Image and Video Understanding

CM0524

Marcello Pelillo

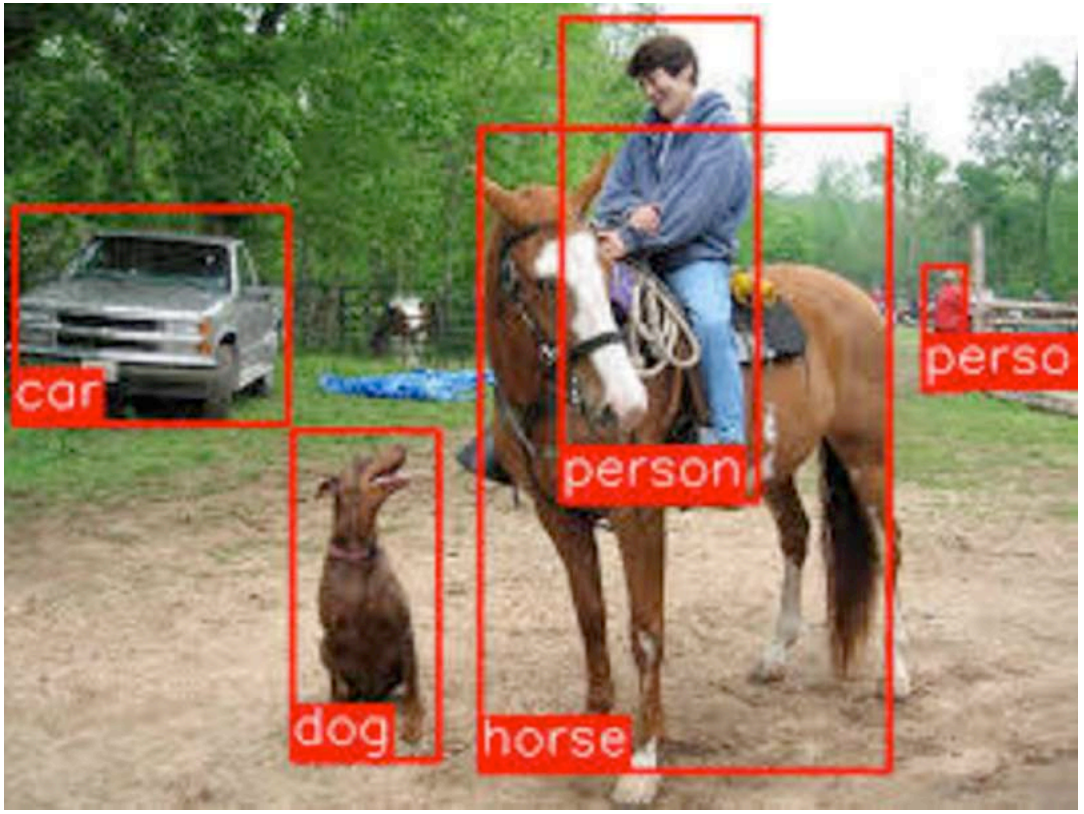
University of Venice

a.y. 2018/2019

What does it mean to see?

The plain man's answer would be, to know what is where by looking.

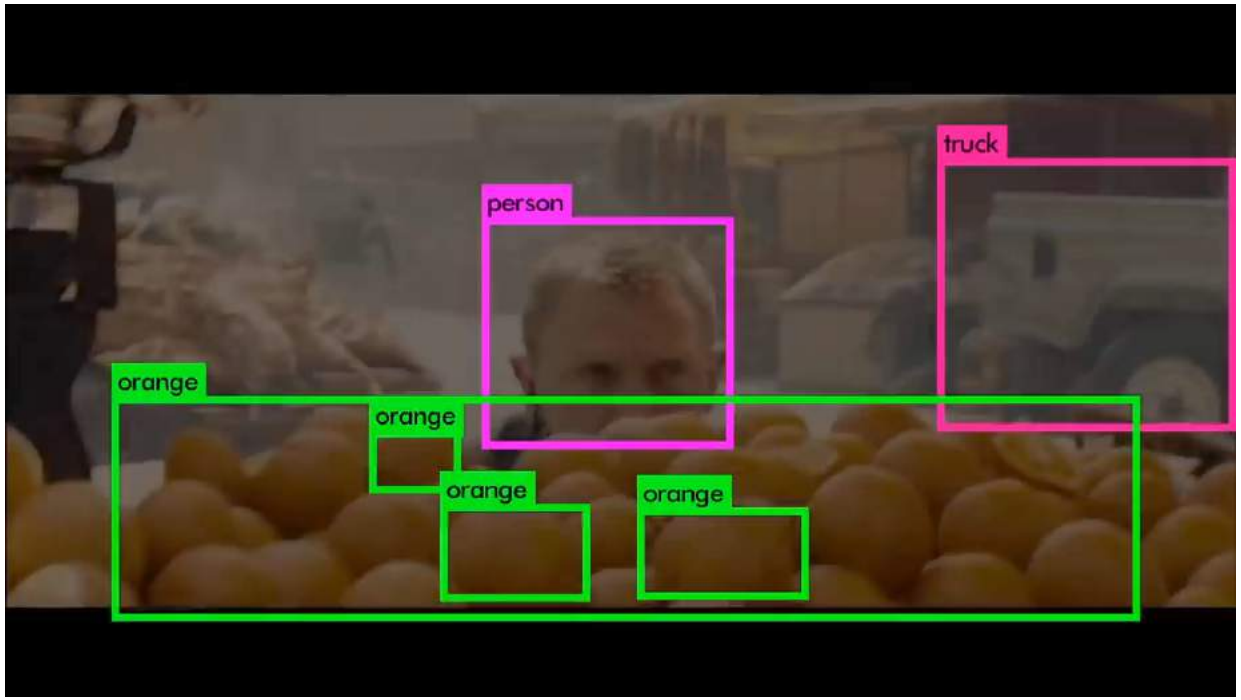
David Marr (1982)



What does it mean to see?

The plain man's answer would be, to know what is where by looking.

David Marr (1982)

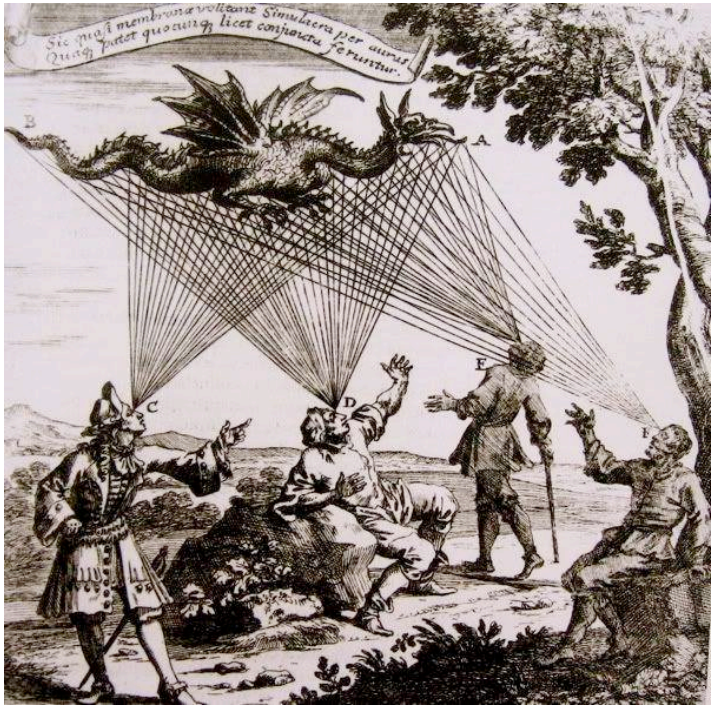


Theories of Vision:
From Pythagoras to Marr

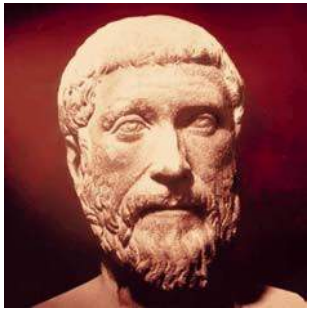
The emission theory

The eye obviously has fire within it,
for when the eye is struck fire flashes out.

Alcmaeon of Croton (ca. 450 BCE)

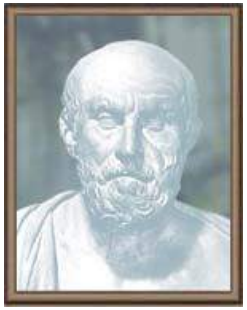


Pythagoras



Euclid

Empedocles



Surprisingly enough …

Fundamentally Misunderstanding Visual Perception

Adults' Belief in Visual Emissions

Gerald A. Winer and Jane E. Cottrell — *The Ohio State University*
Virginia Gregg — *State University of New York at Oswego*
Jody S. Fournier and Lori A. Bica — *The Ohio State University*

June/July 2002 • American Psychologist

 0003-066X/02/$5.00
Vol. 57, No. 6/7, 417–424 DOI: 10.1037//0003-066X.57.6-7.417

> We have shown that many college students believe in visual extramissions, as shown by a variety of measures and probes, and what we find most significant and surprising is that this belief is extremely resistant to standard educational experiences that seem as though they should counteract the misunderstanding.

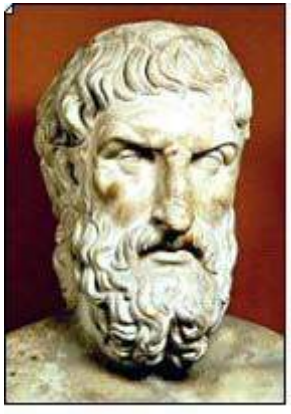
The intromission theory

There are what we call images of things stripped off the surface layers of substances like membranes—these fly to and fro in air.

Lucretius, On the Nature of Things (De Rerum Natura)

Democritus
(460–360 a.C.)

Epicurus

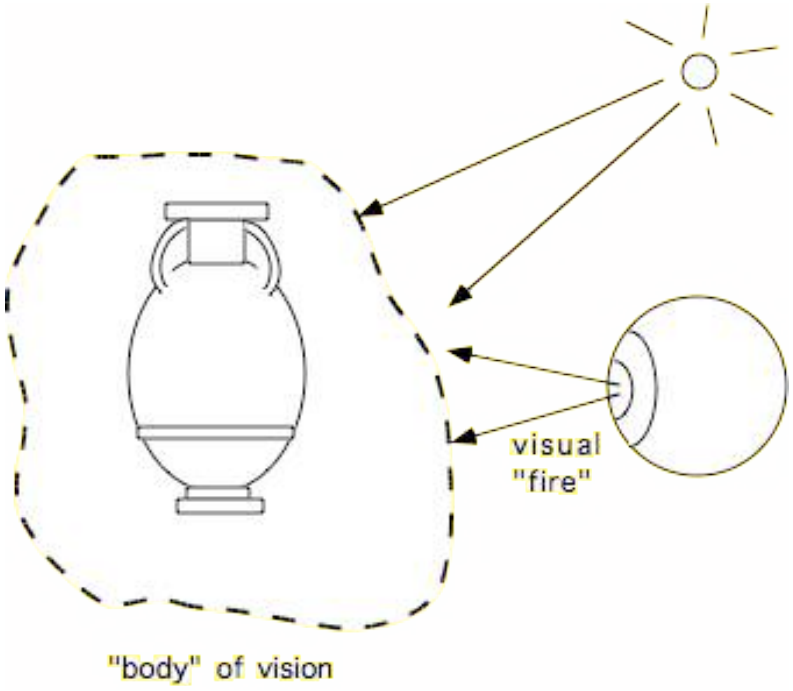


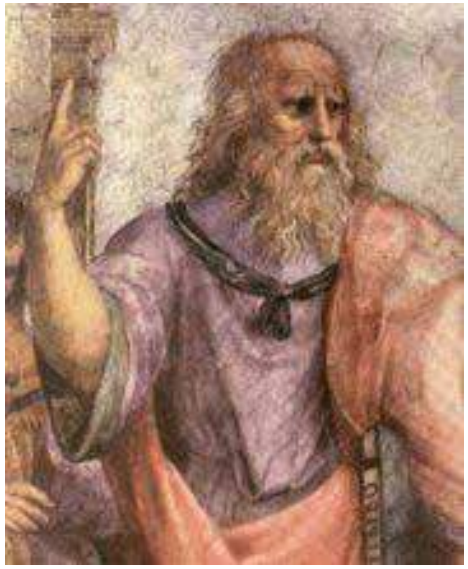
Lucretius

Plato's view

Accordingly, whenever there is daylight round about, the visual current issues forth, and coalesces with the daylight and is formed into a single homogeneous body in direct line with the eyes

Plato, Timeus





After centuries …

The most serious problem facing the Muslim heirs of Greek thought was the extraordinary diversity of their inheritance.

Among theories of optics, for instance, Muslim thinkers had the following choice:

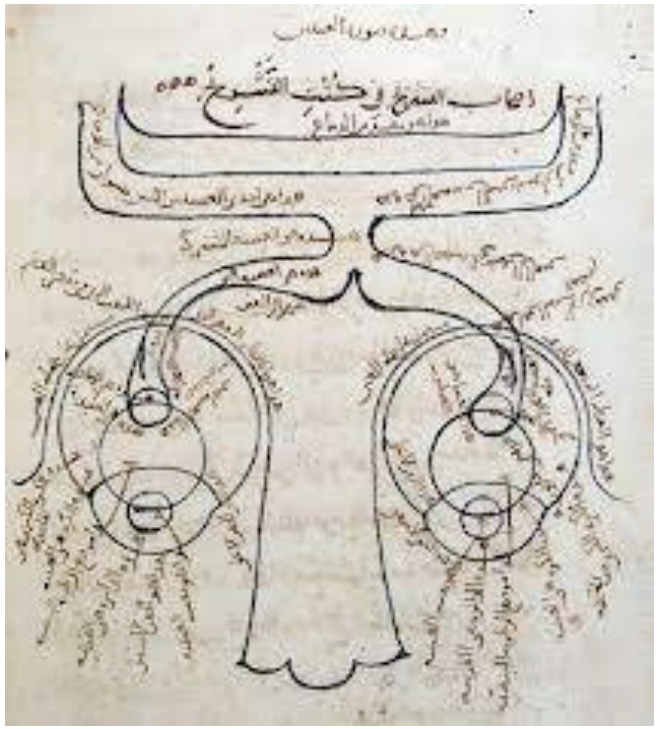
- the emission theory of sight of Euclid and Ptolemy, which postulated visual rays emanating from the observer's eye;
- the older Epicurean intromission theory, which reversed the rays and made them corporeal;
- the combined emission-intromission theories of Plato and Galen;
- some enigmatic statements of Aristotle about light as qualitative change in a medium.

From: D. C. Lindberg (1967)

Alhazen's synthesis

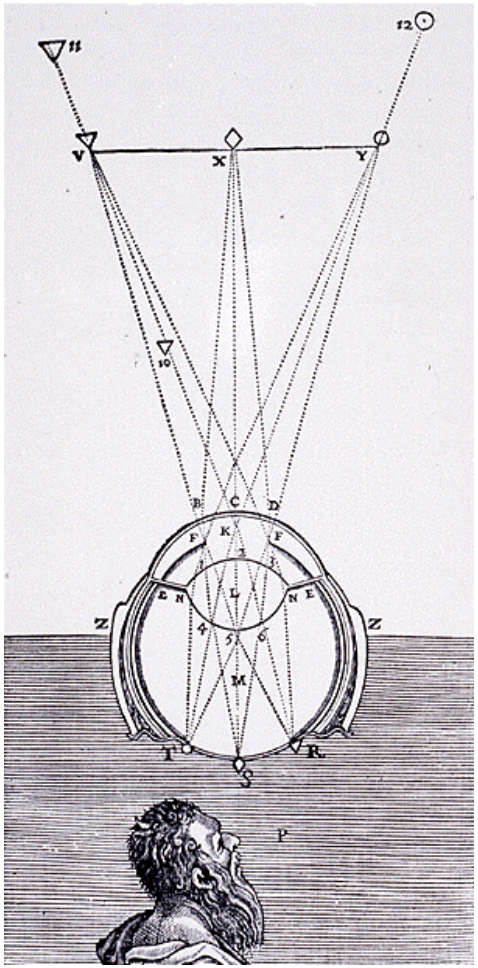
From each point of every coloured body, illumination by any light, issue light and colour along straight lines that can be drawn from that point.

Alhazen, Book of Optics



Alhazen, 965-1039 AD

Kepler's modern theory of retinal images



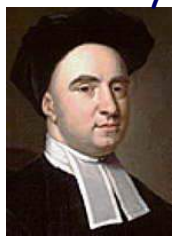
Kepler, 1571-1630

Nativism *vs.* empiricism

Cartesio

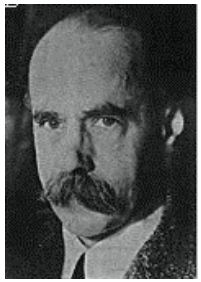
Kant

Berkeley

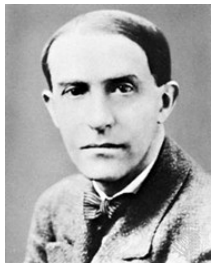


Locke

Wertheimer



Koehler



Mill

Hume

Koffka

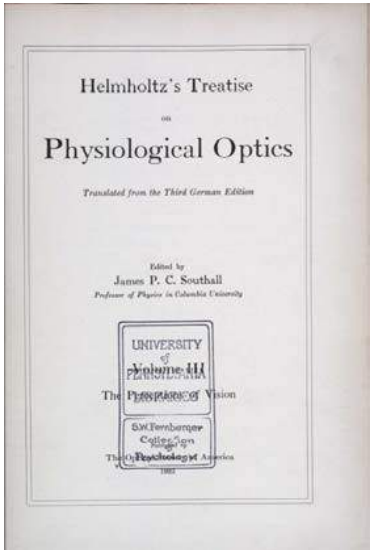
Kanizsa

Helmholtz

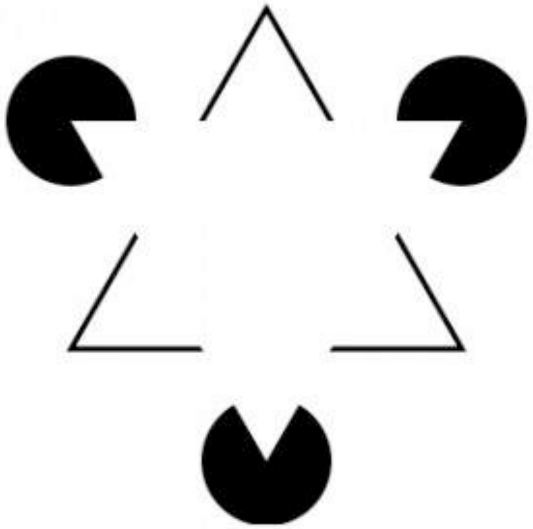
Gregory

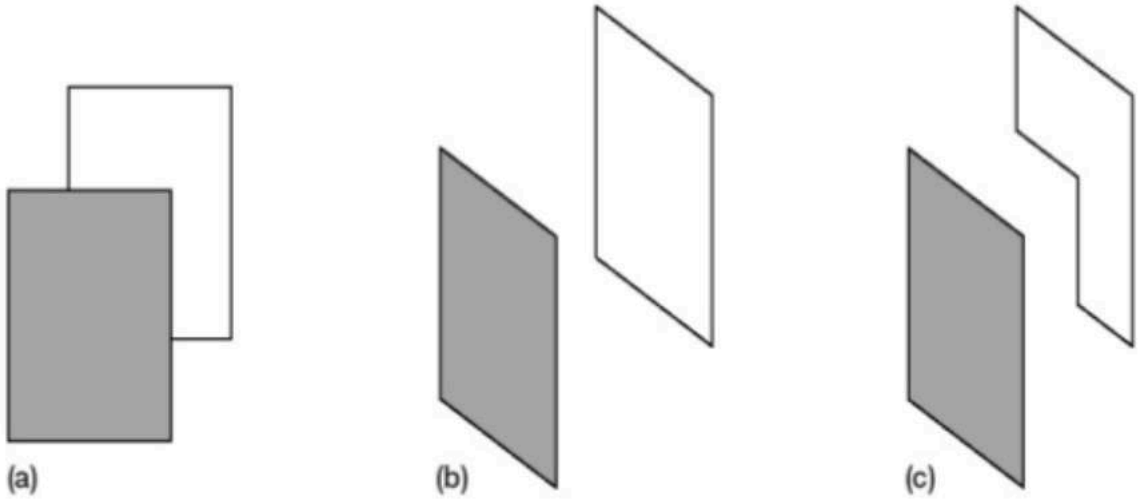
Helmholtz:
Vision as unconscious inference

The psychic activities that lead us to infer that there in front of us at a certain place there is a certain object of a certain character, are generally not conscious activities, but unconscious ones. In their result they are equivalent to a *conclusion*, to the extent that the observed action on our senses enables us to form an idea as to the possible cause of this action; although, as a matter of fact, it is invariably simply the nervous stimulations that are perceived directly, that is, the actions, but never the external objects themselves. But what seems to differentiate them from a conclusion, in the ordinary sense of that word, is that a conclusion is an act of conscious thought. An astronomer, for example, comes to real conscious conclusions of this sort, when he computes the positions of the stars in space, their distances, etc., from the perspective images he has had of them at various times and as they are seen from different parts of the orbit of the earth. His conclusions are based on a conscious knowledge of the laws of optics. In the ordinary acts of vision this knowledge of optics is lacking. Still it may be permissible to speak of the psychic acts of ordinary perception as *unconscious conclusions*, thereby making a distinction of some sort between them and the common so-called conscious conclusions. And while it is true that there has been, and probably always will be, a measure of doubt as to the similarity of the psychic activity in the two cases, there can be no doubt as to the similarity between the results of such unconscious conclusions and those of conscious conclusions.



Helmholtz:
Vision as unconscious inference





Gibson and the ecological approach

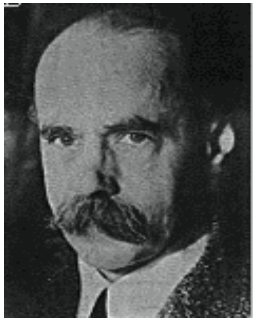
The belief of the empiricists that the perceived meanings and values of things are supplied from past the experience of the observer will not do.

But even worse is the belief of nativists that meanings and values are supplied from the past experience of the race by innate ideas.

J. J. Gibson, 1979

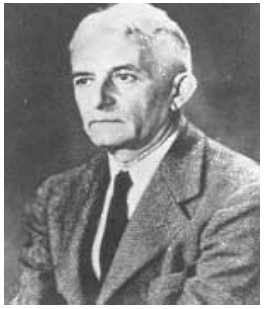
The Gestalt school

Wertheimer



Koehler

Koffka



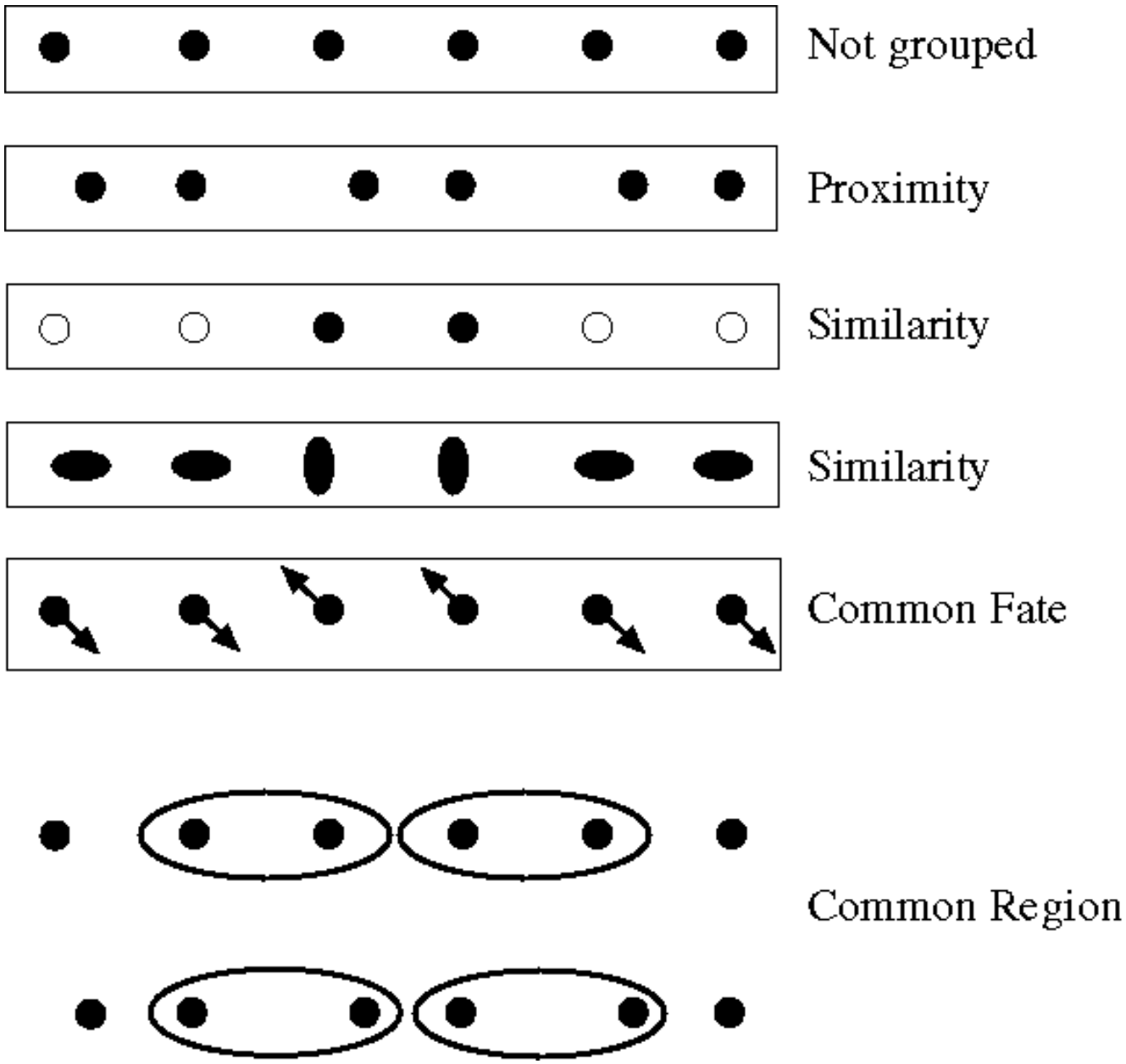
Gestalt properties

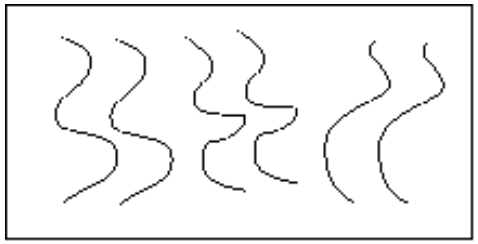
elements in a collection of elements can have properties that result from relationships

- Gestaltqualitat

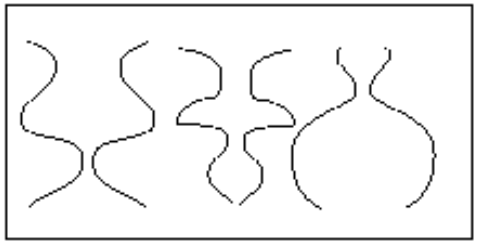
A series of factors affect whether elements should be grouped together

- Gestalt factors

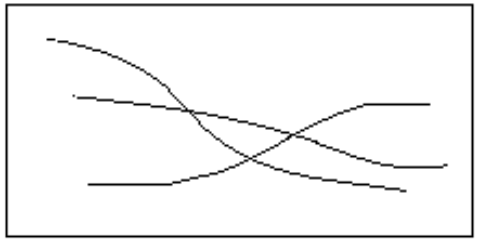

Not grouped
Proximity
Similarity
Similarity
Common Fate
Common Region



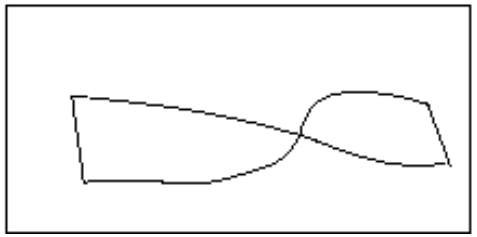
Parallelism



Symmetry



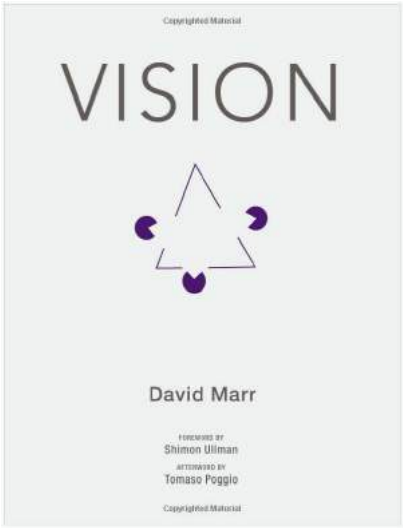
Continuity



Closure

David Marr and the computational approach

- **Computational level:** what does the system do (e.g.: what problems does it solve or overcome) and similarly, why does it do these things
- **Algorithmic/representational level:** how does the system do what it does, specifically, what representations does it use and what processes does it employ to build and manipulate the representations
- **Implementational/physical level:** how is the system physically realised (in the case of biological vision, what neural structures and neuronal activities implement the visual system)



A missing component?

I am not sure that Marr would agree, but I am tempted to add learning as the very top level of understanding, above the computational level.
[...]
Only then may we be able to build intelligent machines that could learn to see—and think—without the need to be programmed to do it.

Tomaso Poggio, 2010

Computer Vision

What is computer vision?

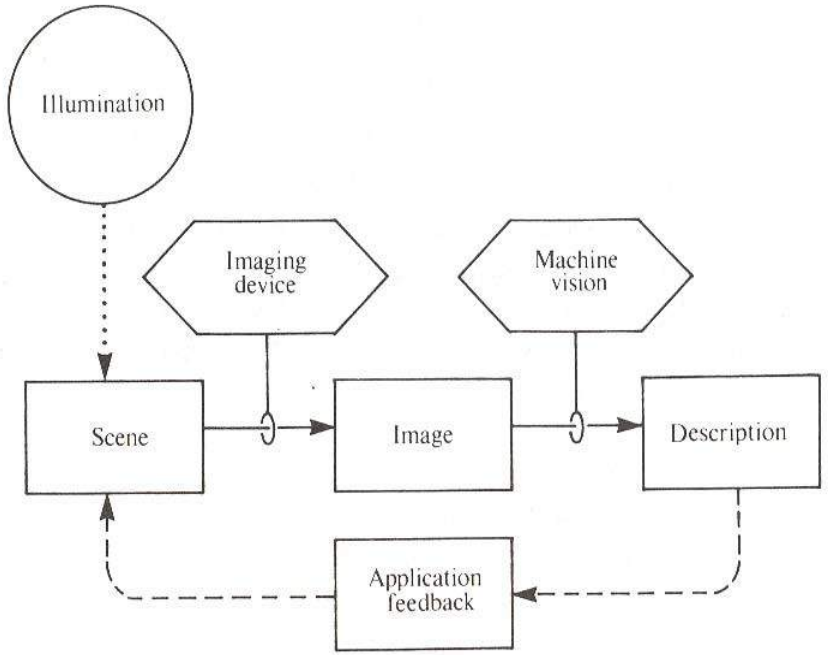
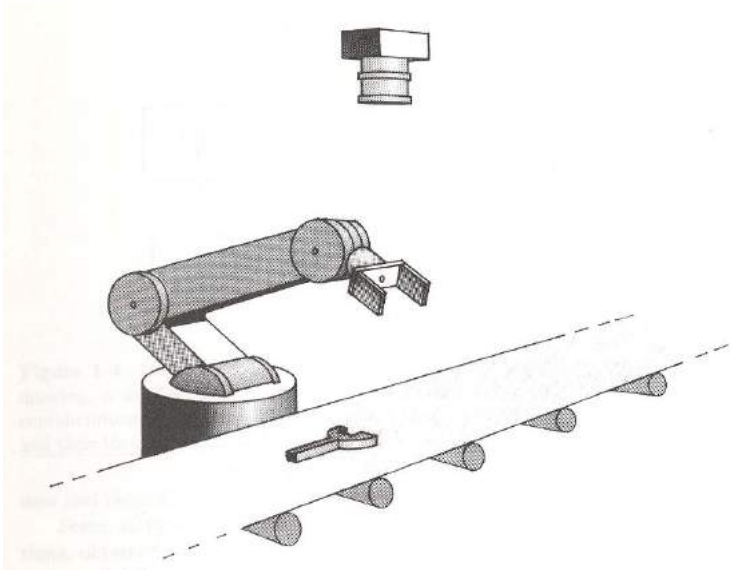


Figure 1-2. The purpose of a machine vision system is to produce a symbolic description of what is being imaged. This description may then be used to direct the interaction of a robotic system with its environment. In some sense, the vision system's task can be viewed as an inversion of the imaging process.



Related disciplines

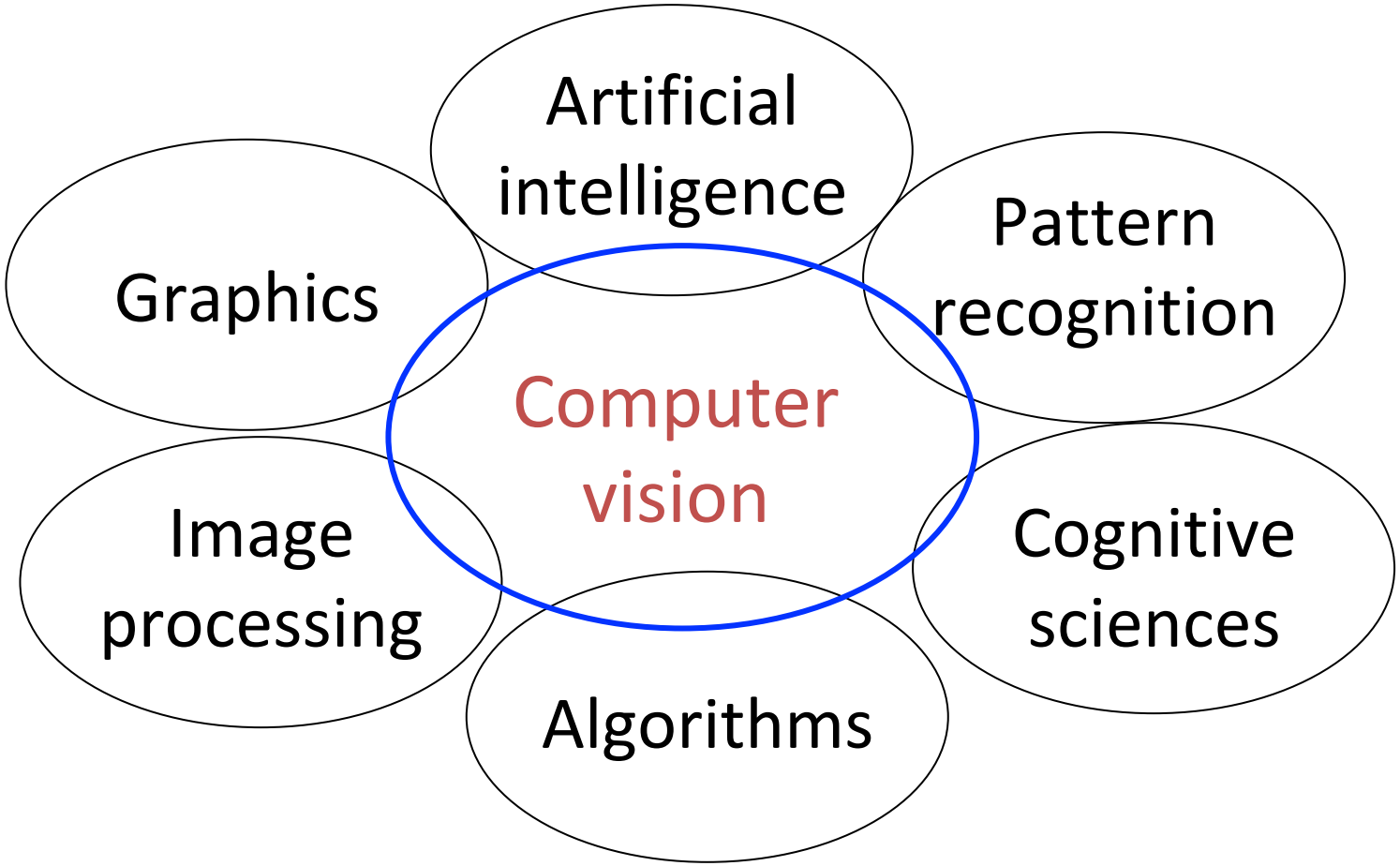
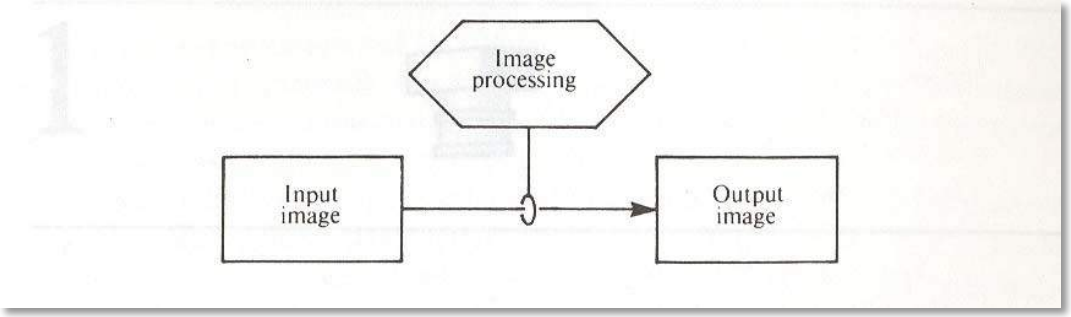
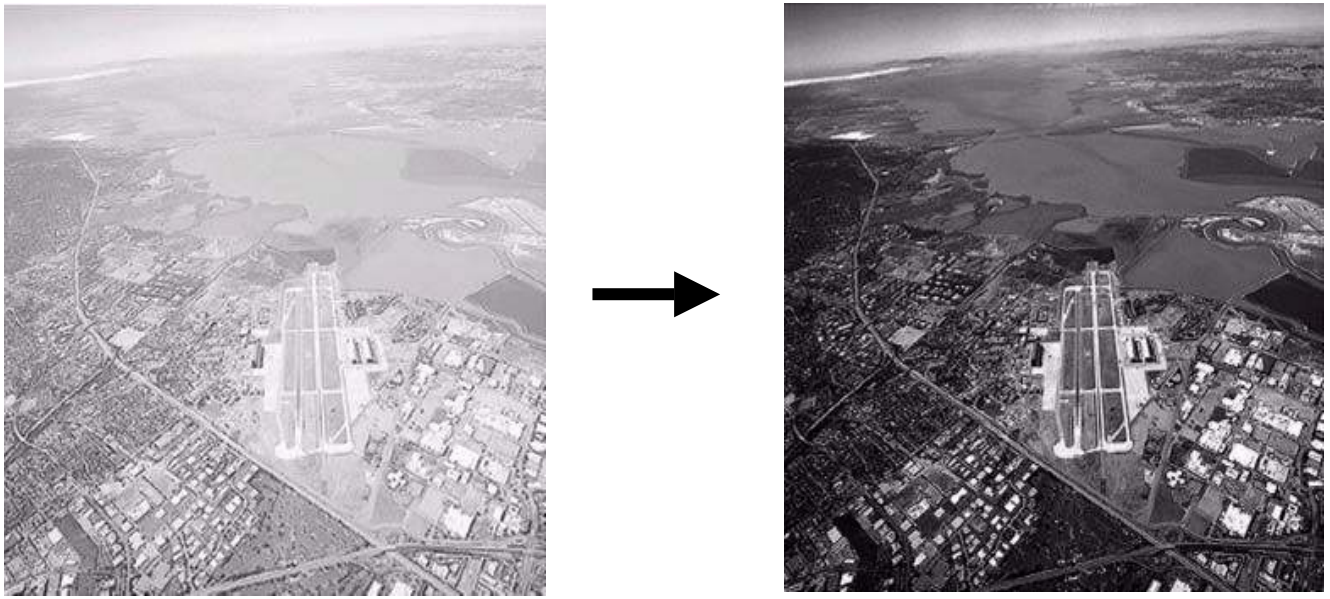
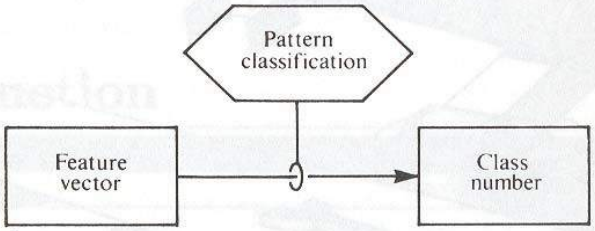


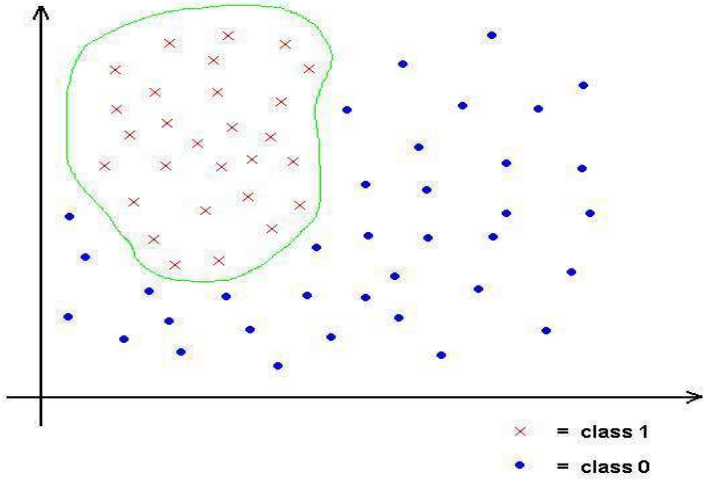
Image processing



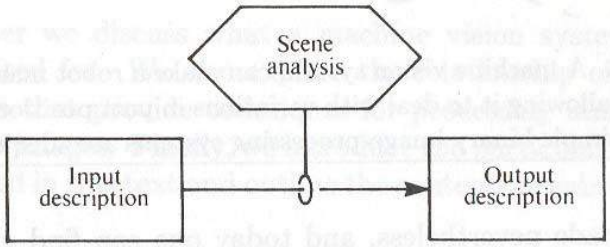


Pattern recognition





Scene analysis



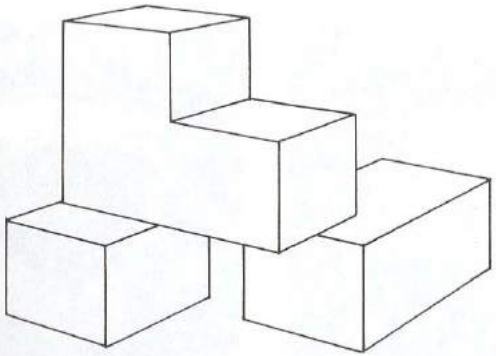
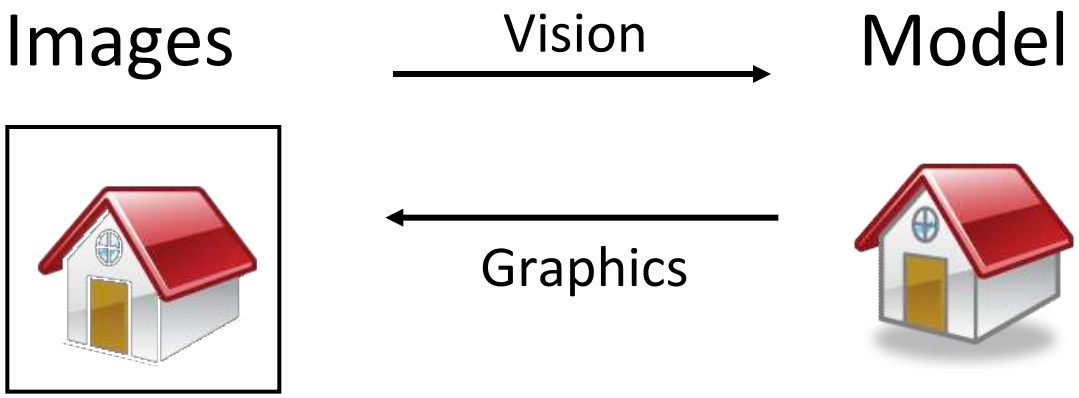


Figure 1-4. In scene analysis, a low-level symbolic description, such as a line drawing, is used to develop a high-level symbolic description. The result may contain information about the spatial relationships between objects, their shapes, and their identities.

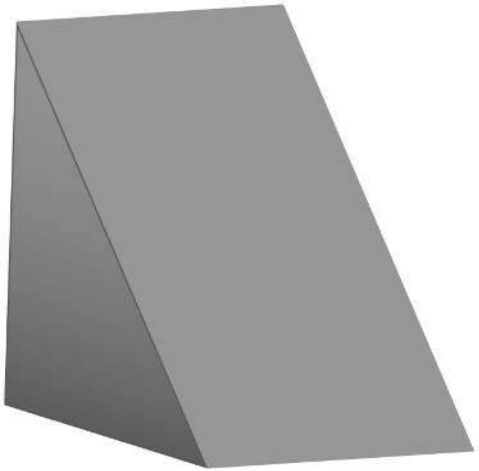
Vision and graphics



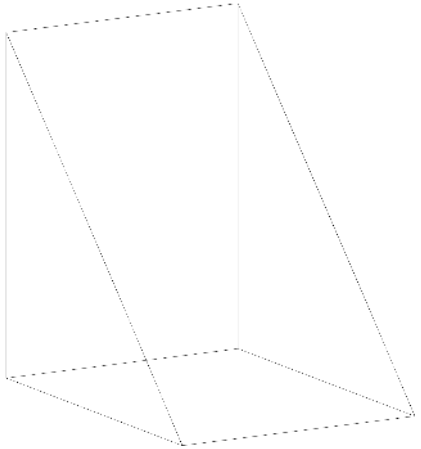
Inverse problems: analysis and synthesis.

From: Trevor Darrell

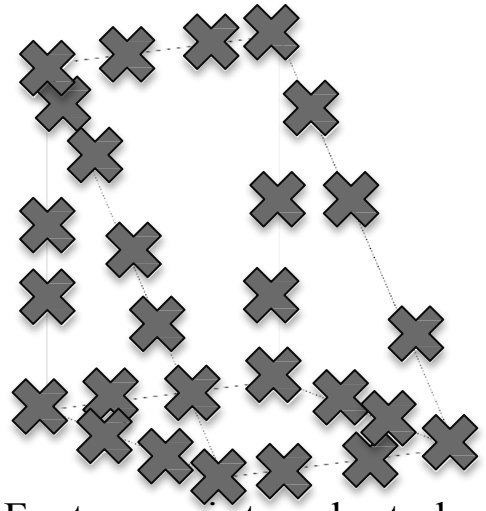
Block worlds



(a) Original picture



(b) Differentiated picture



(c) Feature points selected

Larry Roberts, 1963

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

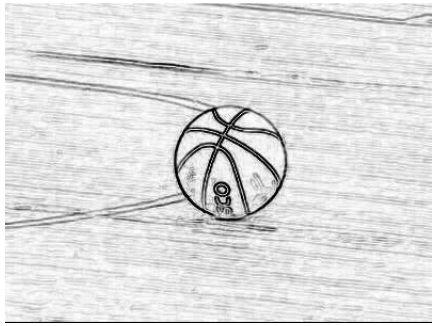
THE SUMMER VISION PROJECT

Seymour Papert

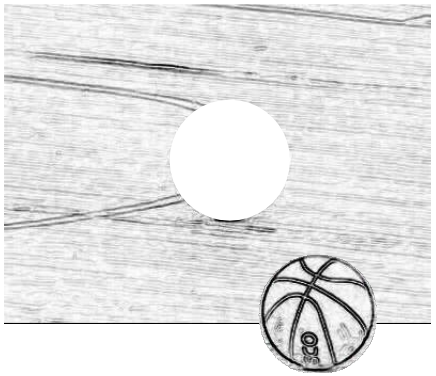
The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Input image

Edge image



2 ½-D sketch



3-D model

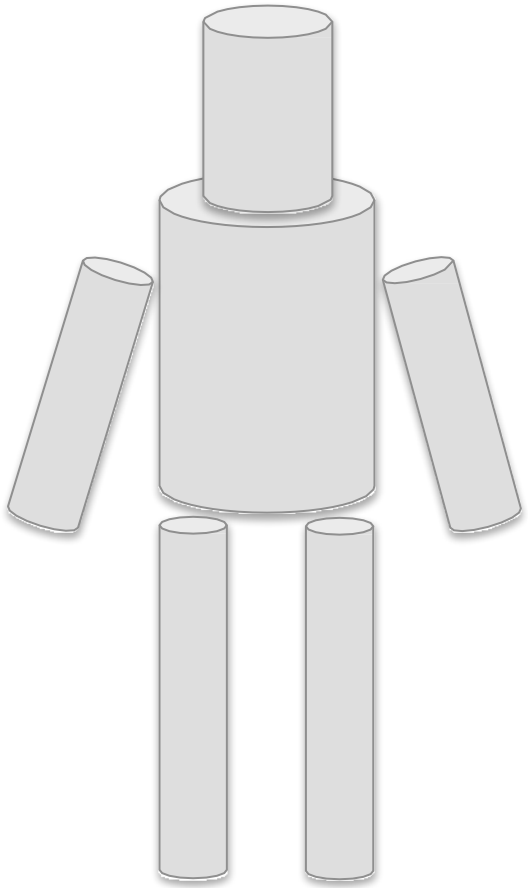
Input Image	Primal Sketch	2 ½-D Sketch	3-D Model Representation
Perceived intensities	Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries	Local surface orientation and discontinuities in depth and in surface orientation	3-D models hierarchically organized in terms of surface and volumetric primitives

Stages of Visual Representation, David Marr, 1970s

From: Fei-Fei Li & Justin Johnson & Serena Yeung

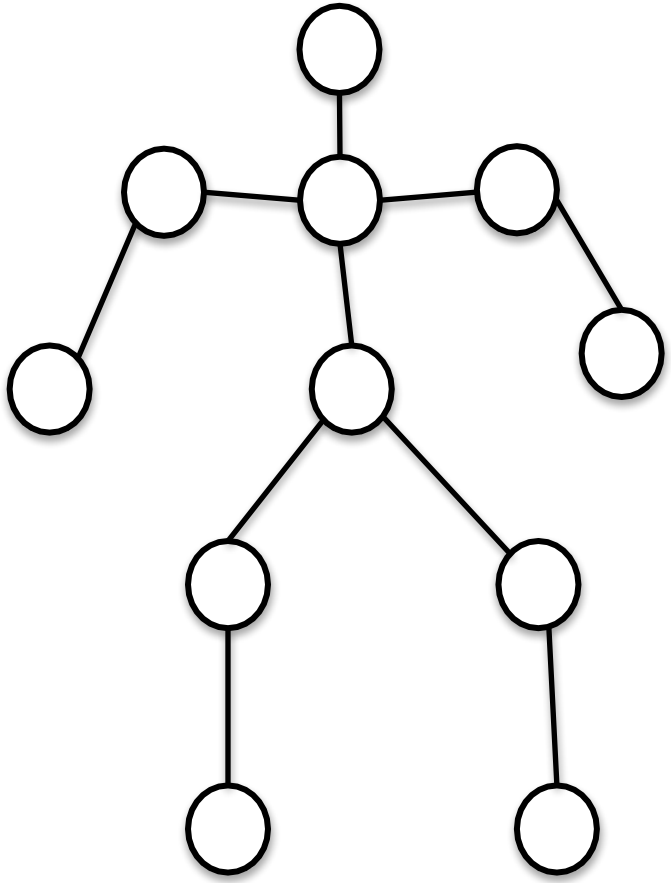
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

Fischler and Elschlager, 1973



David Lowe, 1987

Image segmentation and clustering

Image is CC BY 3.0

Image is public domain

Image is CC-BY 2.0; changes

From: Fei-Fei Li & Justin Johnson & Serena Yeung

Face Detection, Viola & Jones, 2001

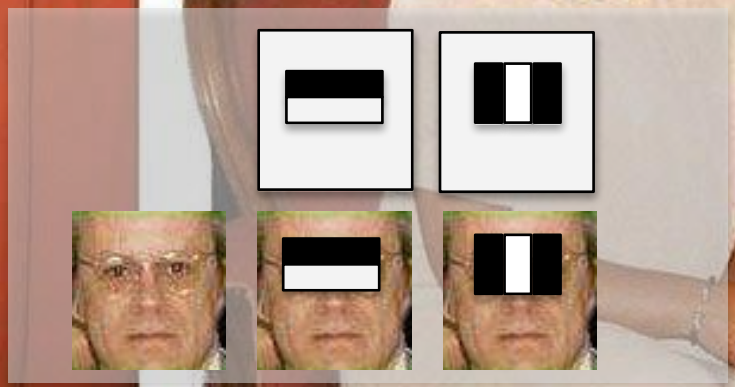
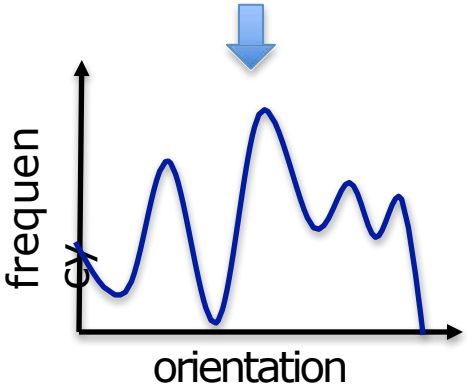


Image is public domain

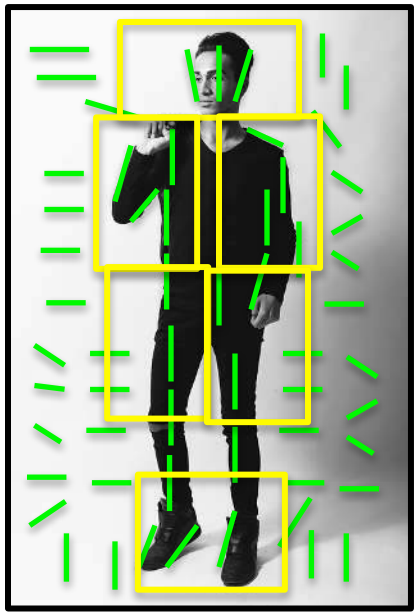
Image is public domain

Image is public domain

“SIFT” & Object Recognition, David Lowe, 1999



Histogram of Gradients (HoG)
Dalal & Triggs, 2005



Deformable Part Model
Felzenswalb, McAllester, Ramanan, 2009

PASCAL Visual Object Challenge (20 object categories)

Image is CC0 1.0 public domain

Image is CC0 1.0 public domain

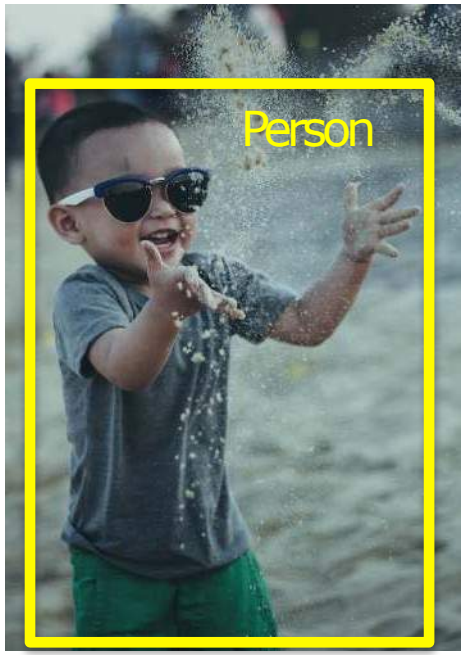
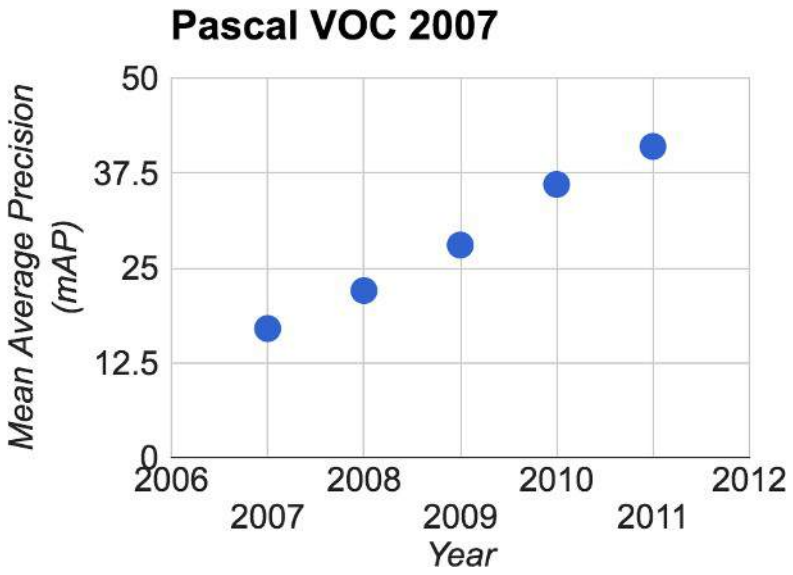


Image is CC0 1.0 public domain



[Everingham et al. 2006-2012]

IMAGENET

www.image-net.org

22K categories and **14M** images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

Russakovsky et al. IJCV 2015

The Age of “Deep Learning”

News & Analysis

Microsoft, Google Beat Humans at Image Recognition

Deep learning algorithms compete at ImageNet challenge

R. Colin Johnson

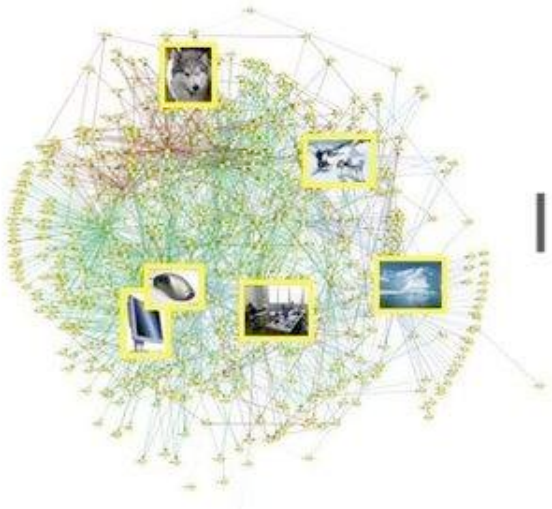
2/18/2015 08:15 AM EST

14 comments

NO RATINGS

1 saves

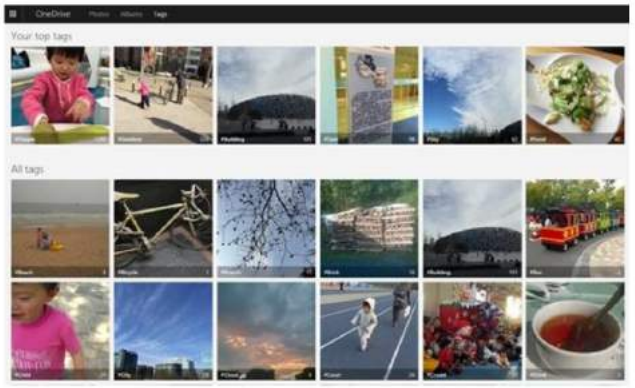
LOGIN TO RATE



IMAGENET

PORTLAND, Ore. -- First computers beat the best of us at chess, then poker, and finally Jeopardy. The next hurdle is image recognition — surely a computer can't do that as well as a human. Check that one off the list, too. Now Microsoft has programmed the first computer to beat the humans at image recognition.

The competition is fierce, with the ImageNet Large Scale Visual Recognition Challenge doing the judging for the 2015 championship on December 17. Between now and then expect to see a stream of papers claiming they have one-upped humans too. For instance, only 5 days after Microsoft announced it had beat the human benchmark of 5.1% errors with a 4.94% error grabbing neural network, Google announced it had one-upped Microsoft by 0.04%.



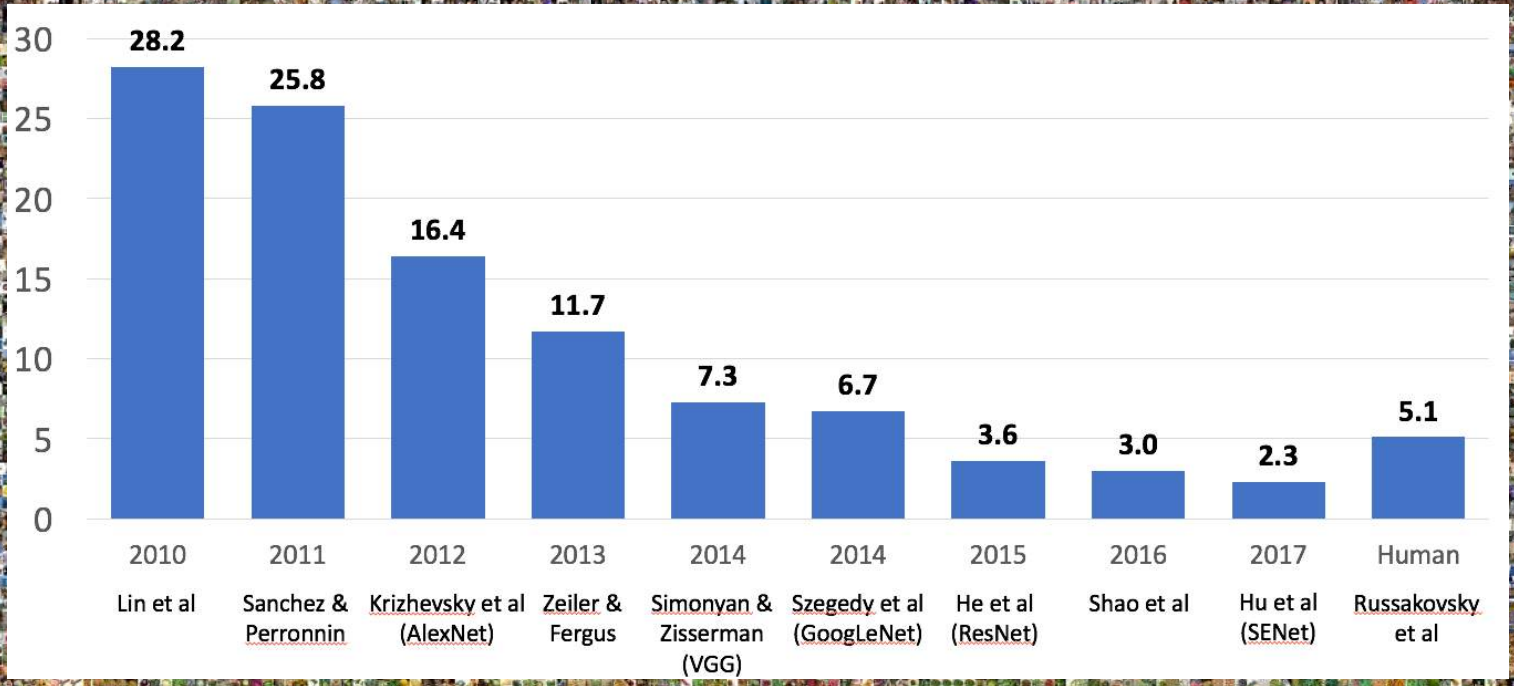
The top row is a representative of the categories that Microsoft's algorithm found in the database and the image columns below are examples that fit. (Source: Microsoft)

IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images



Russakovsky et al. IJCV 2015

The Deep Learning "Philosophy"

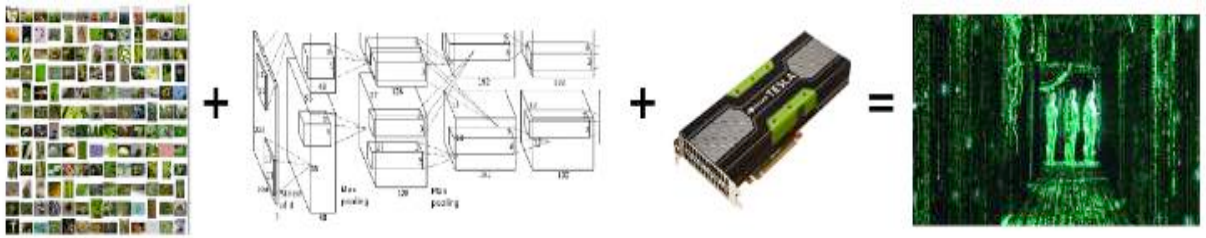
- Learn a feature hierarchy all the way from pixels to classifier
- Each layer extracts features from the output of previous layer
- Train all layers jointly

Old Idea… Why Now?

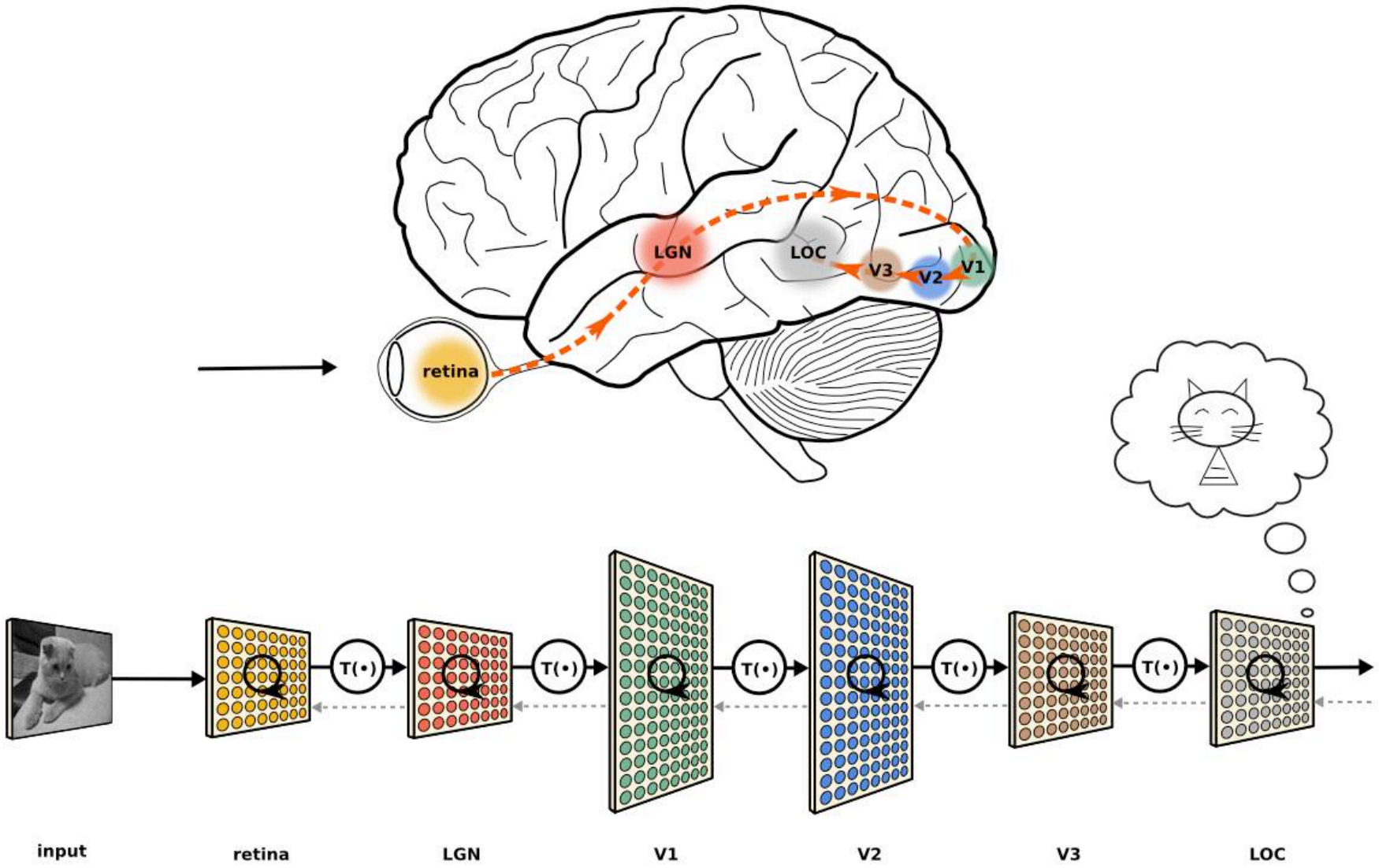
1. We have more data - from Lena to ImageNet.

2. We have more computing power, GPUs are really good at this.

3. Last but not least, we have new ideas

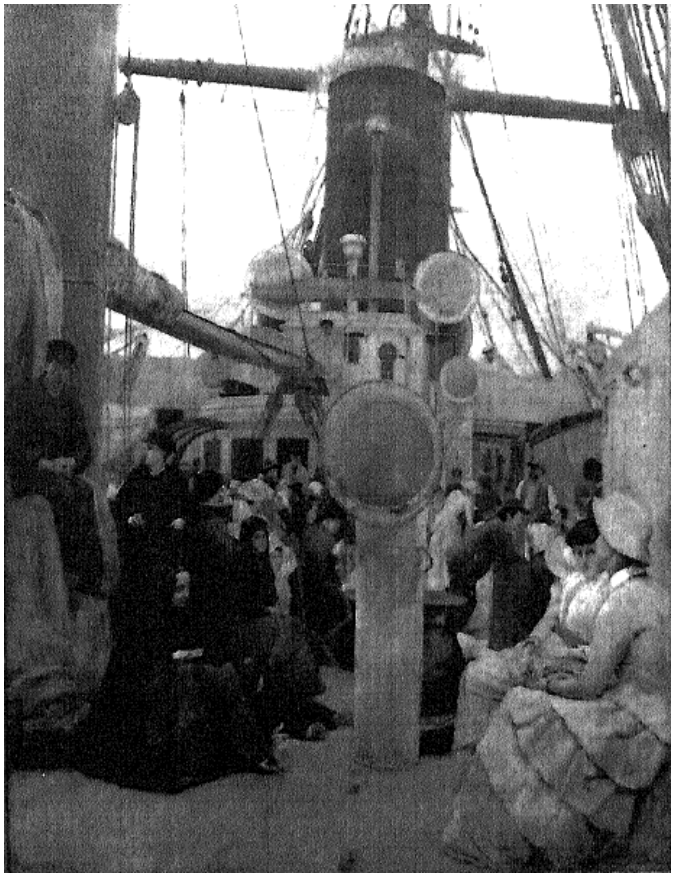


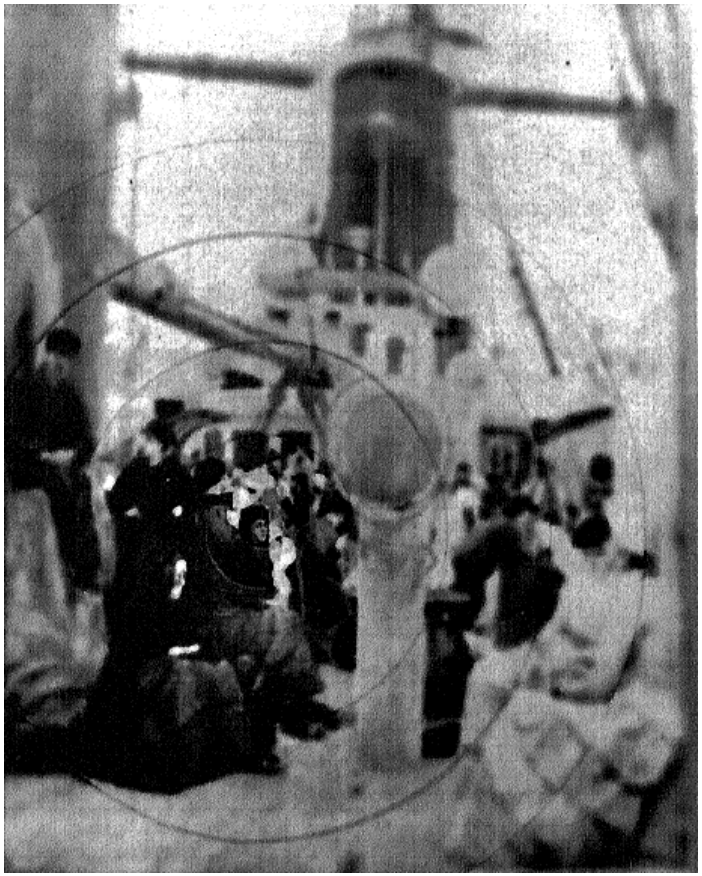
Inspiration from Biology



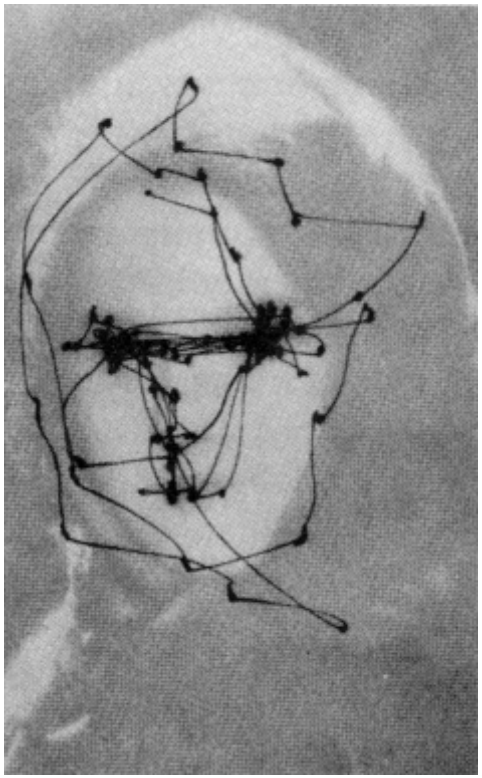
Imitating nature?

Foveal vision





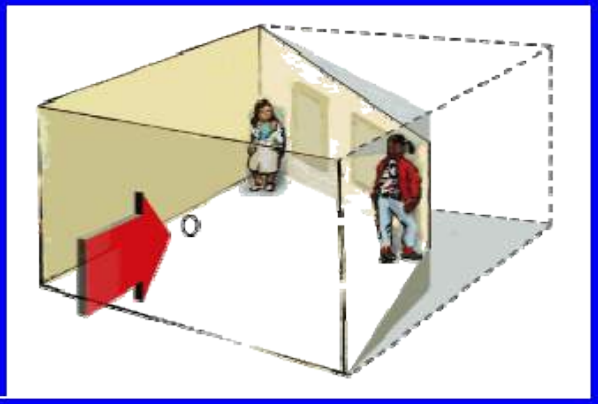
Attention and eye movement



Ambiguities, inconsistencies, illusions

The Ames room

The Ames Room



Necker's cube

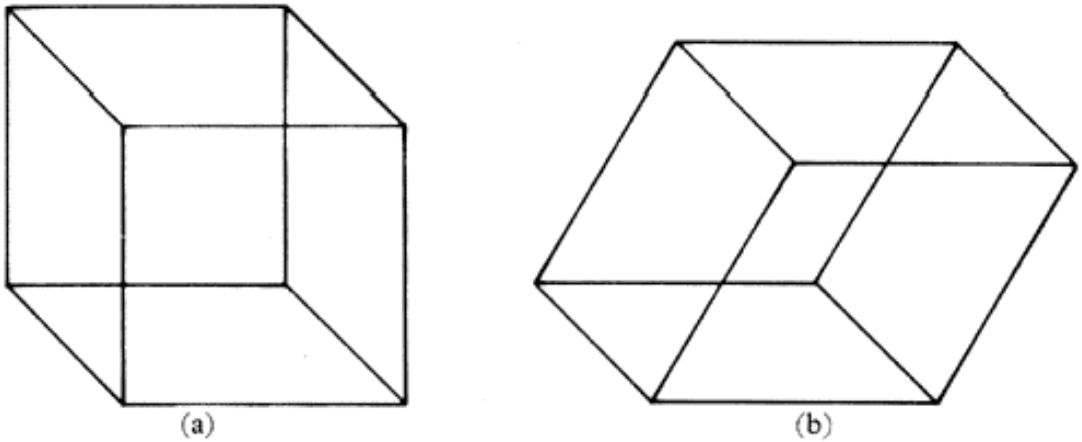
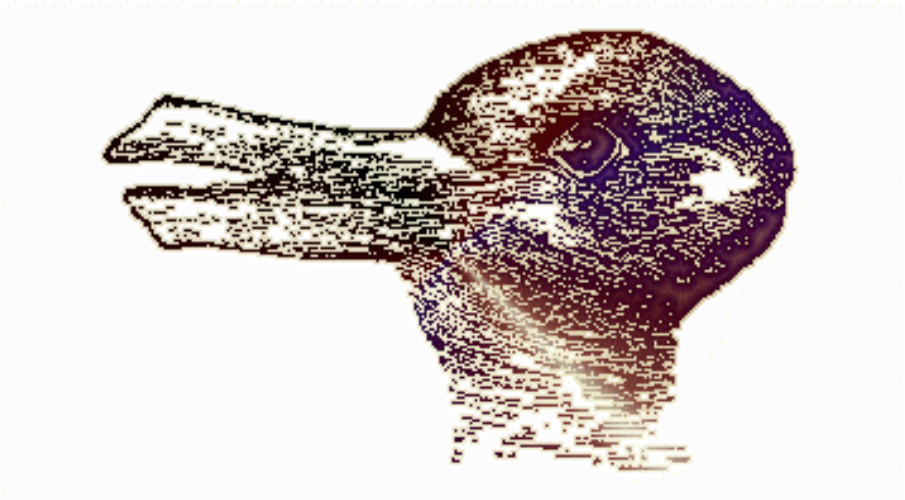


FIG. 2 (a) Necker cube. This is the most famous of many depth-ambiguous, figures. (When presented with no background it changes in shape with each reversal, the apparent back being larger than the apparent front face.) (b) Necker rhomboid. This is the original form, presented by L. A. Necker in 1832.

Bistable perceptions



Mueller-Lyer

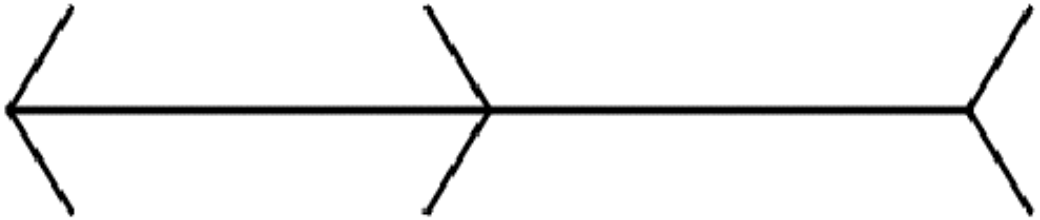


FIG. 5 Müller-Lyer arrows figure 1889. The most famous illusion: the outward-going 'arrow heads' produce expansion of the 'shaft' and the inward-going heads contraction.

Ponzo

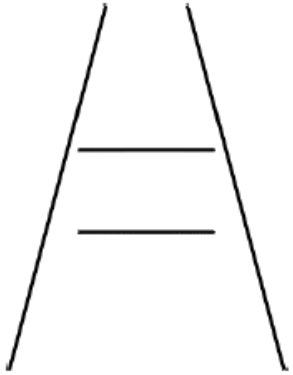
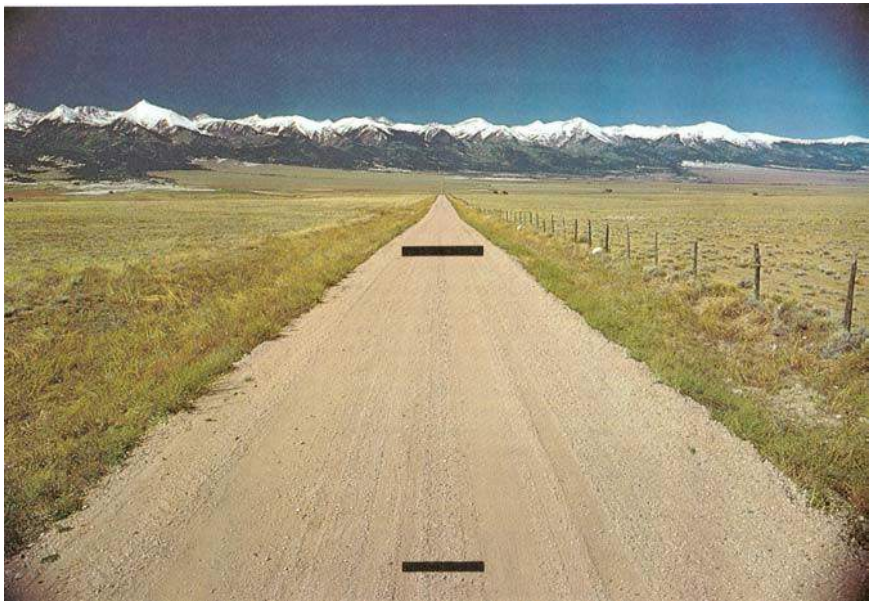
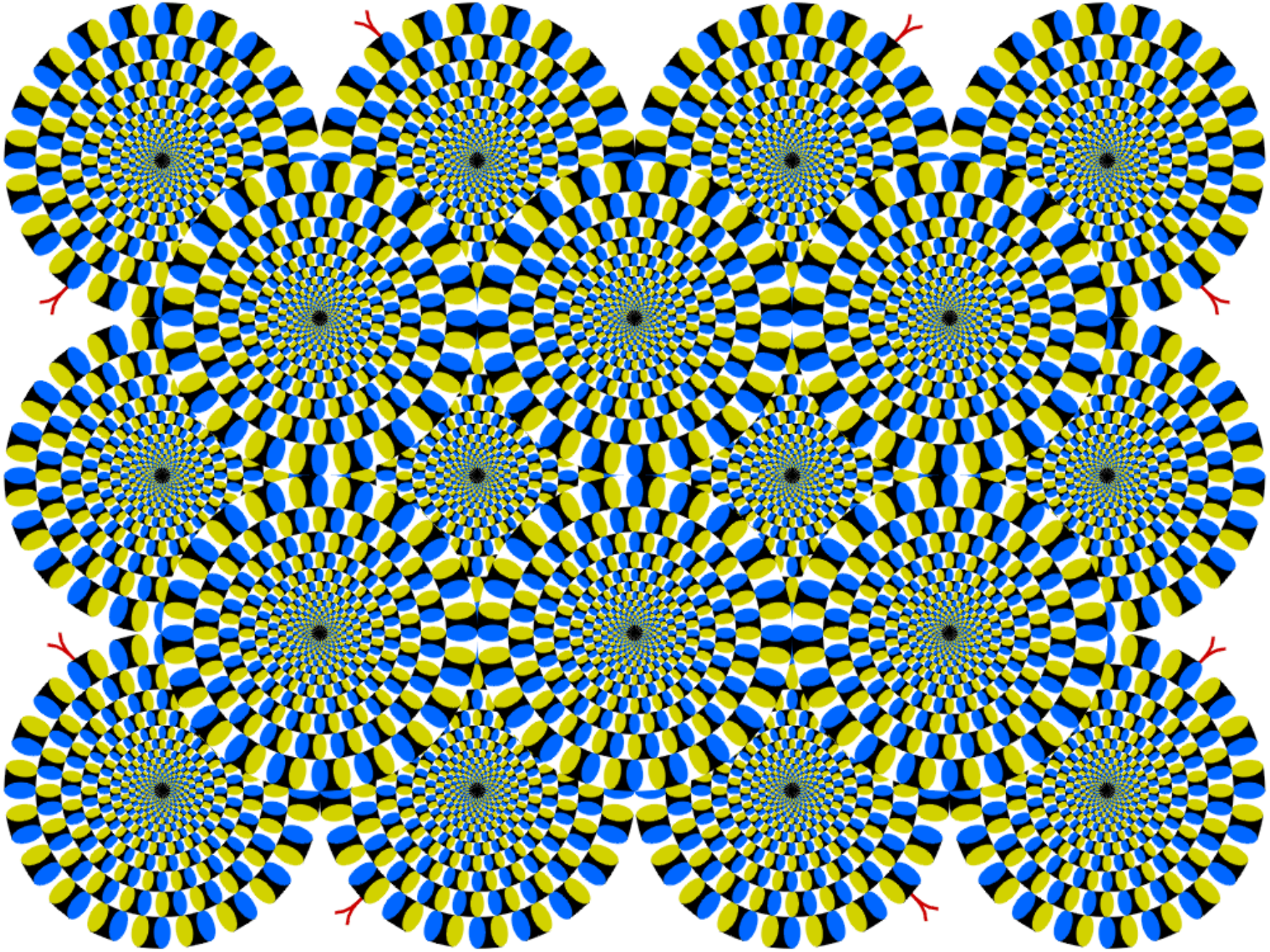
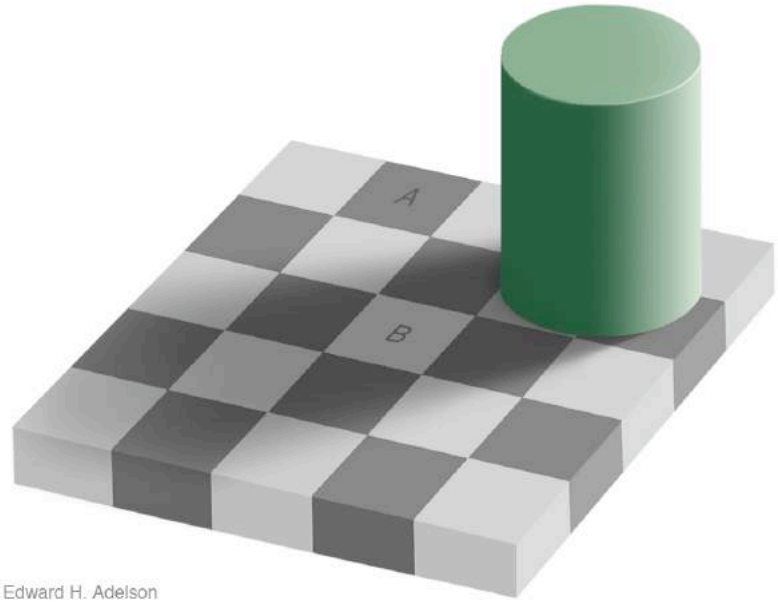


FIG. 6 Ponzo figure. The upper of the parallel lines is expanded with respect to the lower.

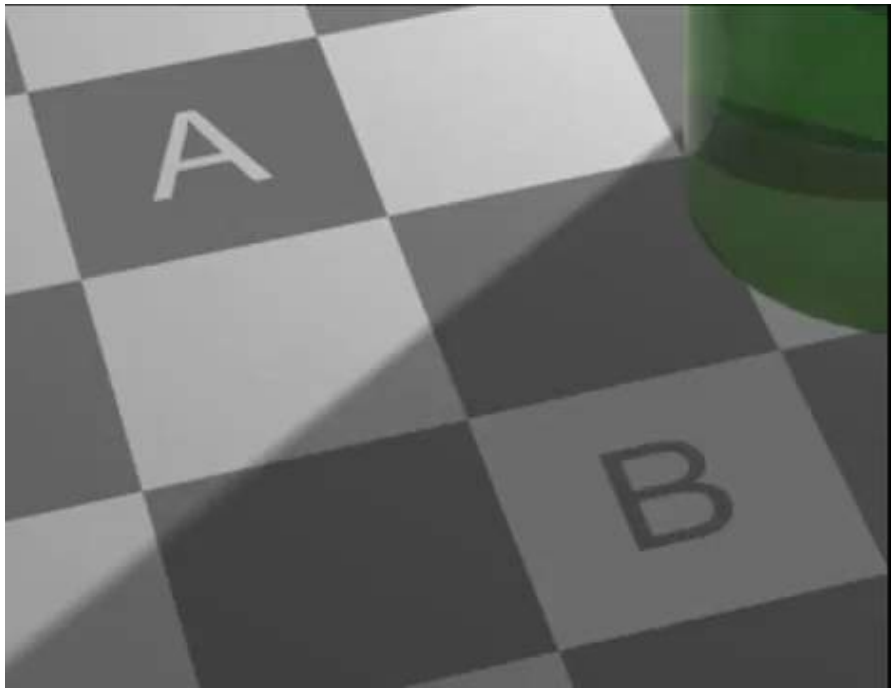




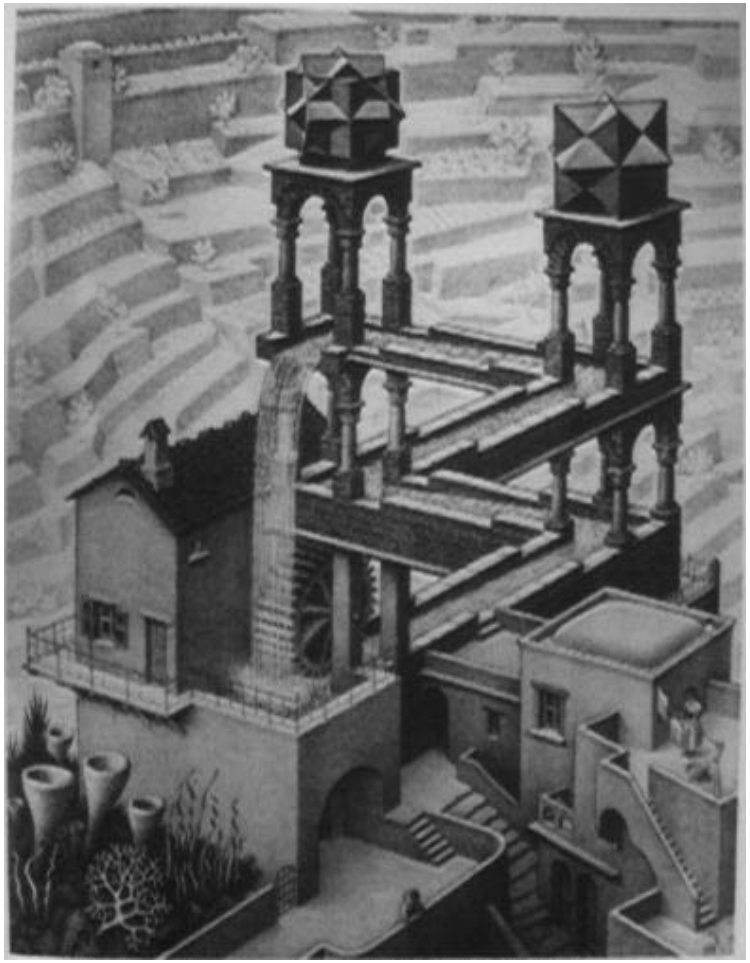
Adelson



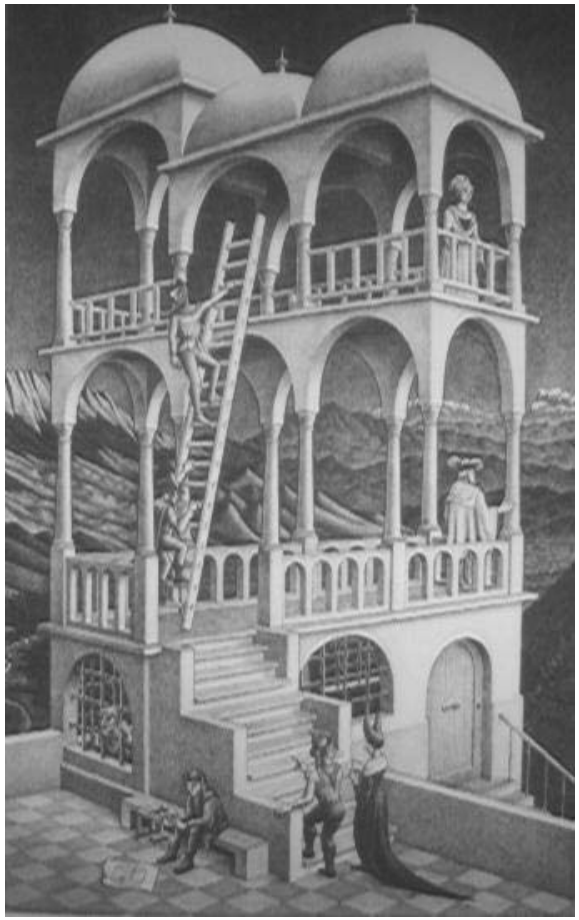
Edward H. Adelson

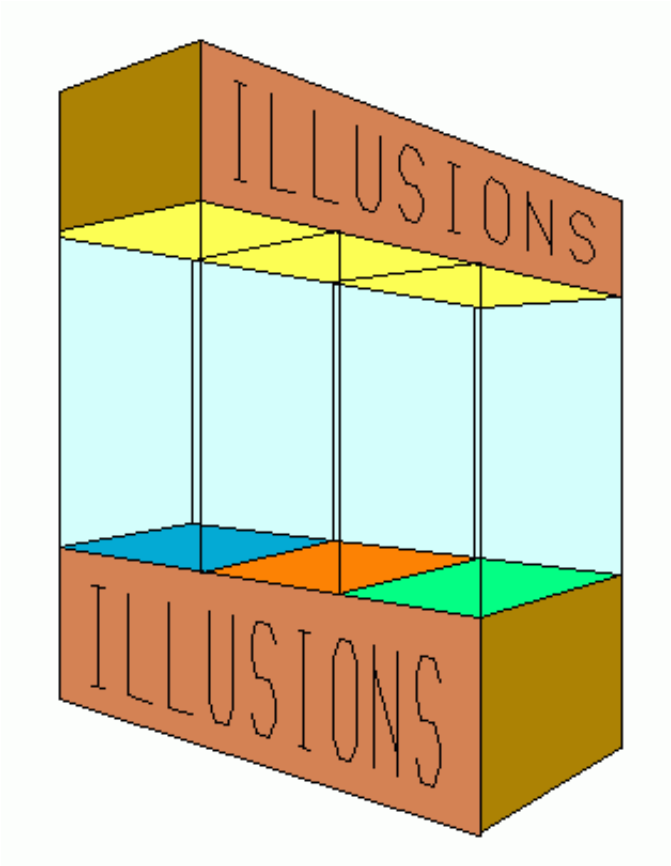


Inconsistencies



Inconsistencies





Inconsistencies

Top-down *vs.* bottom-up

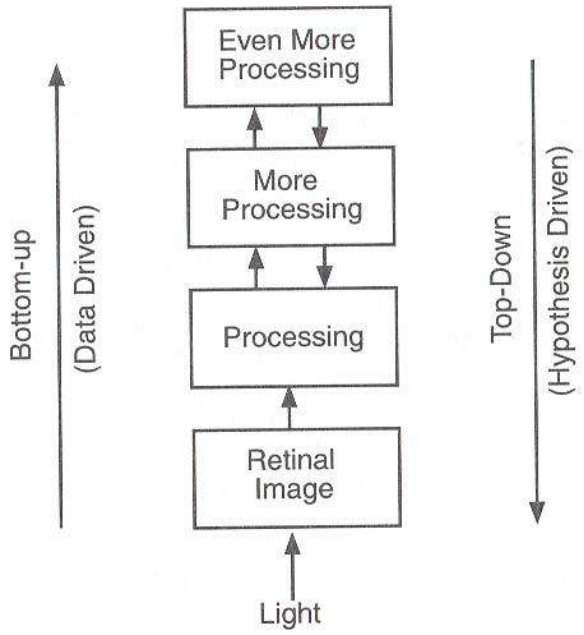
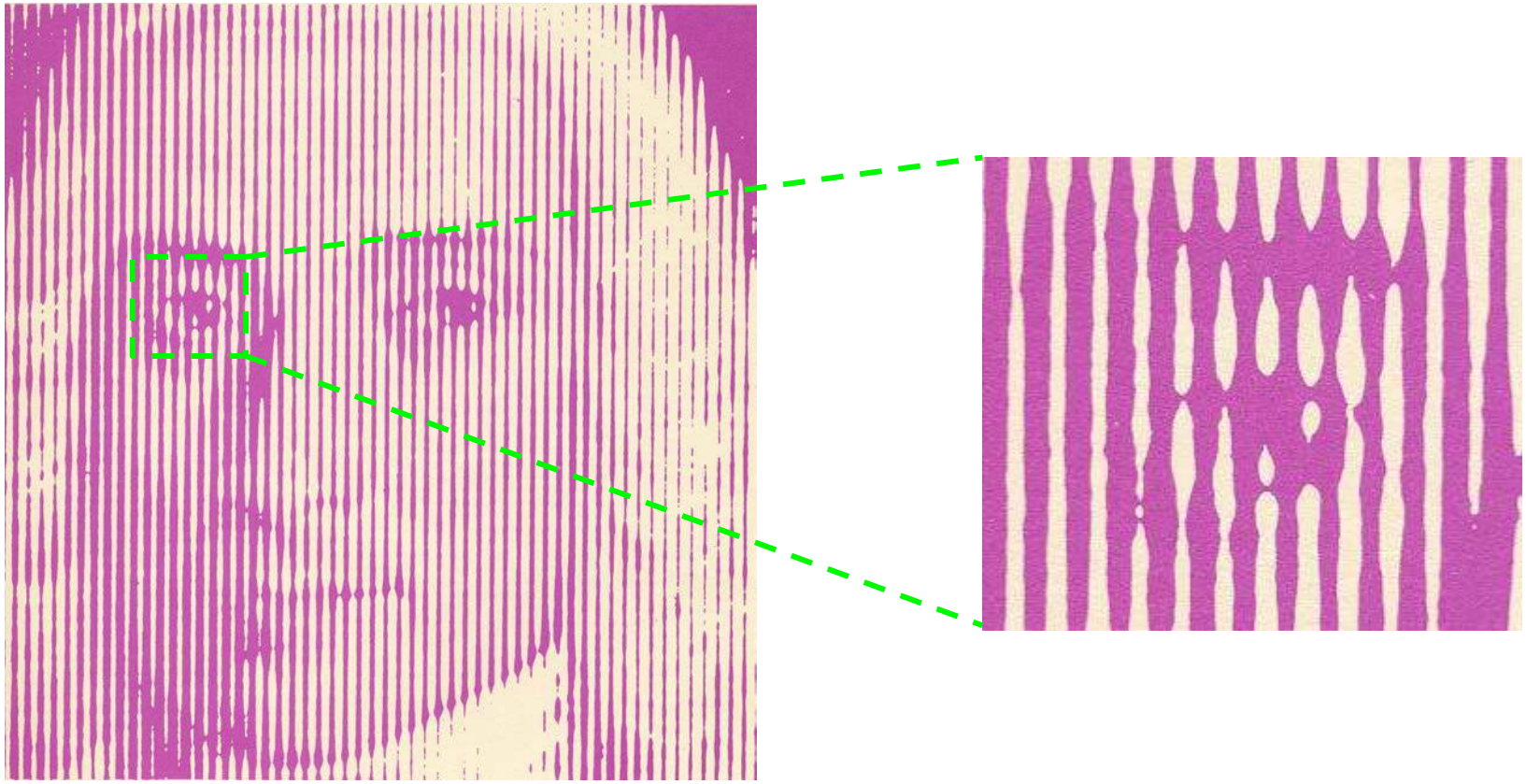


Figure 2.3.11 Bottom-up versus top-down processing. The two directions of processing are referred to as *bottom-up* (or *data driven*) from lower to higher levels of processing and *top-down* (or *hypothesis driven*) from higher to lower levels of processing.

Top-down or bottom-up?

Top-down or bottom-up?

Top-down or bottom-up?



Charlie Chaplin's mask

Figure 1. Photographs of a rotated hollow mask: (a) and (b) (black hat) show the front and side truly convex view; (d) (white hat) shows the inside of the mask; it appears convex although it is truly hollow; (c) is curiously confusing as part of the hollow inside is seen as convex, combined with the truly convex face. This is even more striking with the actual rotating mask. Viewing the hollow mask with both eyes it appeal's convex, until viewed from as close as a metre or so. Top-down knowledge of faces is pitted against bottom-up signalled information. The face reverses each time a critical viewing distance is passed, as 'downwards' knowledge or 'upwards' signals win. (This allows comparison of signals against knowledge by nulling.)

Course topics

Detection and Recognition

- Face detection and recognition
- Pedestrian detection
- Instance recognition
- Category recognition
- Applications

Video understanding

- Tracking
- Person re-identification
- Action recogniton
- Applications

Approach: we'll cover both classical and the latest state-of-the-art methods (= deep learning)

Related courses

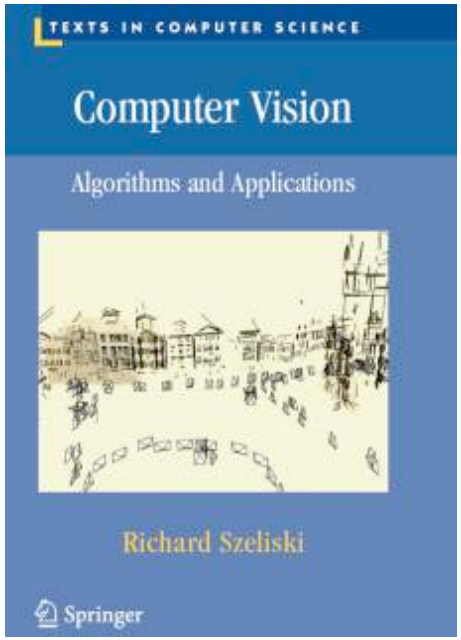
- **Artificial intelligence (CM0472, CM0492)**
- **Computer vision (CM0193)**

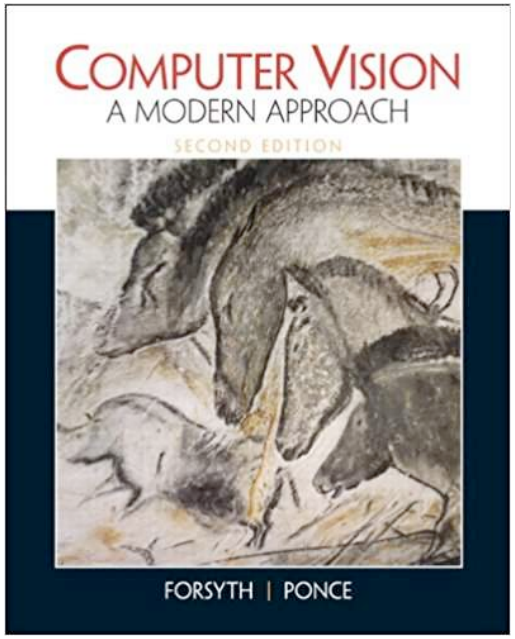
Some intersections, *but can be taken independently*.

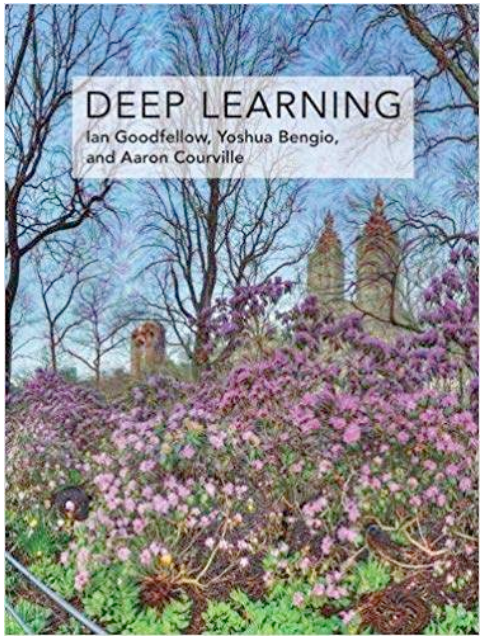
Exam

- **Oral**
- **Small project** (to be agreed together), possibly related to thesis or AI project

Suggested readings







Optional

- Slides of course
- Additional material (recent papers)

All material available in my homepage (follow “teaching”)