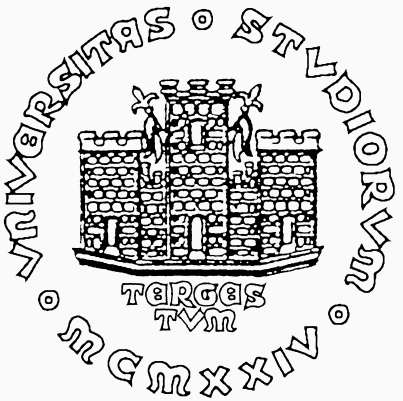
Computer Vision and Pattern Recognition

Course ID: 554SM – Fall 2018

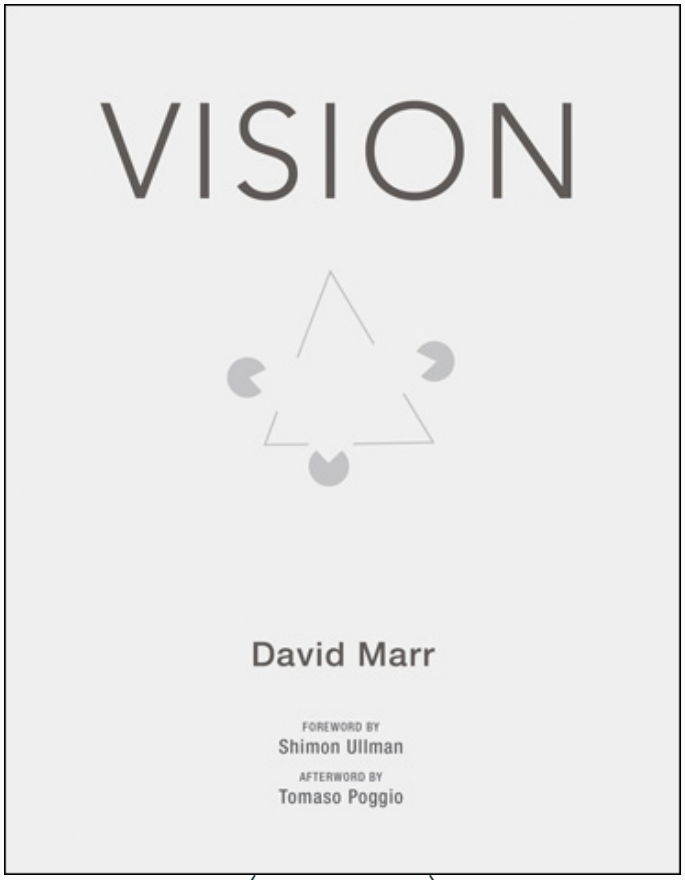
Felice Andrea Pellegrino

University of Trieste
Department of Engineering and Architecture



554SM –Fall 2018
Lecture 1: Introduction to Computer Vision

What and where



(Marr, 1982)

What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking. In other words, vision is the process *of discovering from images what is present in the world, and where it is.*

(David Marr)

Computer Vision

Short definition

Computer vision is the science of getting the machines “see.”

According to The British Machine Vision Association and Society for Pattern Recognition:

Long definition

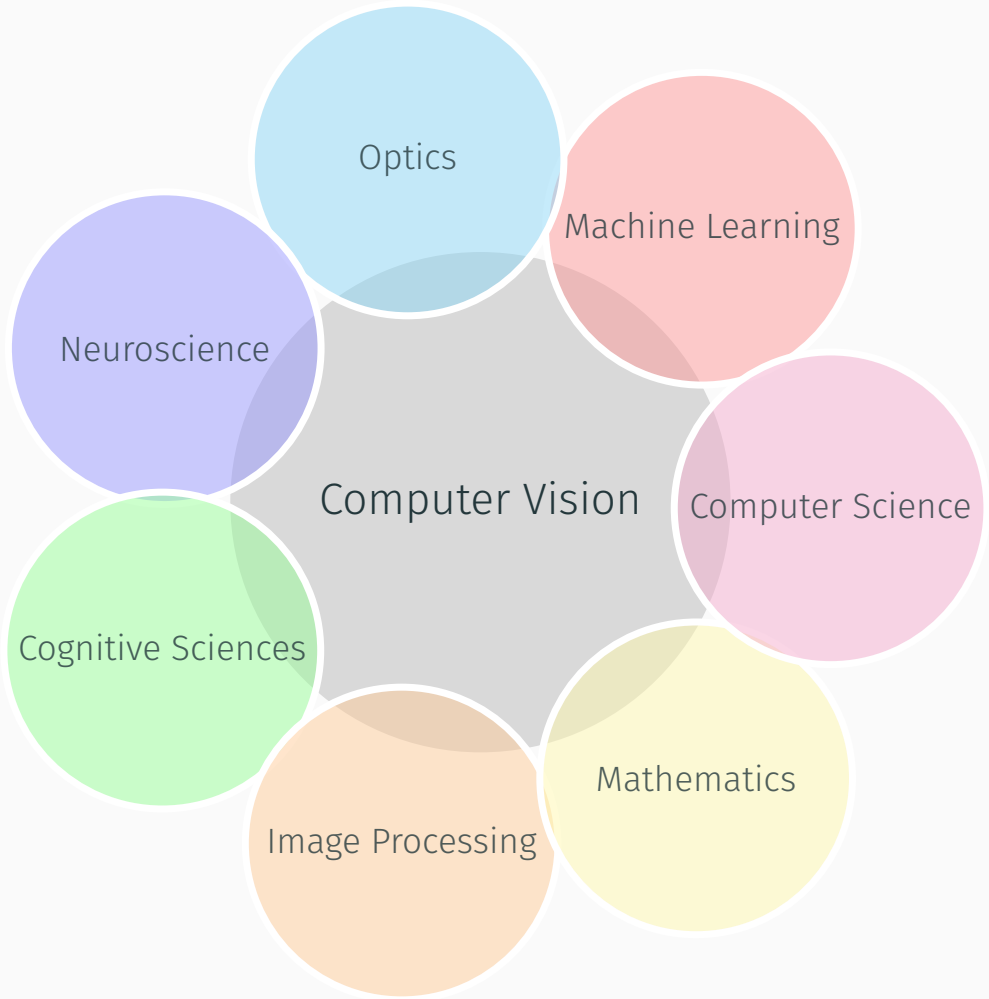
Computer vision is concerned with the automatic extraction, analysis and understanding of useful information from a single image or a sequence of images. It involves the development of a theoretical and algorithmic basis to achieve automatic visual understanding.

Computer Vision (cont.)

Other definitions:

- (Trucco and Verri, 1998): Computing properties of the 3-D world from one or more digital images.
- (Stockman and Shapiro, 2001): To make useful decisions about real physical objects and scenes based on sensed images.
- (Ballard and Brown, 1982): The construction of explicit, meaningful description of physical objects from images.
- (Forsyth and Ponce, 2012): Extracting descriptions of the world from pictures or sequences of pictures.

What is related to?



What and where

What kind of information can we extract from an image?

- Semantic information (“what”);
- metric 3D information (“where”).

What

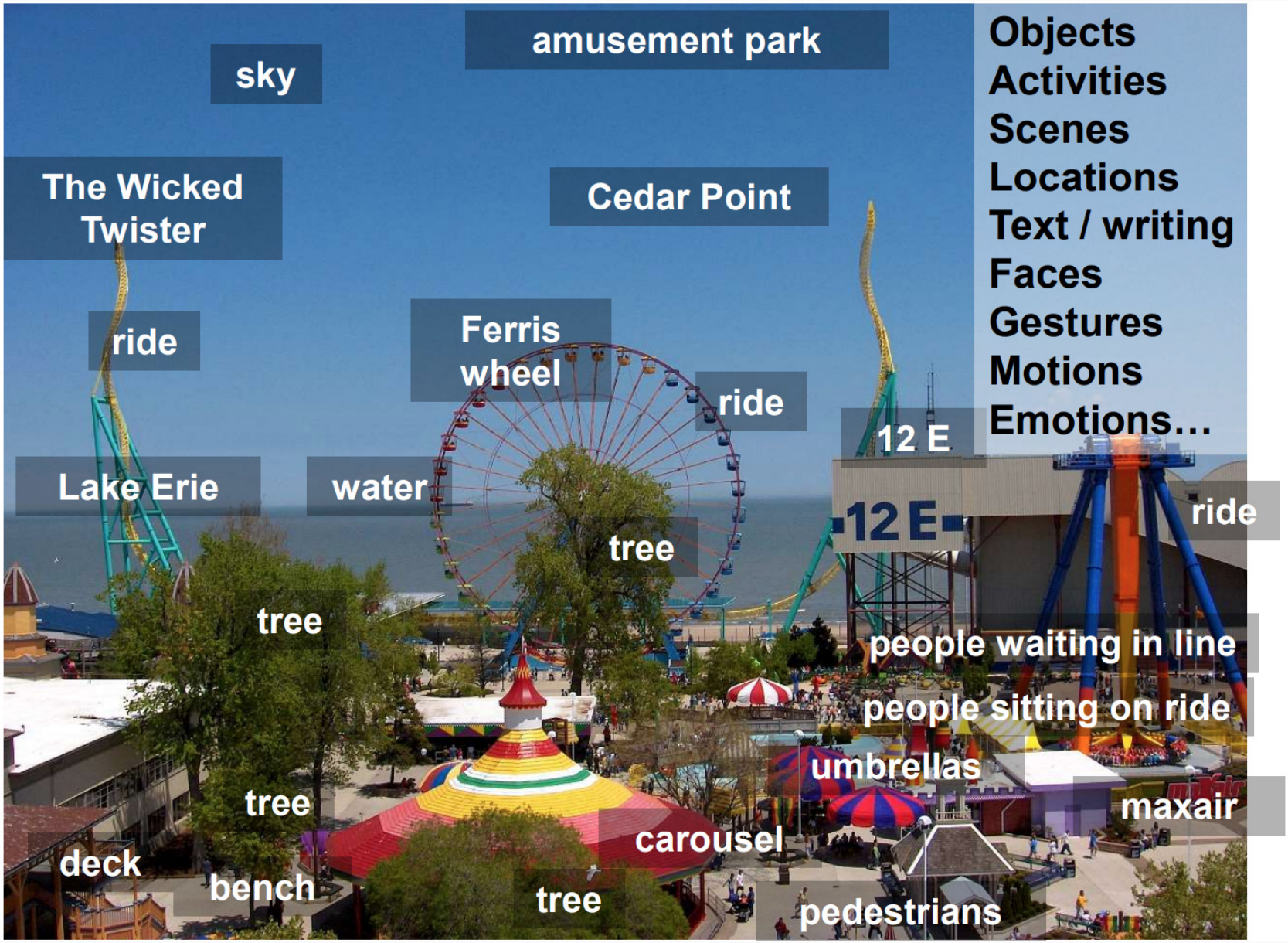
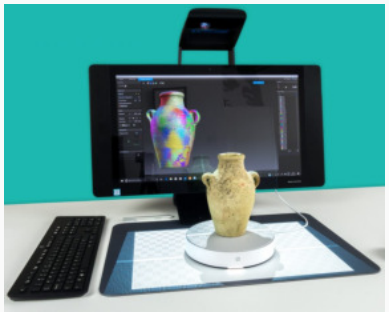
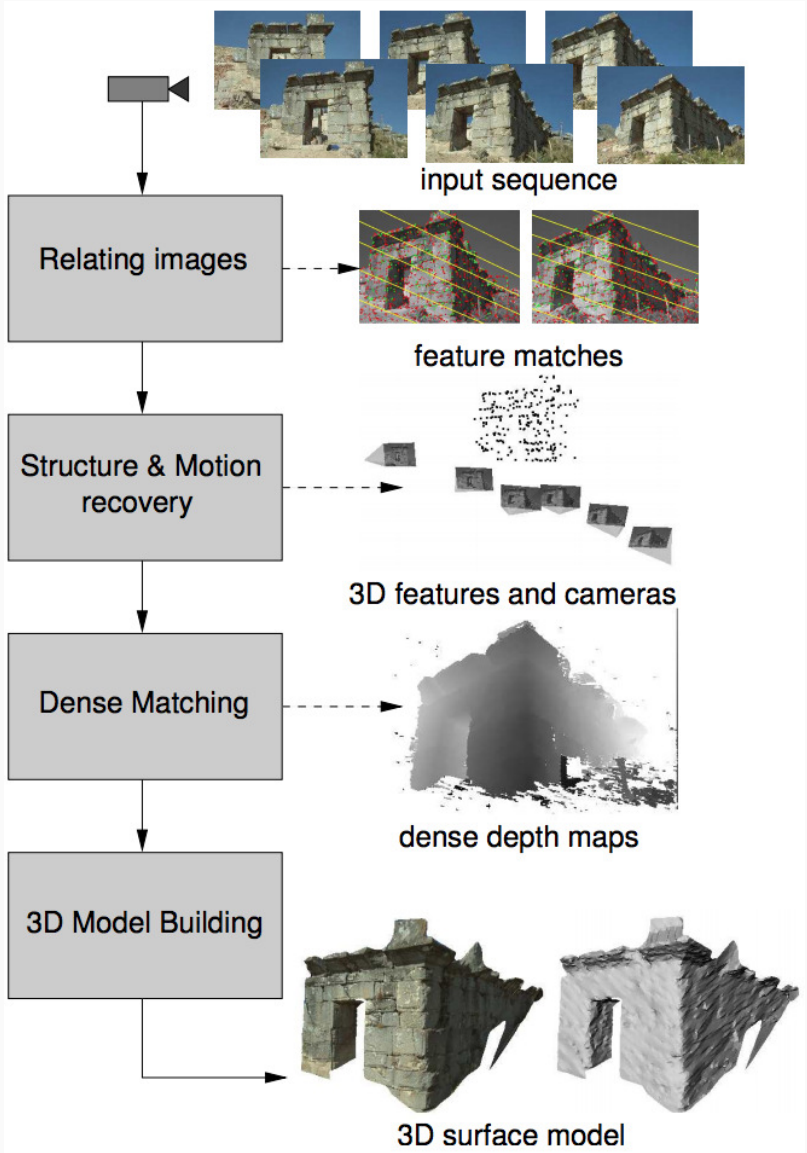


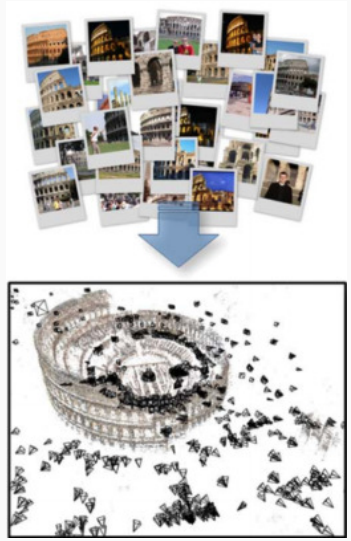
image credit: Kristen Grauman

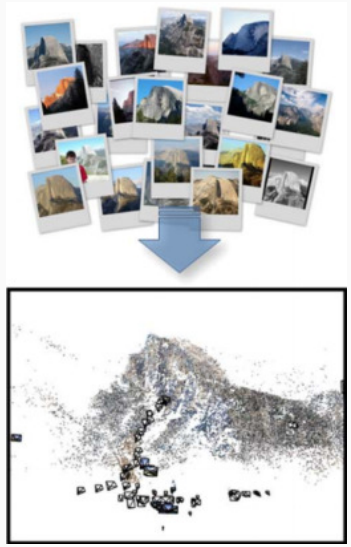
Where





(Pollefeys et al., 2002)





(Snavely et al., 2010)

“What” is difficult

Giuseppe Arcimboldo, *Vertumnus*, 1590.

What is in this image?

- Fruits and vegetables?
- A portrait of a man?
- A painting of Giovanni Arcimboldo?

Many levels of interpretation.

What is correct depends on the specific task.

The semantic gap

What we see.

97	24	144	34	91	125
208	153	238	185	255	225
136	121	178	28	57	90
90	178	149	30	167	115
240	179	209	164	155	247
224	163	225	84	99	11
141	9	253	167	36	249
159	18	0	192	6	48
150	82	222	149	108	171
53	136	157	189	47	150
77	168	253	60	186	173
121	104	135	188	95	92
59	210	123	248	215	159
216	184	205	222	188	208
50	248	58	22	146	5
58	136	128	94	45	21
44	83	231	95	245	250
58	27	147	175	68	167
112	156	216	153	237	59
80	199	189	202	57	103
236	108	150	94	96	31
110	23	63	53	22	69
47	68	171	22	164	66
232	39	21	198	46	85
251	72	160	53	12	39
112	113	169	99	185	89
28	135	187	141	89	31
66	117	228	59	169	226

What the computer sees.

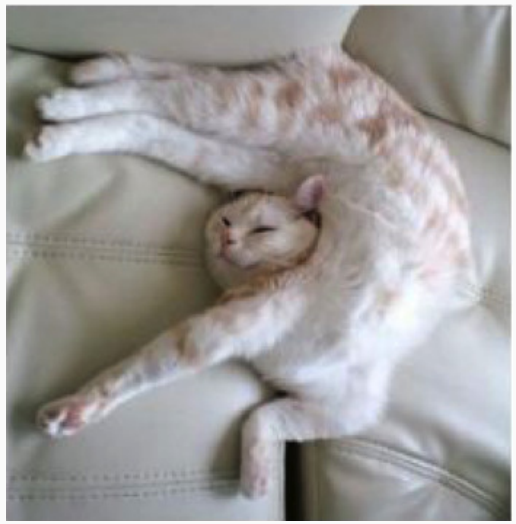
Images are sensed and therefore represented in computers as arrays of numbers.
Such low-level representation is far from semantics.

One of the goals of Computer Vision: bridge the gap between pixels and “meaning.”

“What” task: some challenges

Illumination

Occlusion



Deformation

Background

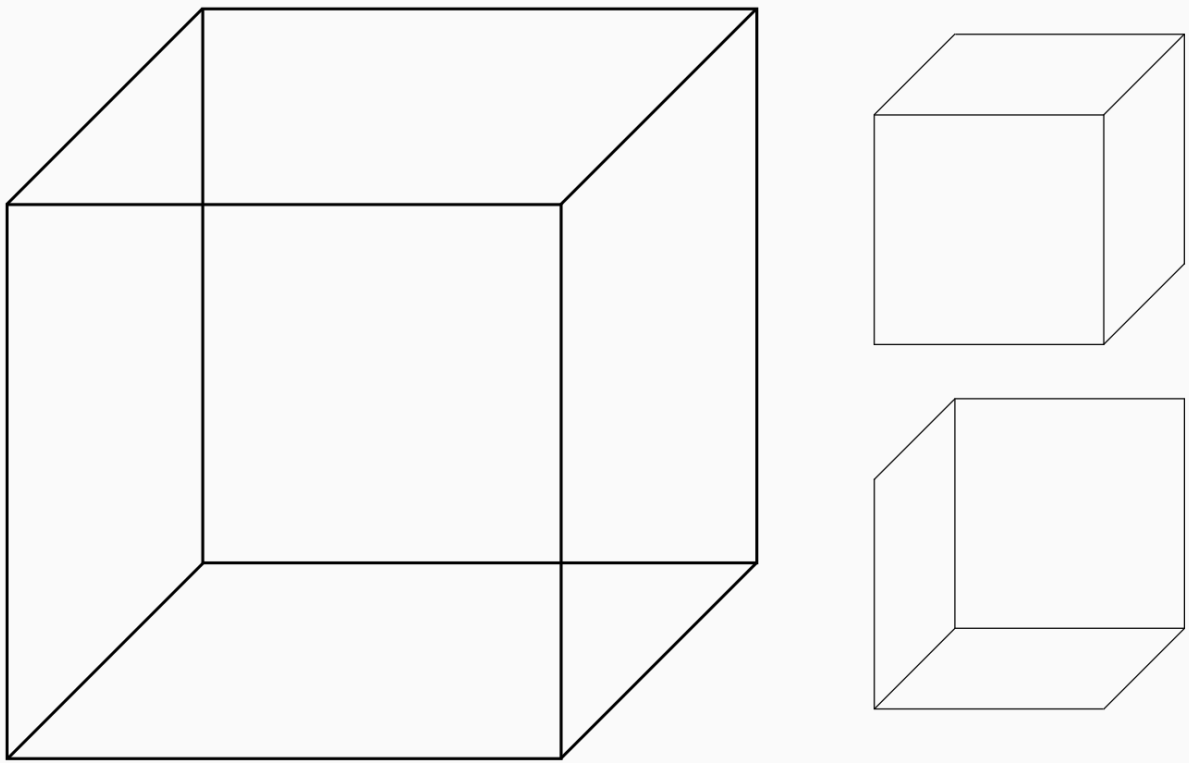
Intraclass variation

slide credit: Fei-Fei Li, Andrej Karpathy and Justin Johnson

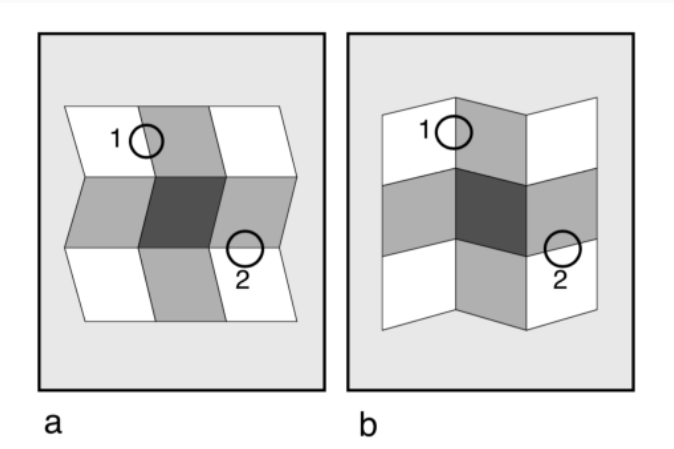
“Where” task

The “where” task is also difficult.

- The forward problem (computer graphics) is well-posed (from 3D to 2D);
- the “inverse problem” (computer vision) is ill-posed (from 2D to 3D).

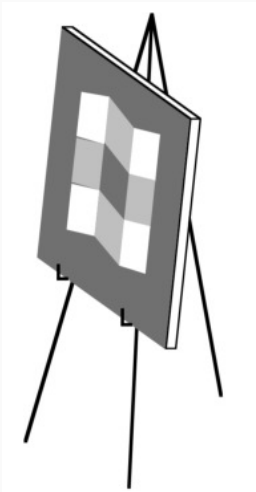


Many explanations

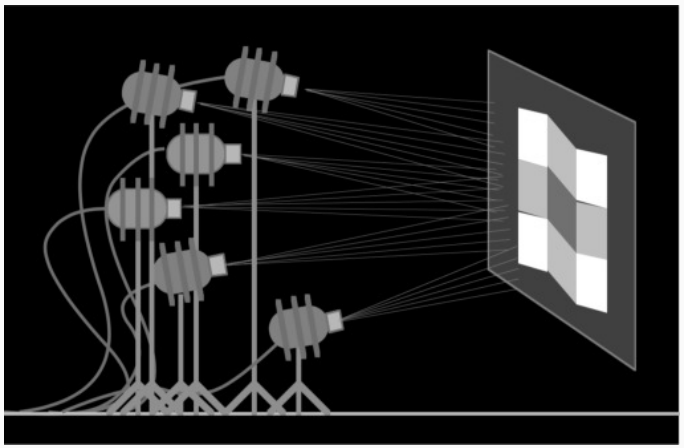


(Adelson and Pentland, 1996)

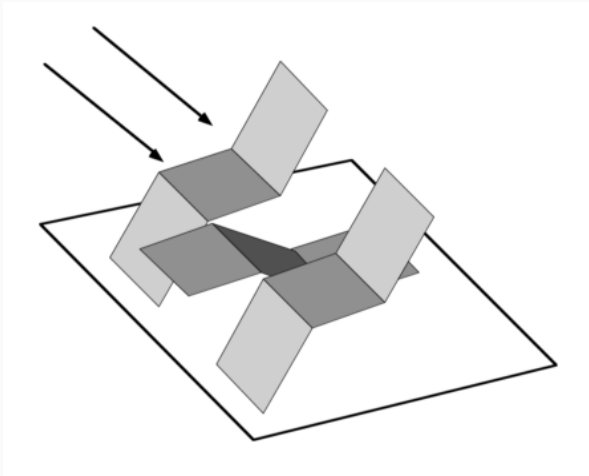
How can the appearance (b) be explained?



Painter's explanation (reflectance).



Lighting designer's explanation (local illumination).



Sheet-metal worker's explanation (shading).

Many explanations (cont.)

Joe Hill, “Ropebridge”, from http://joehill-art.com/

Joe Hill, “Ant”, from http://joehill-art.com/

A summer project

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".

(Papert, 1966)

Underestimation of the difficulty

The first great revelation was that the problems are difficult. Of course, these days this fact is a commonplace. But in the 1960s almost no one realized that machine vision was difficult. The field had to go through the same experience as the machine translation field did in its fiascoes of the 1950s before it was at least realized that here were some problems that had to be taken seriously.

(David Marr)

People who have not worked in the field often underestimate the difficulty of the problem. [...] This misperception that vision should be easy dates back to the early days of artificial intelligence [...], when it was initially believed that the cognitive *(logic proving and planning) parts of intelligence were intrinsically more difficult than the* perceptual *components.*

(Richard Szeliski)

Human vision is awesomely effective...

Why people usually underestimates the difficulties of vision?

> *The reason for this misperception is that we humans are ourselves so good at vision.* *(David Marr)*

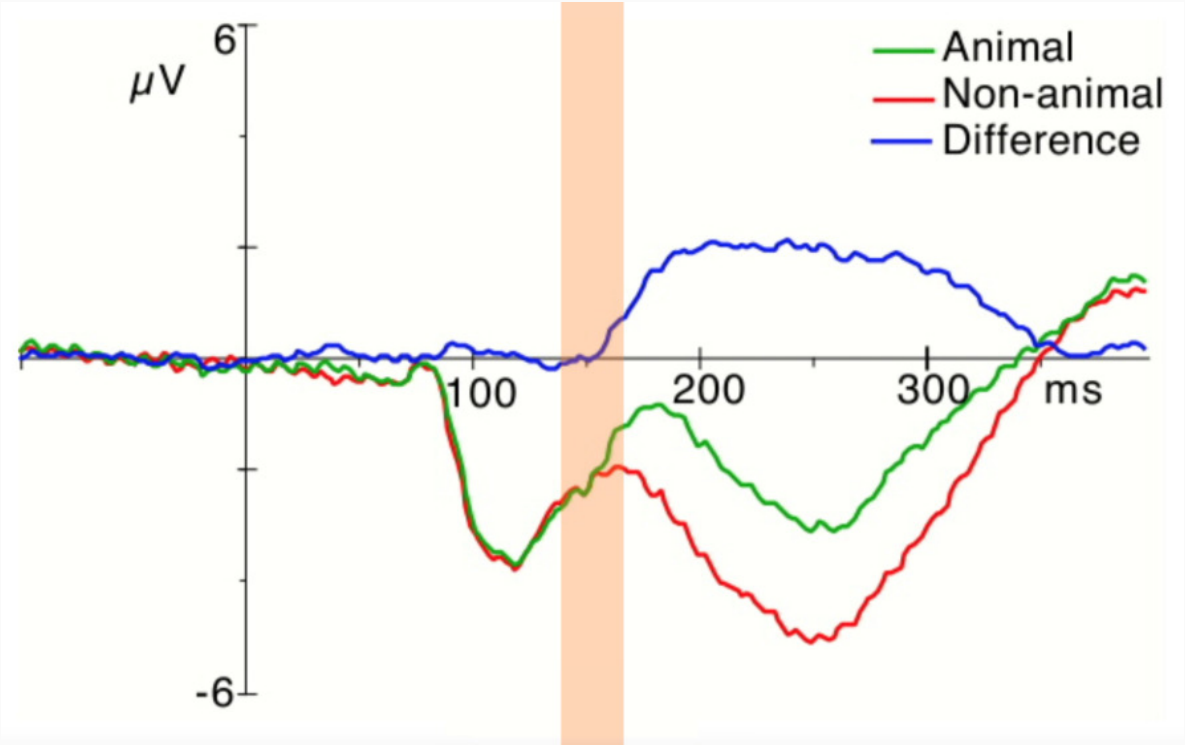
Ronald C. James, “Dalmatian”, Life Magazine, February 19, 1965

... and efficient

How long does it take for the human visual system to process a complex natural image?

... and efficient (cont.)

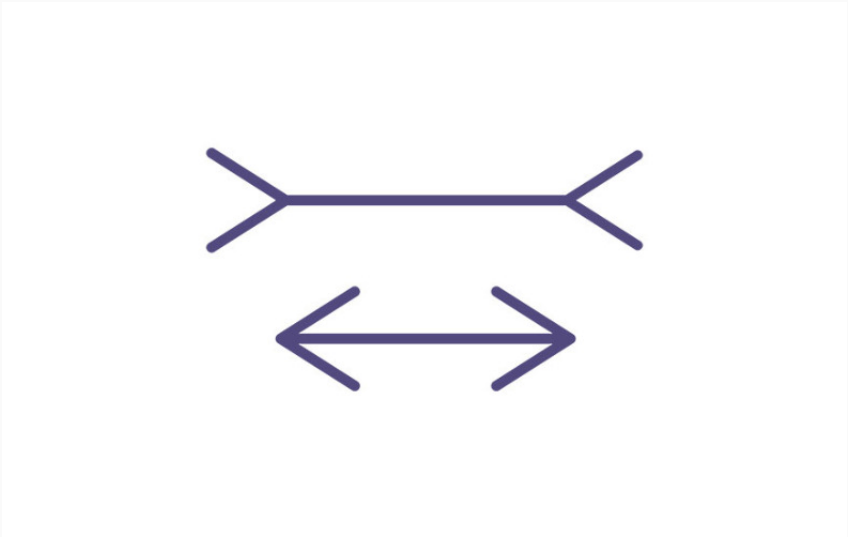
Experiment with Event Related Potentials (Thorpe et al., 1996): go/nogo categorization task, in which subjects have to decide whether a previously unseen photograph, flashed for just 20 ms, contains an animal.



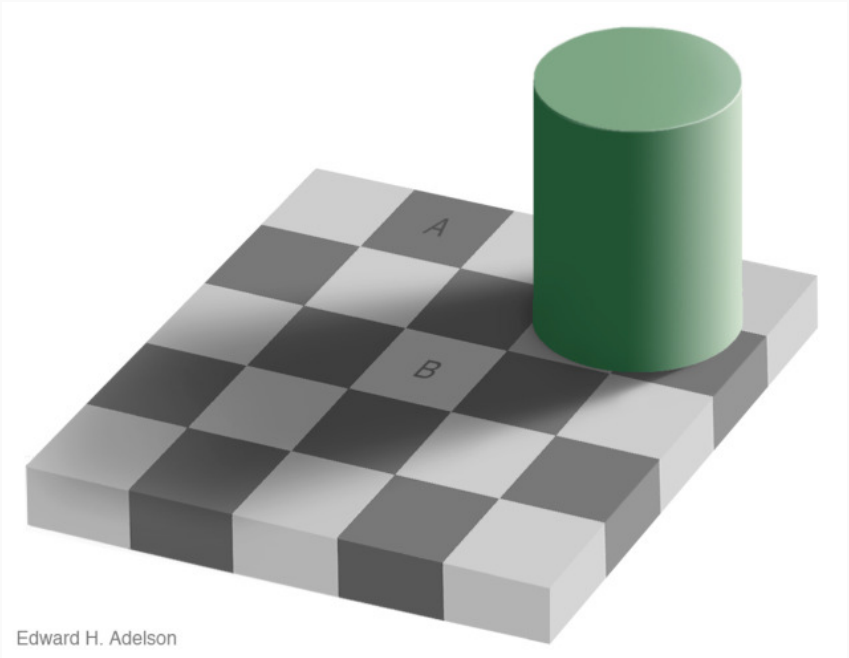
(Thorpe et al., 1996)

Sensing and perceiving

There are special mechanisms in human vision. Even low-level features such as size (length) and intensity are processed in a non-trivial way.



Müller-Lyer illusion. Probably due to the imagined perspective effects.



Checker shadow illusion. Human visual system tries to discount illumination when interpreting colors. From http://persci.mit.edu/gallery/checkershadow

Many approaches and algorithms

- Human vision may inspire, but most Computer Vision deals with finding effective ways for solving specific problems (engineering perspective);
- there is no “the” Computer Vision algorithm but instead **a collection of algorithms/approaches** for tackling **specific problems** in **restricted domains**.

Two classes of problems come from the “what” and “where” tasks:

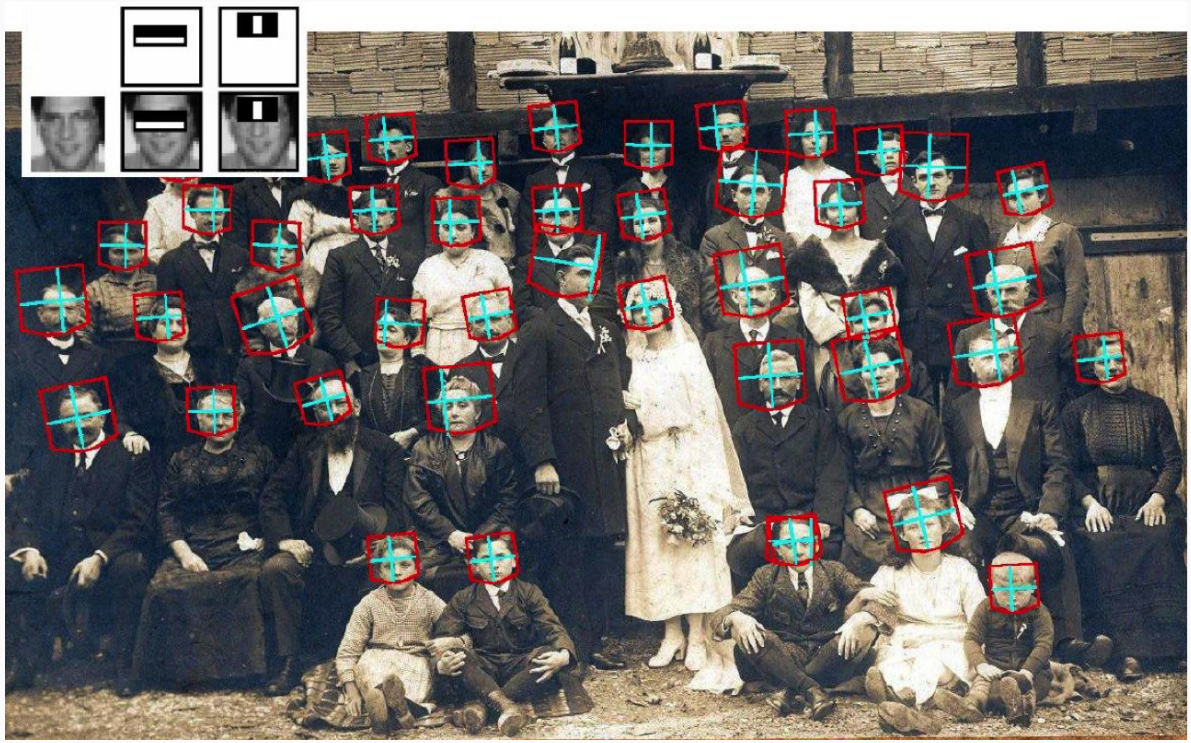
- Recognition (in broad sense)
  - *object detection* (find all the regions in an image where a specific kind of object is likely to occur, for instance face detection and pedestrian detection);
  - *instance recognition* (recognize a known specific object potentially being viewed from a novel viewpoint);
  - *category-level recognition* (categorize images as belonging to a general class such as “cat” or “bicycle,” among many possible classes.)
- Reconstruction
  - recovering the 3D structure of an object or a scene, given a sufficient amount of images of the object/scene.

A more detailed taxonomy

A more detailed taxonomy is as follows. Notice that here “recognize” refers to the category-level recognition.

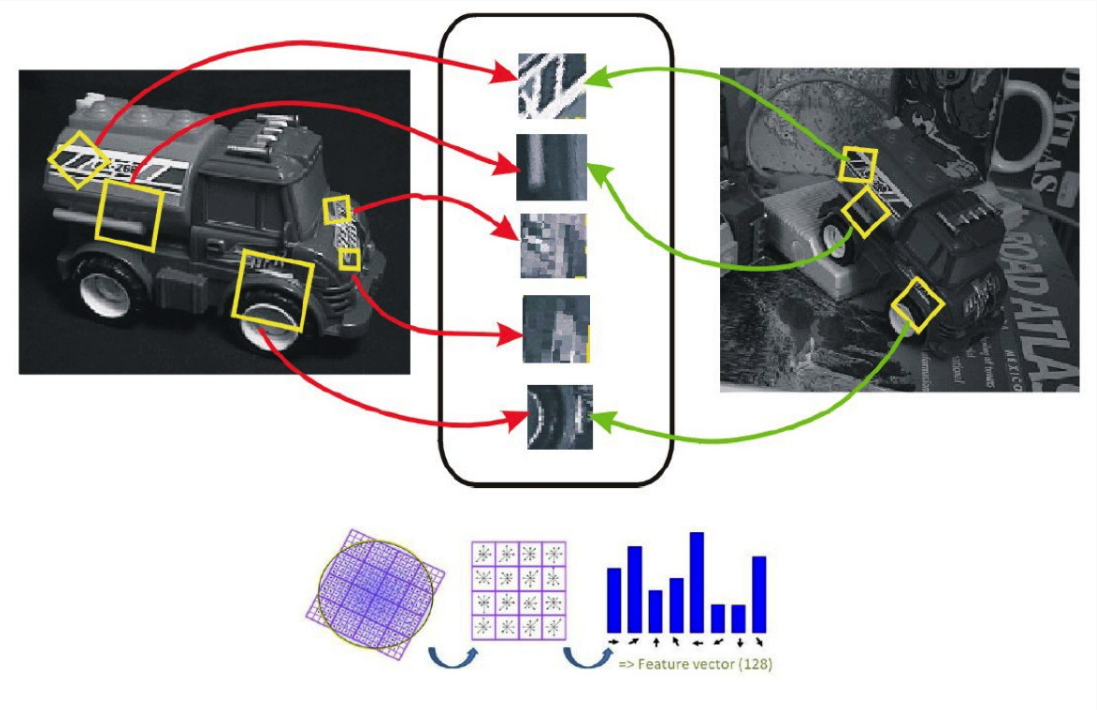
- Reconstruct: given two or more 2D images of a scene, compute a 3D model of it.
- Detect: is there one or more instance of x in this image?
- Localize: where are the instances?
- Segment: where are the boundaries of each instance?
- Track: how is this instance moving from one image to the next?
- Recognize = classify: what is this?

Object detection



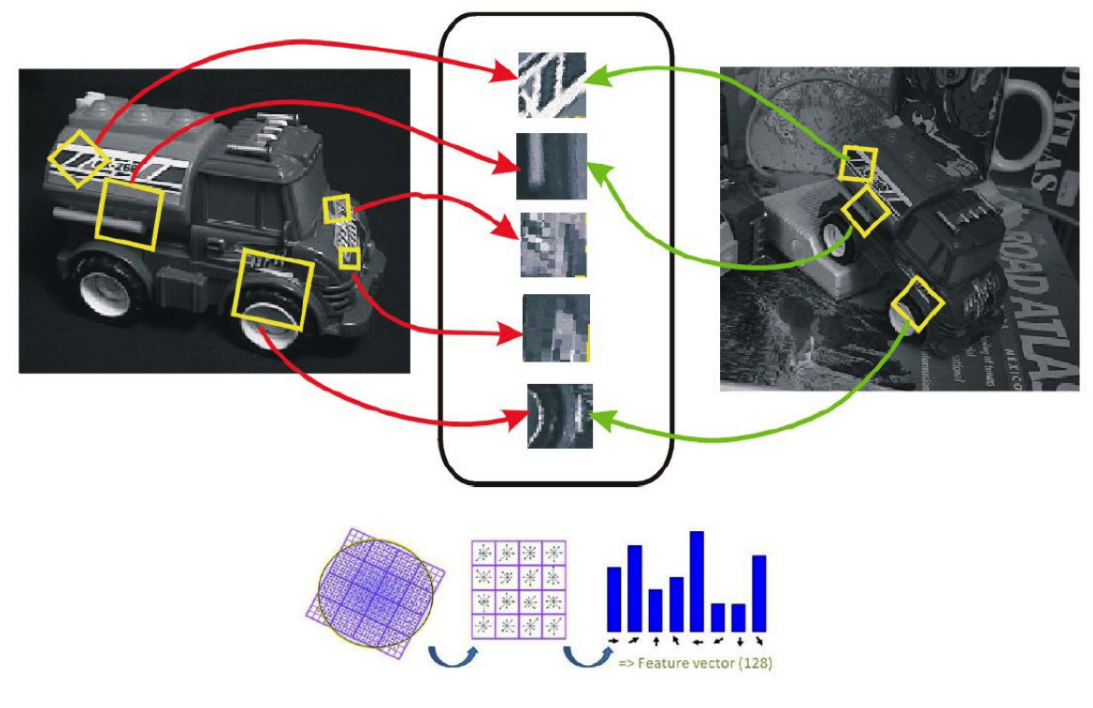
Face detection, (Viola and Jones, 2001)

Instance recognition

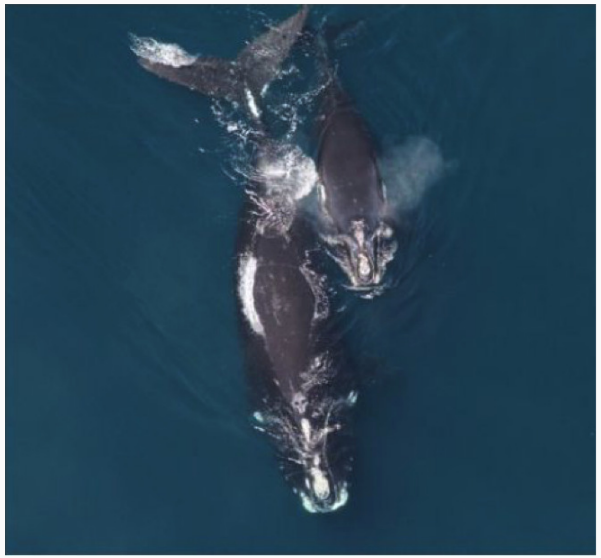


SIFT and object recognition, (Lowe, 1999)

Instance recognition

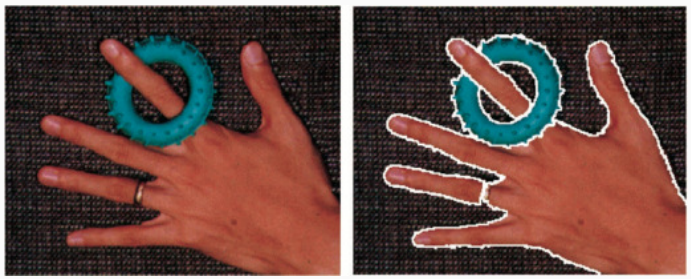


SIFT and object recognition, (Lowe, 1999)

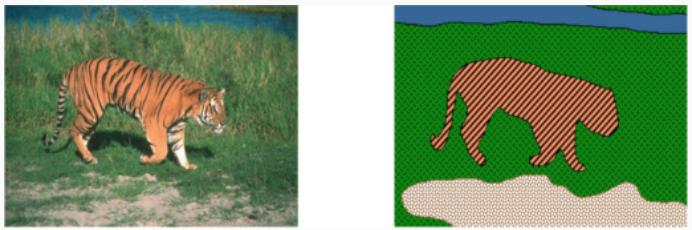


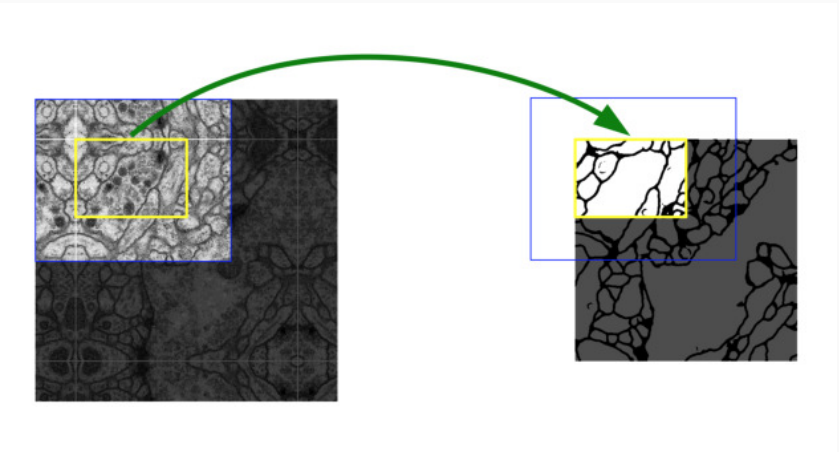
Right whale recognition, Kaggle Challenge, https://www.kaggle.com/c/noaa-right-whale-recognition

Segmentation



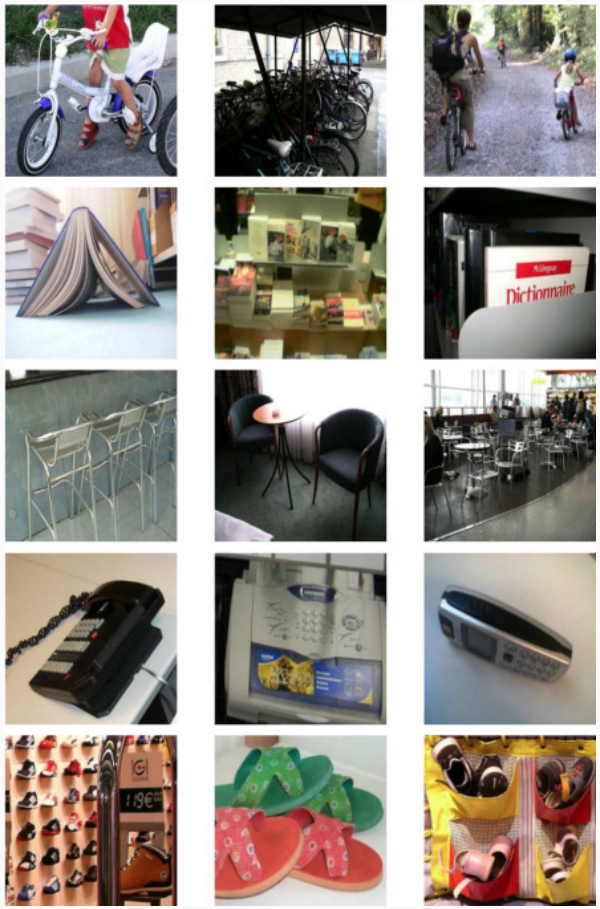
(Comaniciu and Meer, 2002)



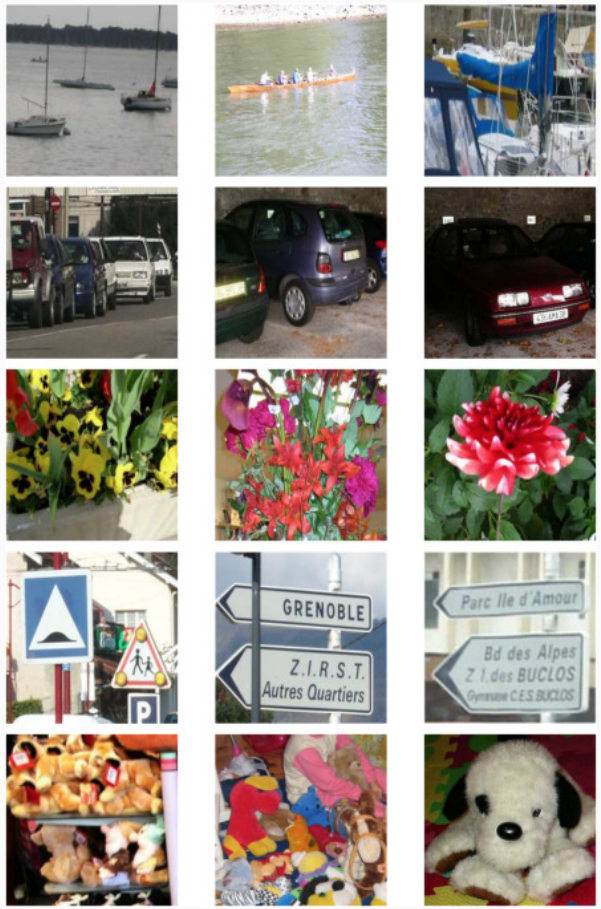


(Ronneberger et al., 2015)

Category-level recognition



(Csurka et al., 2006)



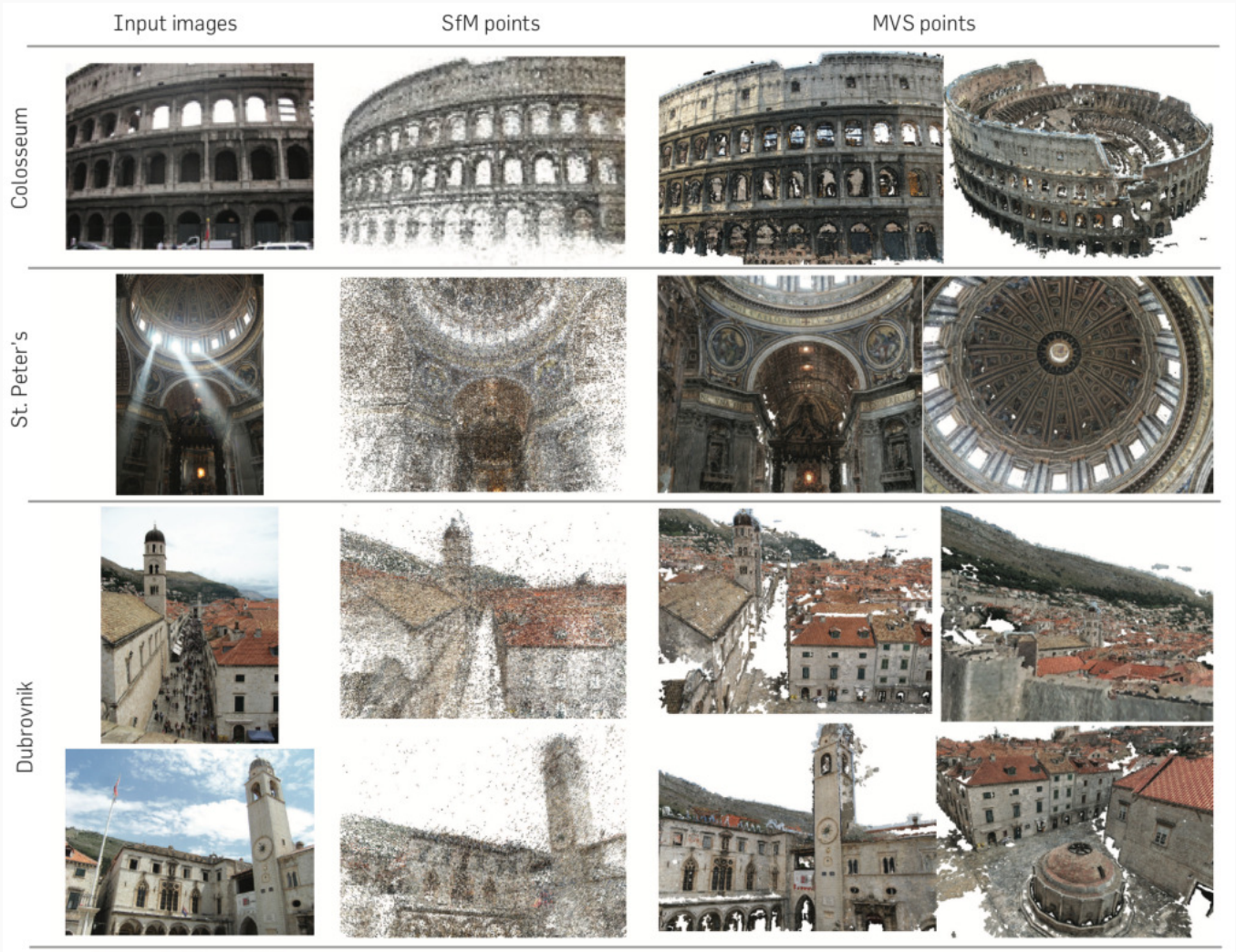
(Csurka et al., 2006)

Category-level recognition is *extremely challenging*, and remains still largely unsolved. However, dramatic progresses have been made, especially in the last decade.

Two noticeable achievements

- Uncalibrated reconstruction (Agarwal et al., 2011)
- Image categorization (Krizhevsky et al., 2012)

Uncalibrated reconstruction



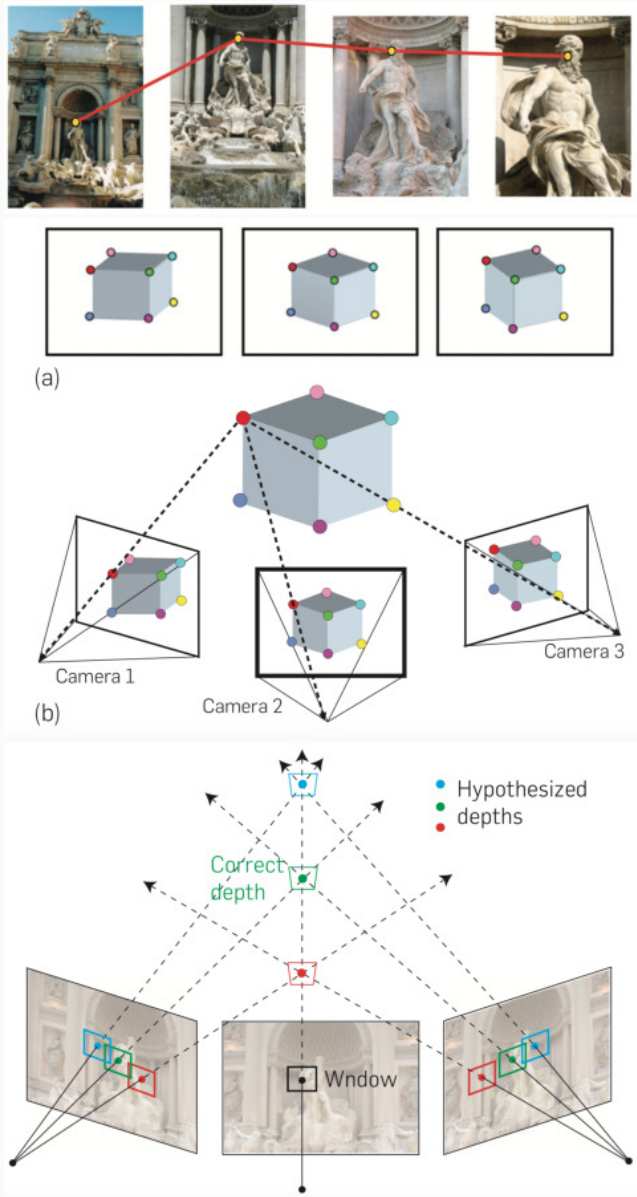
(Agarwal et al., 2011)

Uncalibrated reconstruction (cont.)

- 496 processors;
- 1984 gigabytes of total memory;
- 62 terabytes of disk;
- 460000 Flickr pictures of Rome, Venice and Dubrovnik;
- $\approx$ 100 hours of computation.
- Output: detailed three-dimensional geometry and colors of monuments in Rome, Venice and Dubrovnik.

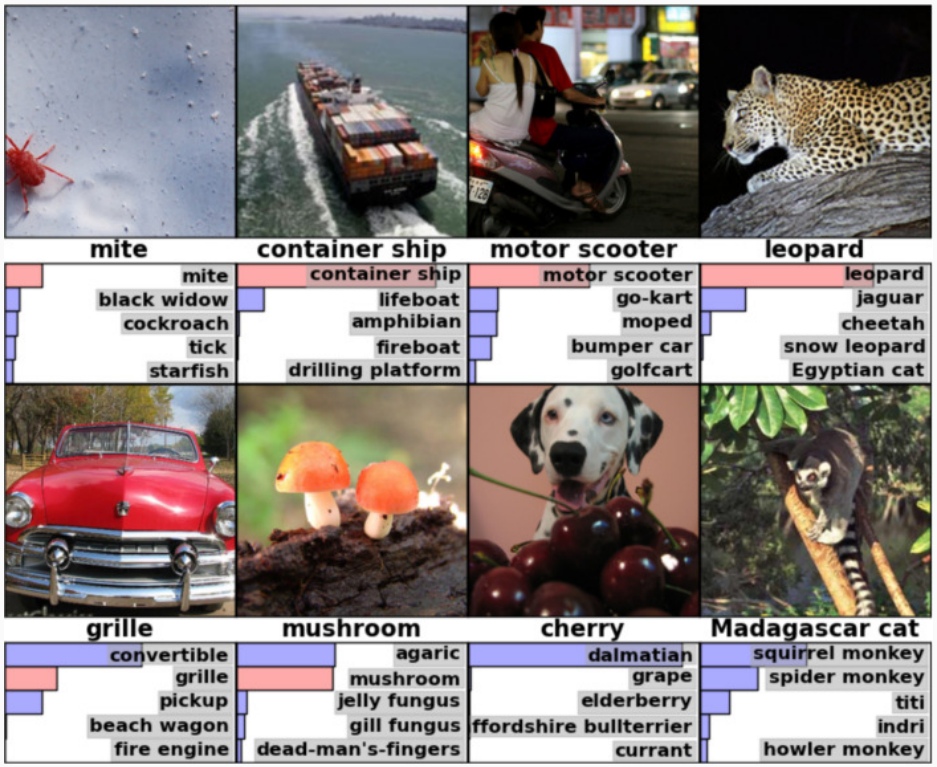
Three well-known Computer Vision problems, solved efficiently and effectively:

- correspondence;
- structure from motion;
- multiview stereo.



(Agarwal et al., 2011)

AlexNet



ImageNet (http://www.image-net.org)

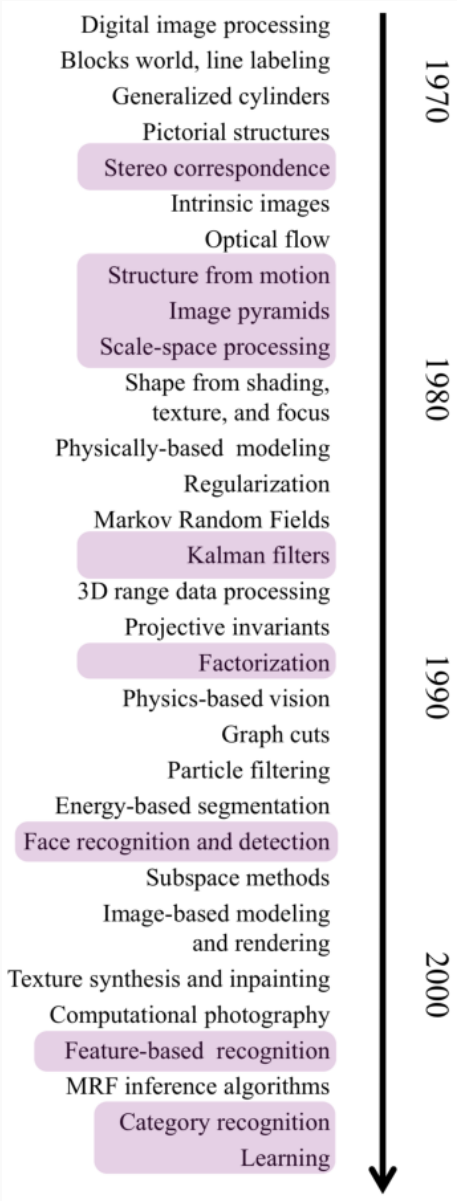
- 14M images
- 22K categories
  - Animal (fish, bird, mammal, invertebrate)
  - Plant (tree, flower, vegetable)
  - Instrumentation (utensile, appliance, tool, musical instrument)
  - ...
- annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC). From 2018 on, the challenges are hosted on Kaggle.

In 2012 a convolutional neural network won the challenge for the first time and by a wide margin, bringing down the state-of-the-art top-5 error rate from 26.1 percent to 15.3 percent (Krizhevsky et al., 2012). Since then, these competitions are consistently won by deep convolutional nets.

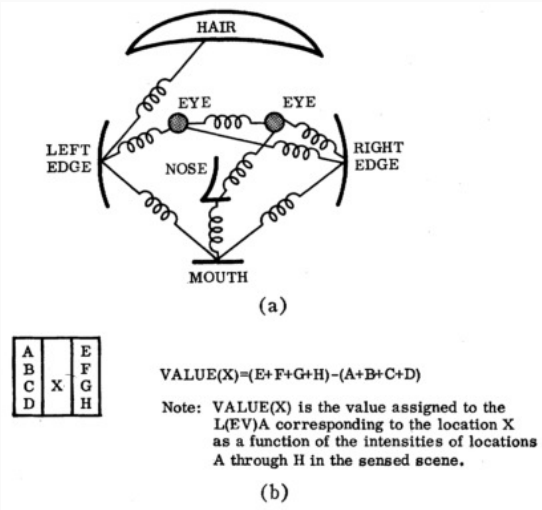
Historical sketches

Historical sketches:

- 1966: Marvin Minsky assigns computer vision as an undergrad summer project;
- 1960's: interpretation of synthetic worlds;
- 1970's: some progress on interpreting selected images;
- 1980's: Artificial Neural Networks come and go; shift toward geometry and increased mathematical rigor;
- 1990's: face recognition; statistical analysis in vogue;
- 2000's: broader recognition; large annotated datasets available; video processing starts;
- 2010's: Deep learning with Convolutional Neural Networks.

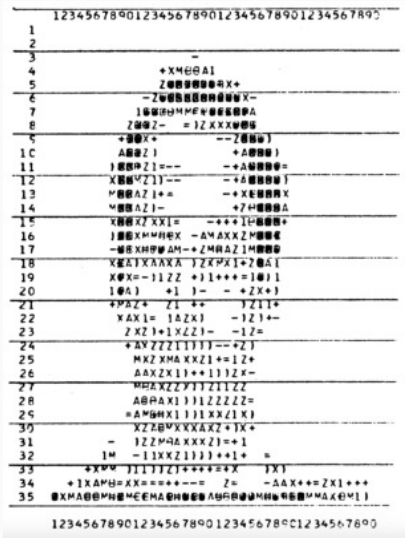


A pioneering approach



"Pictorial structure" - (Fischer and Elschlager, 1973)

- ad-hoc detectors, tailored for the different parts
- ad-hoc relation between parts
- detection is formulated as an *optimization problem*:
  - maximize the local matching
  - minimize the springs' stretching



HAIR WAS LOCATED AT (6, 18)
L/EDGE WAS LOCATED AT (18, 10)
R/EDGE WAS LOCATED AT (18, 25)
L/EYE WAS LOCATED AT (17, 13)
R/EYE WAS LOCATED AT (17, 21)
NOSE WAS LOCATED AT (22, 18)
MOUTH WAS LOCATED AT (24, 17)

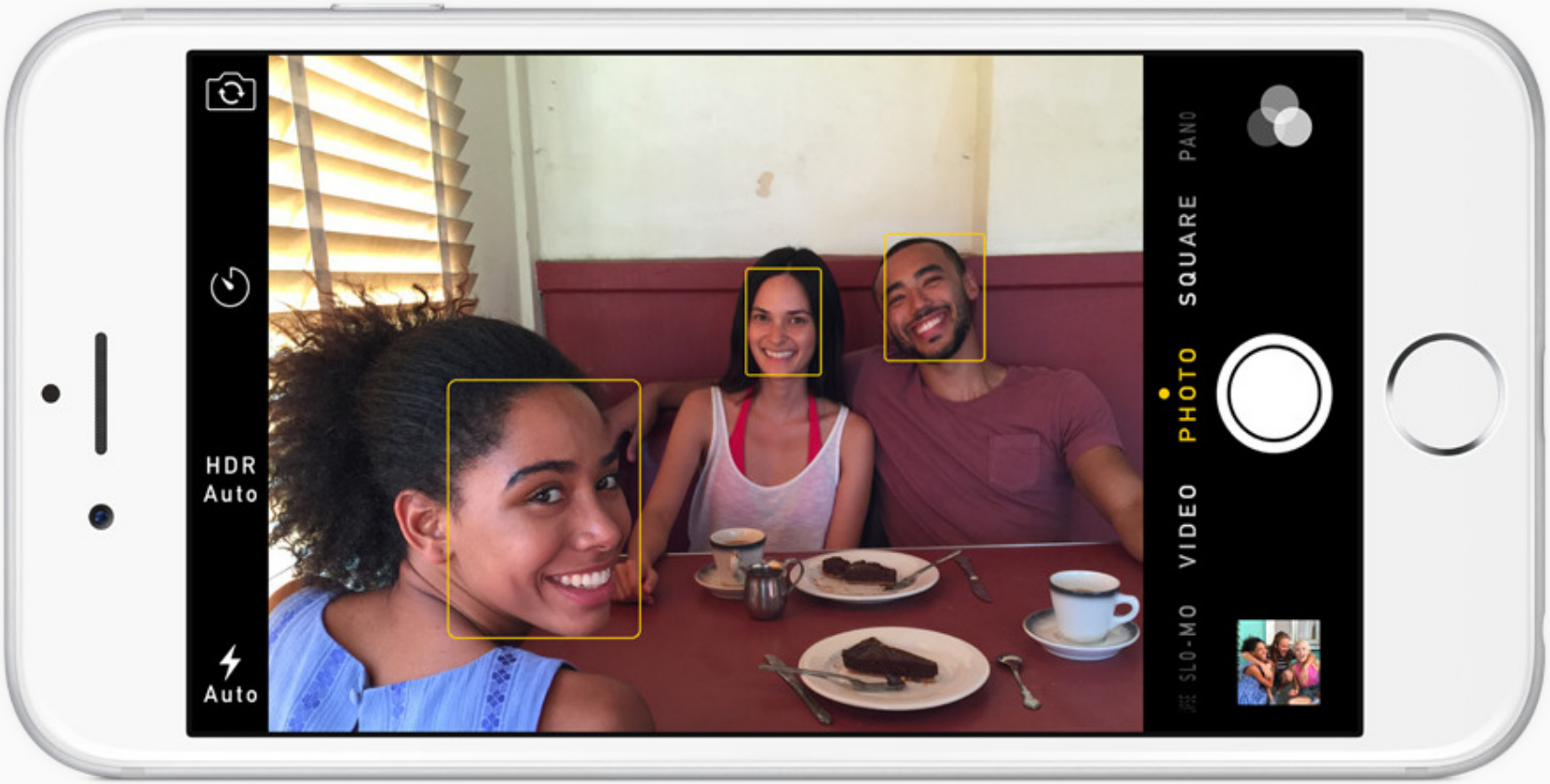
Applications

- Optical Character Recognition
- Movies
- Surveillance
- Human Computer Interface – hand gestures
- Aids to the blind
- Face recognition and biometrics
- Industrial inspection
- Virtual Earth; Street view
- Robotic control
- Autonomous driving
- Space: planetary exploration, docking
- Medicine – pathology, surgery, diagnosis
- Microscopy
- Behavioral studies
- Remote Sensing
- Digital photography
- Video games

Optical character recognition and other simple patterns recognition

Slide credit: Svetlana Lazebnik

Face detection

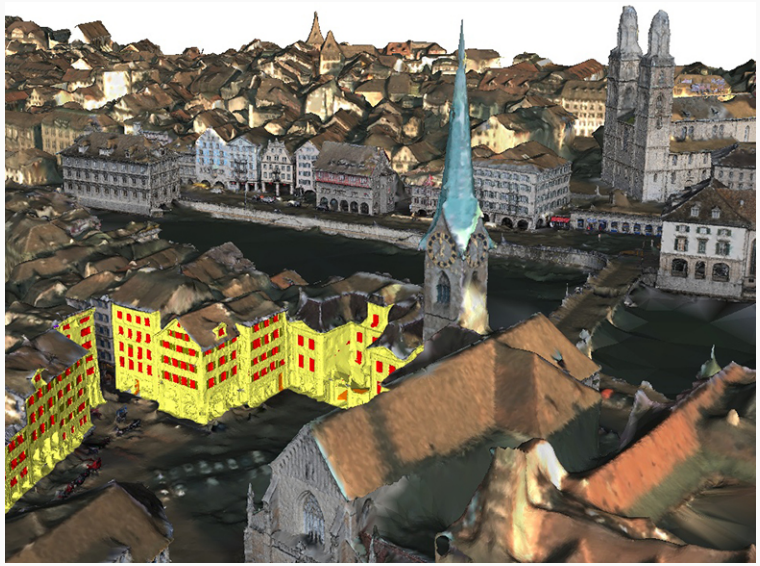


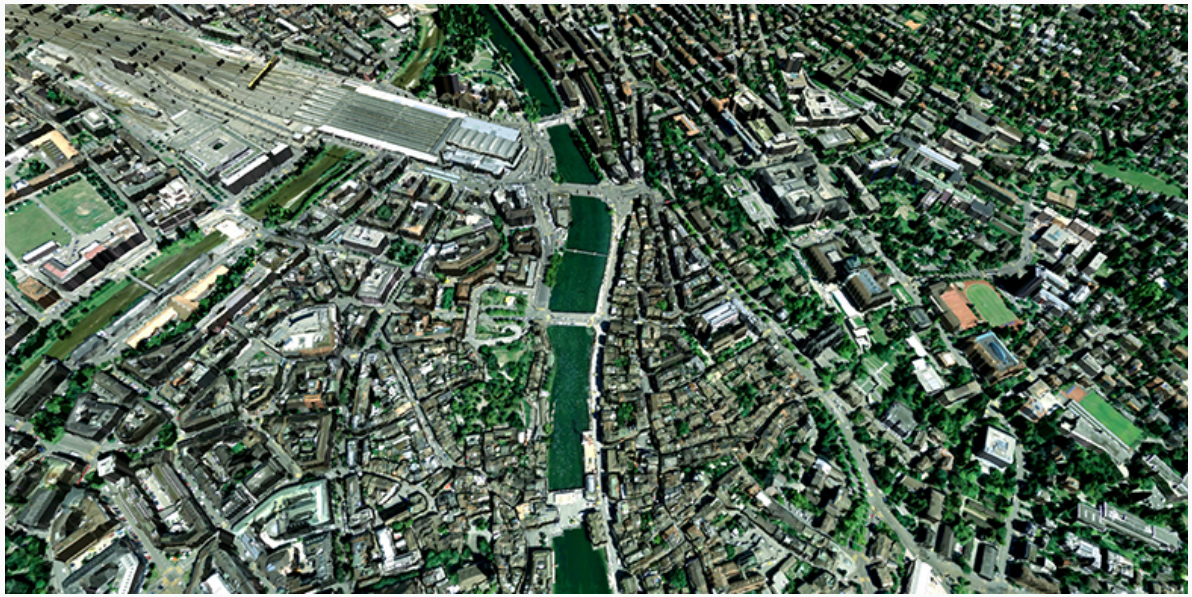
Face recognition

Surveillance

Privacy masking

Urban modelling





Images from https://varcity.ethz.ch/.

Robotics

NASA's Mars Curiosity Rover.

- Panorama stitching;
- 3D terrain modeling;
- obstacle detection, position tracking.

Field guides

leafsnap

Home | Species | Collectors | About

Leaf of the Bottlebrush Buckeye

Leafsnap: An Electronic Field Guide

Leafsnap is the first in a series of electronic field guides being developed by researchers from Columbia University, the University of Maryland, and the Smithsonian Institution. This free mobile app uses visual recognition software to help identify tree species from photographs of their leaves.

Leafsnap contains beautiful high-resolution images of leaves, flowers, fruit, petiole, seeds, and bark. Leafsnap currently includes the trees of the Northeast and will soon grow to include the trees of the entire continental United States.

This website shows the tree species included in Leafsnap, the collections of its users, and the team of research volunteers working to produce it.

Free for iPhone: Available on the App Store **and iPad:** Available on the iPad App Store

YouTube

guardian.co.uk

Motion capture

Andy Serkis as the ape Caesar in “War for the Planet of the Apes.”

Self-driving cars

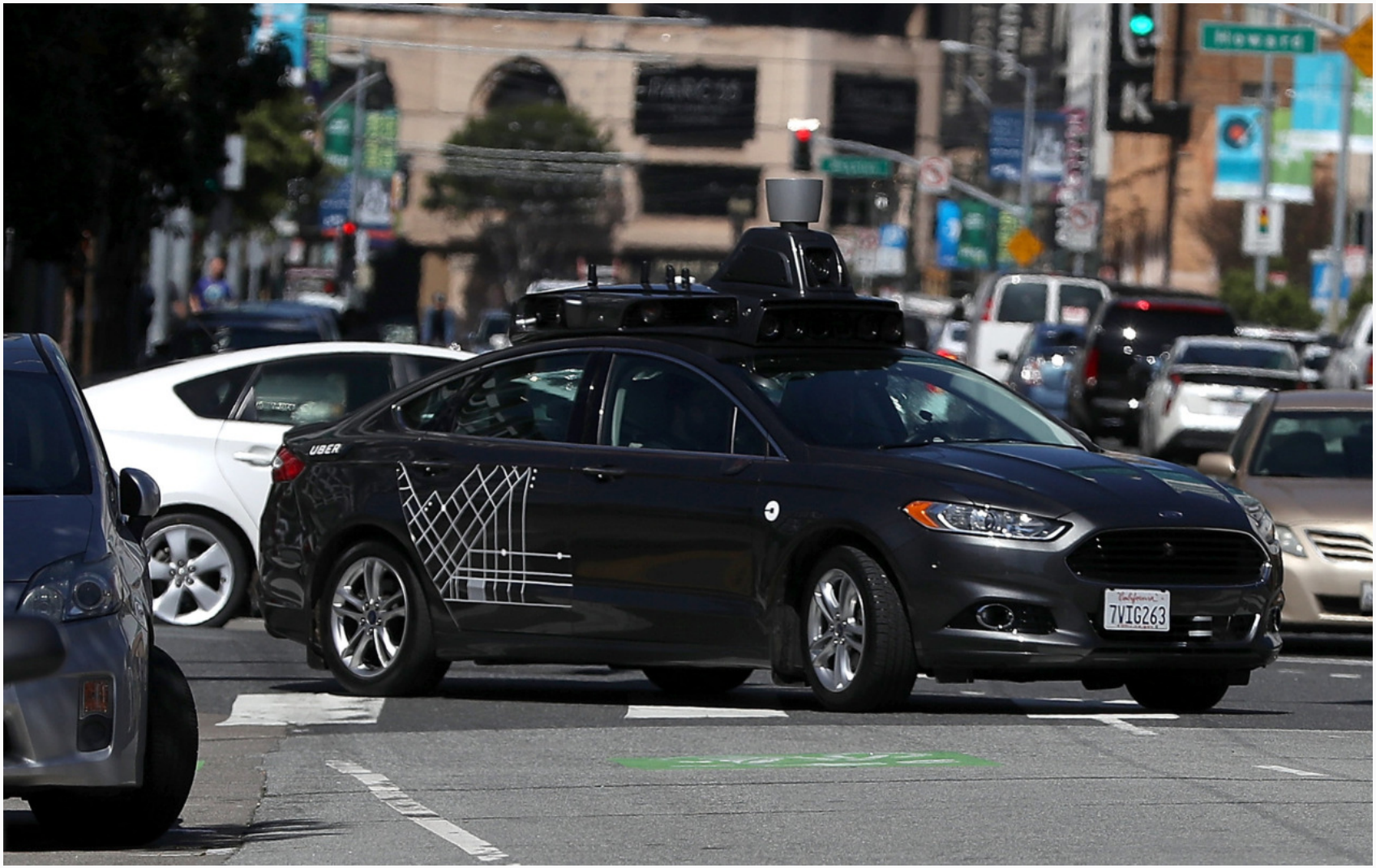
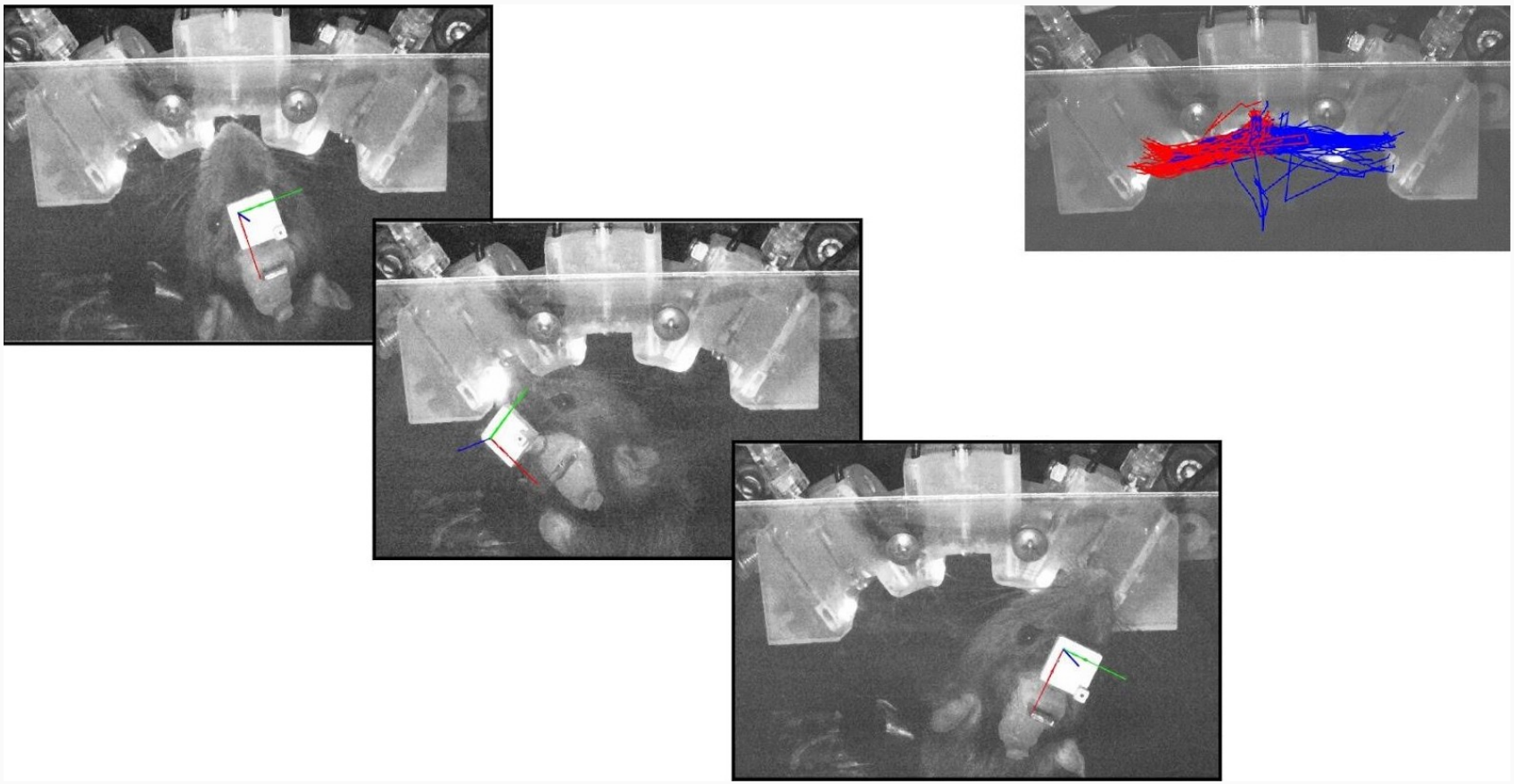


Image source: Justin Sullivan/Getty Images.

Vision-based interaction

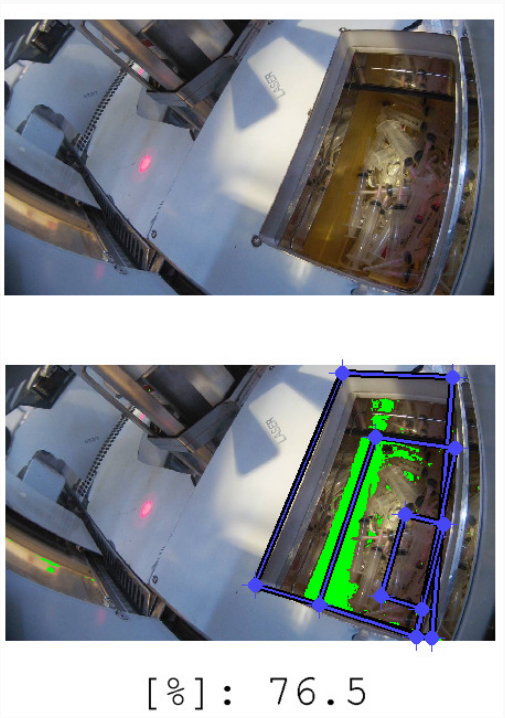
Agriculture

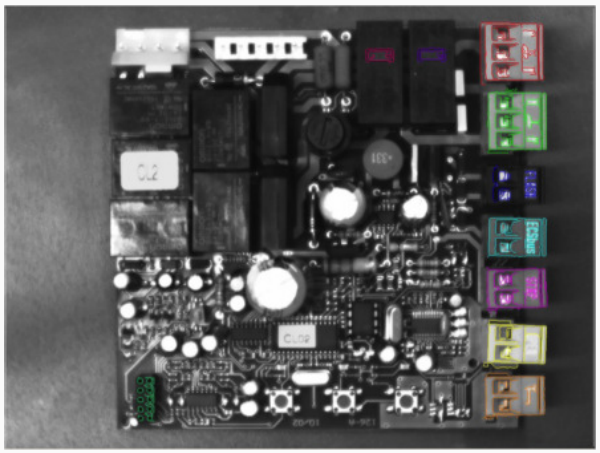
Animal behavior study



Real-time head tracking application. Image courtesy of Visual Neuroscience Lab, International School for Advanced Studies (SISSA), Trieste.

Inspection





About notation

In most of the cases:

- vectors are denoted by lower case italic (e.g. v);
- scalars are denoted by mixed case italic, (e.g. γ, A),
- matrices are denoted by upper case italic, (e.g. M);
- vectors operate as column vectors, i.e., they post-multiply matrices, (e.g. Mv);
- v_1 may denote the first component of vector v, or the first column of matrix V;
- $v_1^\top$ denotes the first row of matrix V;
- $P \succ 0$ means that the matrix P is positive definite;
- $P \succeq 0$ means that the matrix P is positive semi-definite;
- ∇f is the gradient of f and is a column vector.

However, we will not be picky about notation. Sometimes, to be close to the notation of the original paper, we will sacrifice consistency.

References

References

Adelson, E. H. and Pentland, a. P. (1996). The perception of shading and reflectance. In Knill, D. and Richards, W., editors, *Perception as Baysian Inference*, pages 409–423. Cambridge University Press, New York.

Agarwal, S., Furukawa, Y., and Snavely, N. (2011). Building rome in a day. *Communications of the ...*, pages 105–112.

Ballard, D. H. and Brown, C. M. (1982). *Computer Vision*. Prentice Hall Professional Technical Reference, 1st edition.

Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

Csurka, G., Dance, C. R., Perronnin, F., and Willamowski, J. (2006). Generic visual categorization using weak geometry. In *Toward Category-Level Object Recognition*, pages 207–224. Springer.

Fischer, M. and Elschlager, R. (1973). The representation and matching of pictorial structure. *IEEE Trans. Comput*, 1:67–92.

Forsyth, D. and Ponce, J. (2012). *Computer vision: a modern approach (2nd edition)*. Pearson Education Limited.

References (cont.)

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The proceedings of the seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. WH San Francisco: Freeman and Company.

Papert, S. (1966). The summer vision project. Technical report, Massachusetts Institute of Technology.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., and Tops, J. (2002). Video-to-3d. *International archives of photogrammetry remote sensing and spatial information sciences*, 34(3/A):252–257.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer.

Snavely, N., Simon, I., Goesele, M., Szeliski, R., and Seitz, S. M. (2010). Scene reconstruction and visualization from community photo collections. *Proceedings of the IEEE*, 98(8):1370–1390.

References (cont.)

Stockman, G. and Shapiro, L. G. (2001). *Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522.

Trucco, E. and Verri, A. (1998). *Introductory techniques for 3-D computer vision*. Prentice Hall Englewood Cliffs.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE.

554SM –Fall 2018

Lecture 1
Introduction to Computer Vision

END