledge of what objects are most likely to be present in the world (cube are more likely than the other objects which are consistent 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, or intuitive, physics (Tenenbaum). Perhaps only for situations which we are familiar with, like catching a ball or balancing a stack of books.

Bayes Models
Knowledge of the 3D World

- How humans acquire this knowledge? Development and Learning.
- Human vision uses an enormous amount of knowledge which is learnt over a lifetime (unlike current AI-Vision systems).
- This knowledge is mostly learnt by infants and children 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 on exploiting image sequences, searching for causal structure, taking actions in the world, and exploits other senses. The Theory of Mind (Gopnik) proposes that infants are baby scientists exploring the world by making predictions which they explore by experiment.

Knowledge of the 3D world.

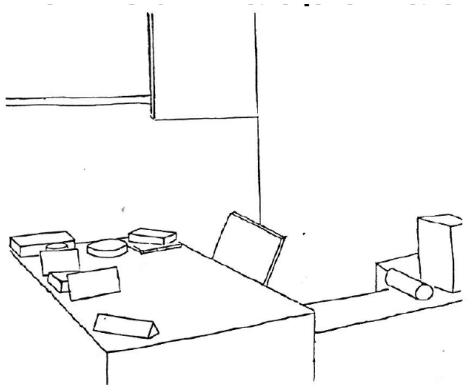
- Kersten's Digital Embryos: Humans are able to learn and detect camouflaged objects.
- Tenenbaum's Tufas: humans can rapidly learn what a "tufa" is from few examples, and organize these (unfamiliar) objects hierarchy (consistent for different people).

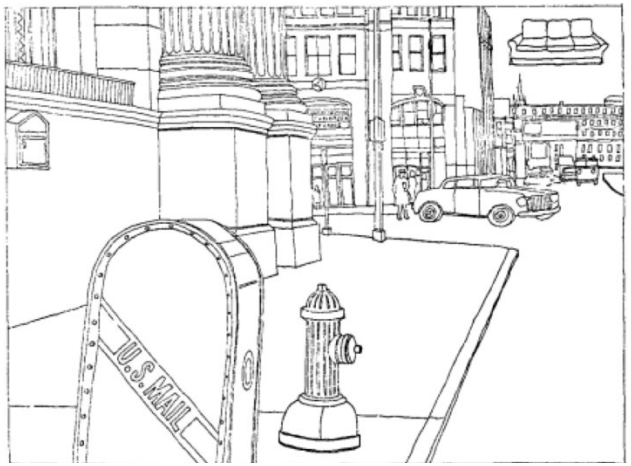


Ability to Exploit Context

- Humans have the ability to use exploit context for impoverished images (left). The more realistic the image the less the amount of context needed.

Object recognition:
50% by context





- Object recognition: 50% by context..

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).

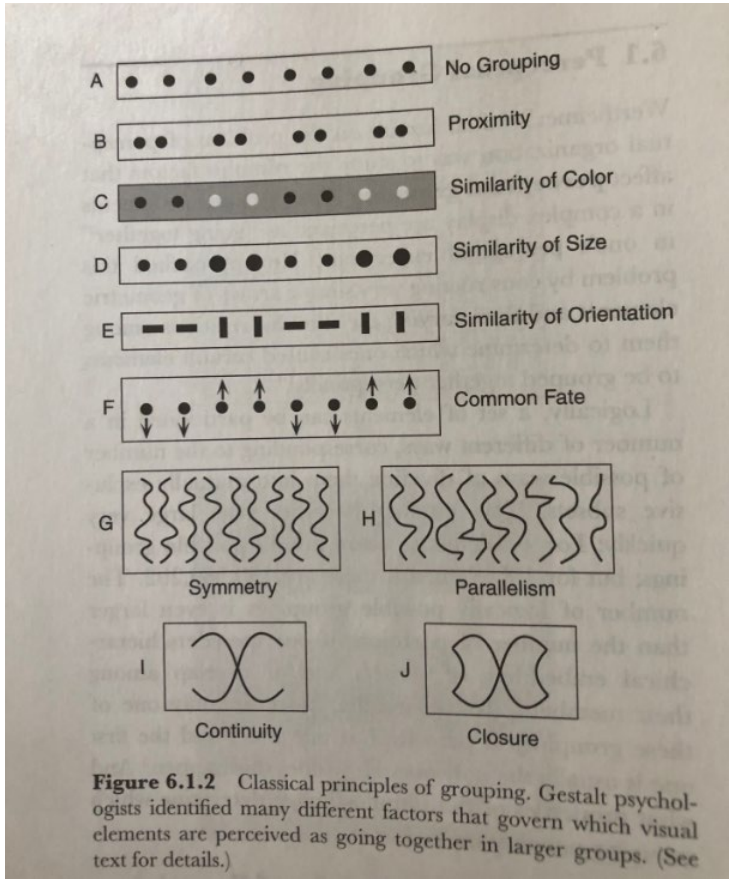
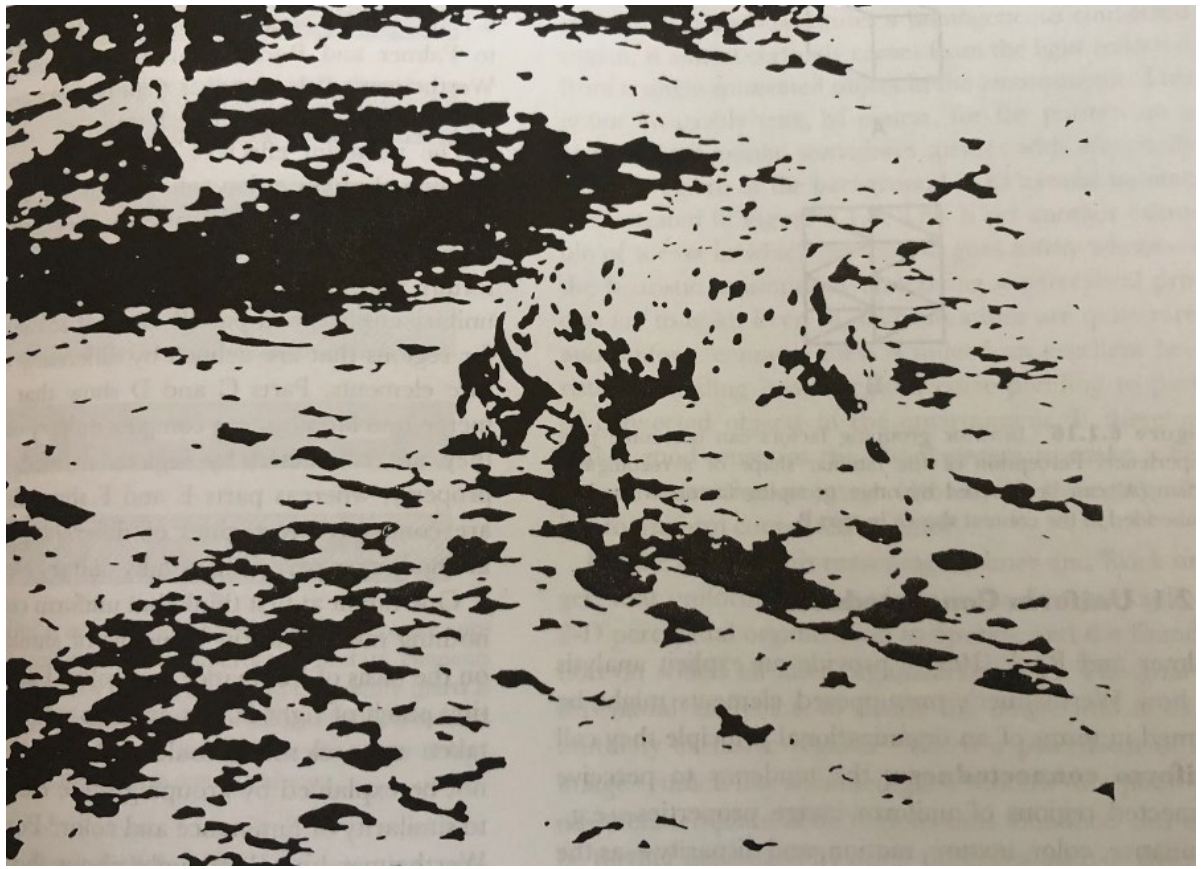
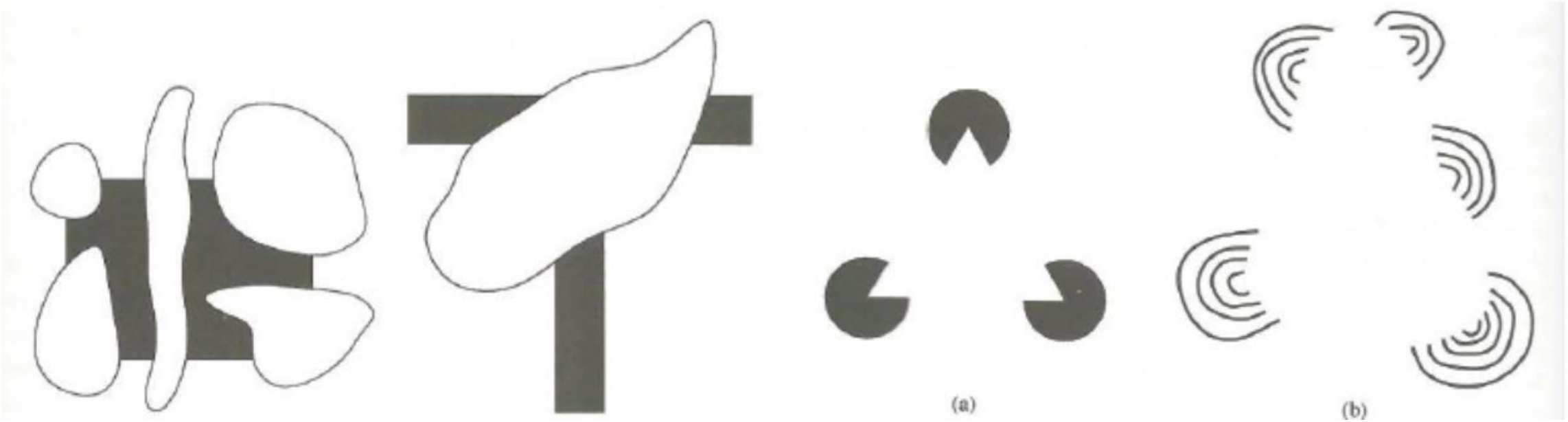


Figure 6.1.2 Classical principles of grouping. Gestalt psychologists identified many different factors that govern which visual elements are perceived as going together in larger groups. (See text for details.)



Gestalt Organization

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



The Variety of 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 strongly coupled.
- Many of these cues are now embedded in AI-Vision models (and were helpful for motivating AI-Vision algorithms in the 1980's), but some are not.

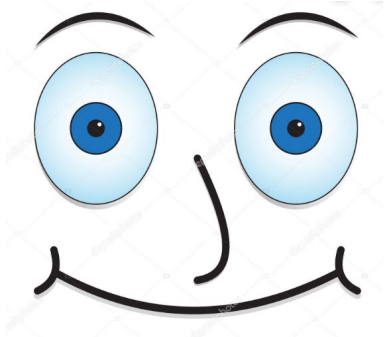
Abstraction and Domain Transfer

- Humans can understand an object from an image, from a drawing, from an highly abstract sketch. 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).

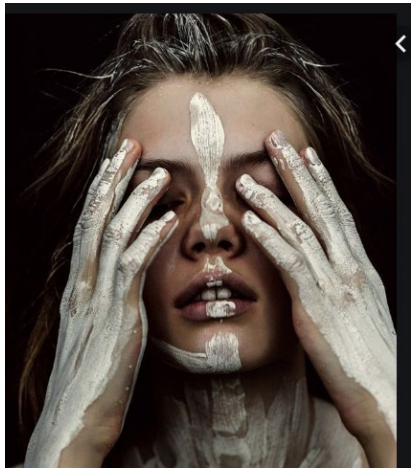
- Humans can perform analogical reasoning. We can not only recognize visual similarity between objects, but also relationships (e.g., part-whole: paw to cat, hand to person), and functional relations (e.g., hammer is in toolbox, notebook is in backpack), and other relations (e.g., woman chases child is analogous to a cat chasing a mouse). This requires abstraction and domain transfer.

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.




Abstraction and Domain Transfer.

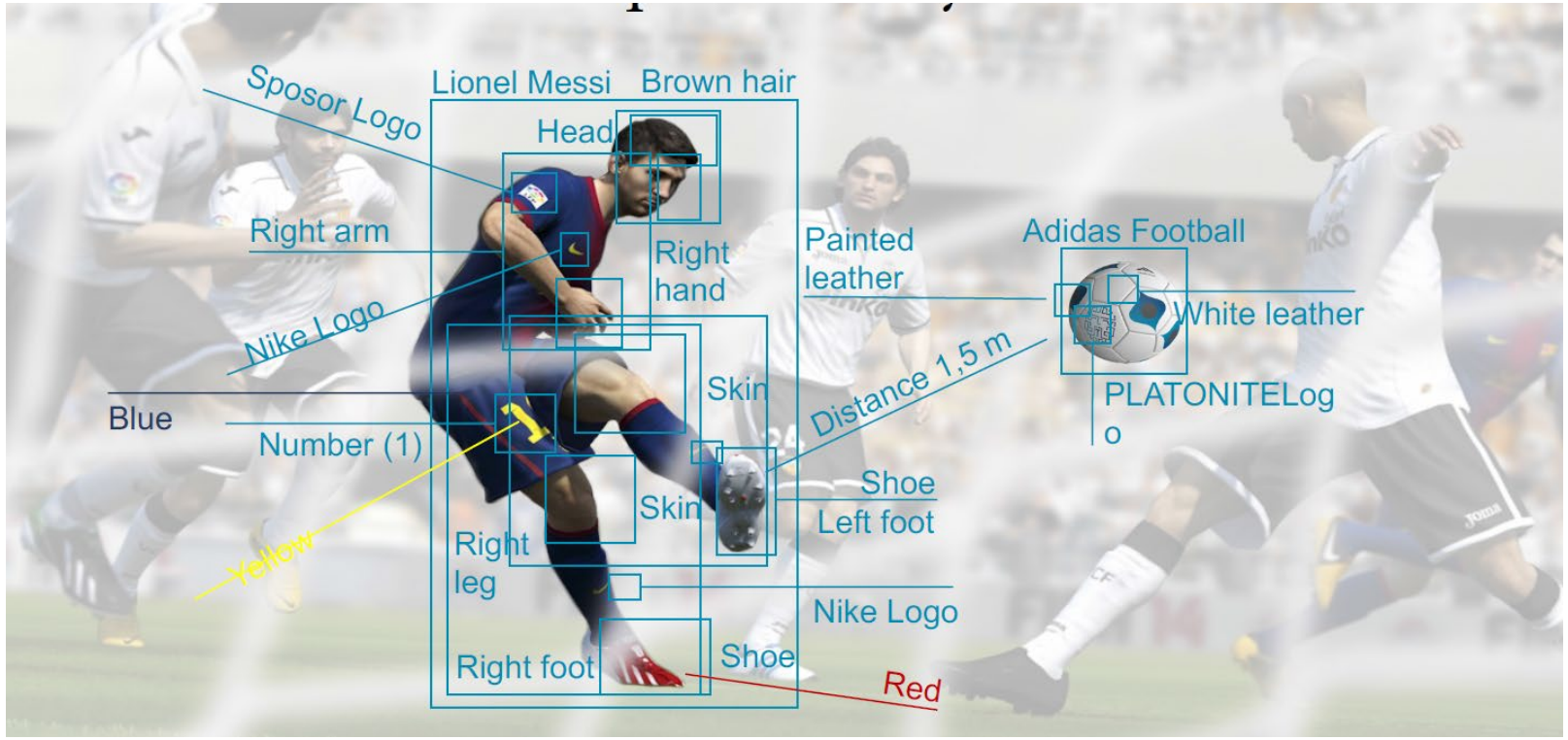
- Text Fonts. Humans can easily read text/digits in new fonts.
- Real Humans and Point Light Sources. Point-light-sources are very different from real humans, but NI-vision can perceive human motion from point-light-source stimuli.

Structured Representations of Parts.

- 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.

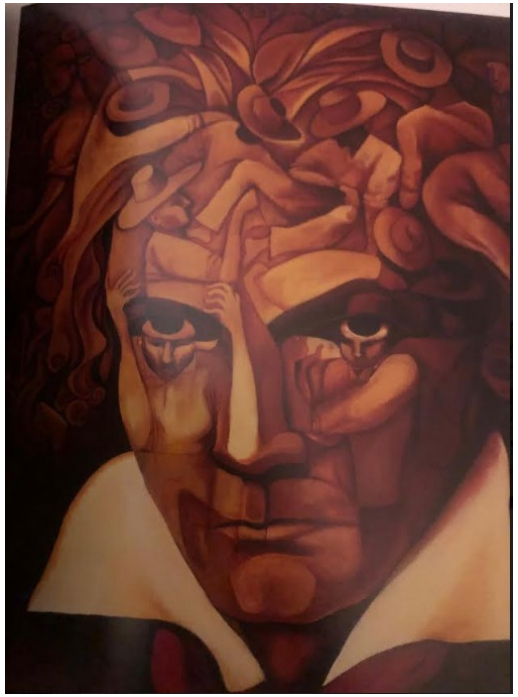
Structured Representations of Parts.

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



Structured Representations of Parts.

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



The Big Picture: AI Vision Bayes.

- The CogSci literature suggests that humans perform cognition by building models of the 3D. These are learnt during development using multi-modal cues and interactions with the environment. *Infants are tiny scientists who learn about the world by performing experiments on the world and on their parents.*
- A similar idea is the concept of *Embodied AI.* Vision, Language, and Robotics should be unified into a theory of agents and object interacting in the 3D world.

This course.

- The goal of the course is to argue that Vision should be formulated in terms of world model and Bayesian inference.
- The course is being updated to take into account recent progress: Transformers, Self-supervised learning, GANs, Diffusion models, Sign-distance functions, Nerf, vision-language models, Large Language models, Auto-Encoders.
- The study of human vision gives challenges and inspiration to computer vision. The human visual system remains the gold standard for visual systems.