

Lecture 1: Introduction

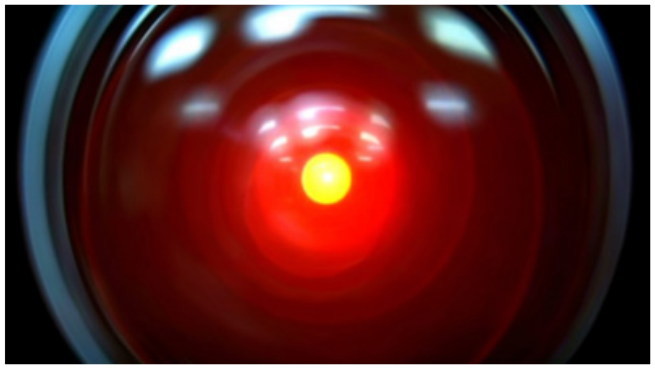
Welcome to CS231n











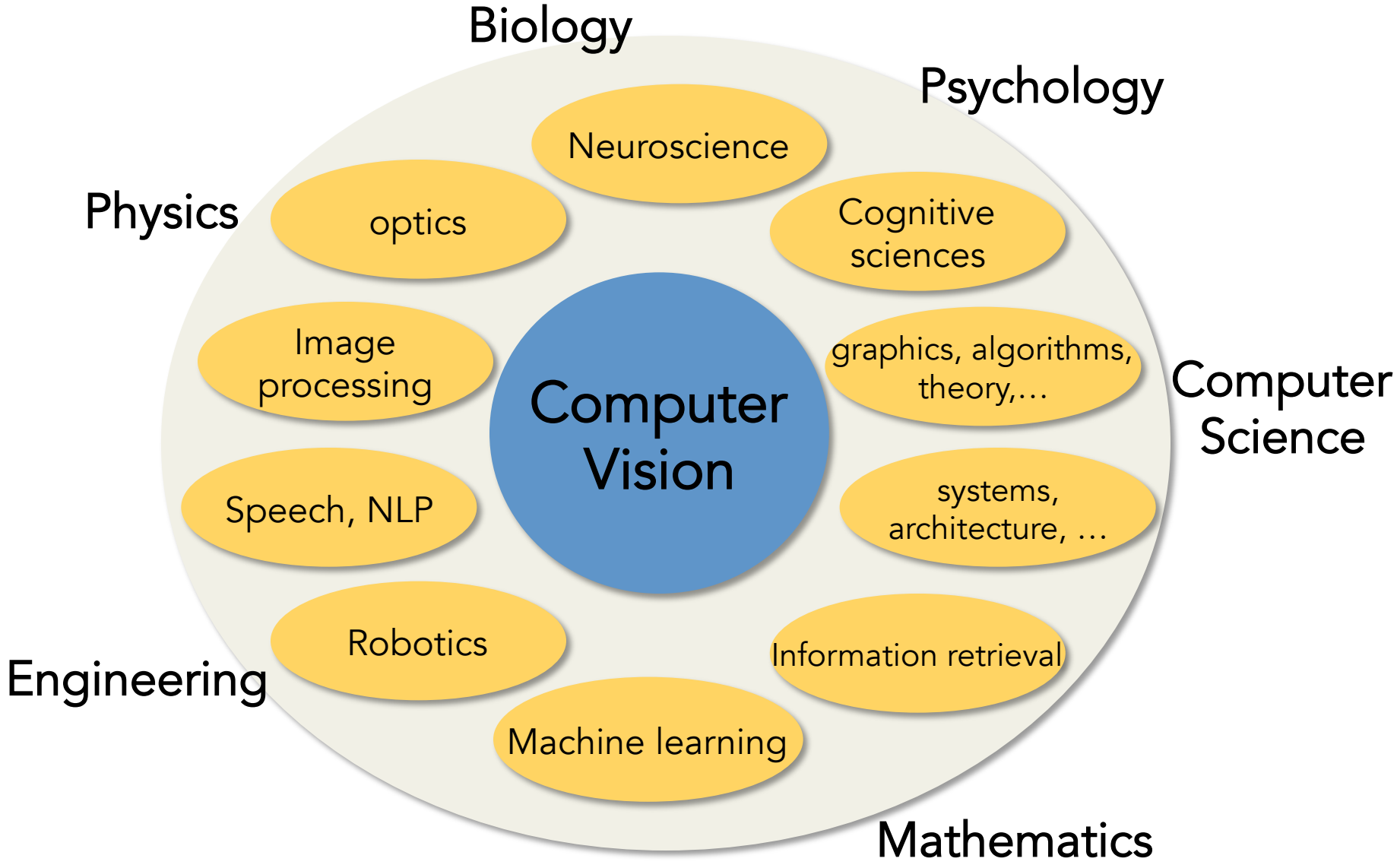


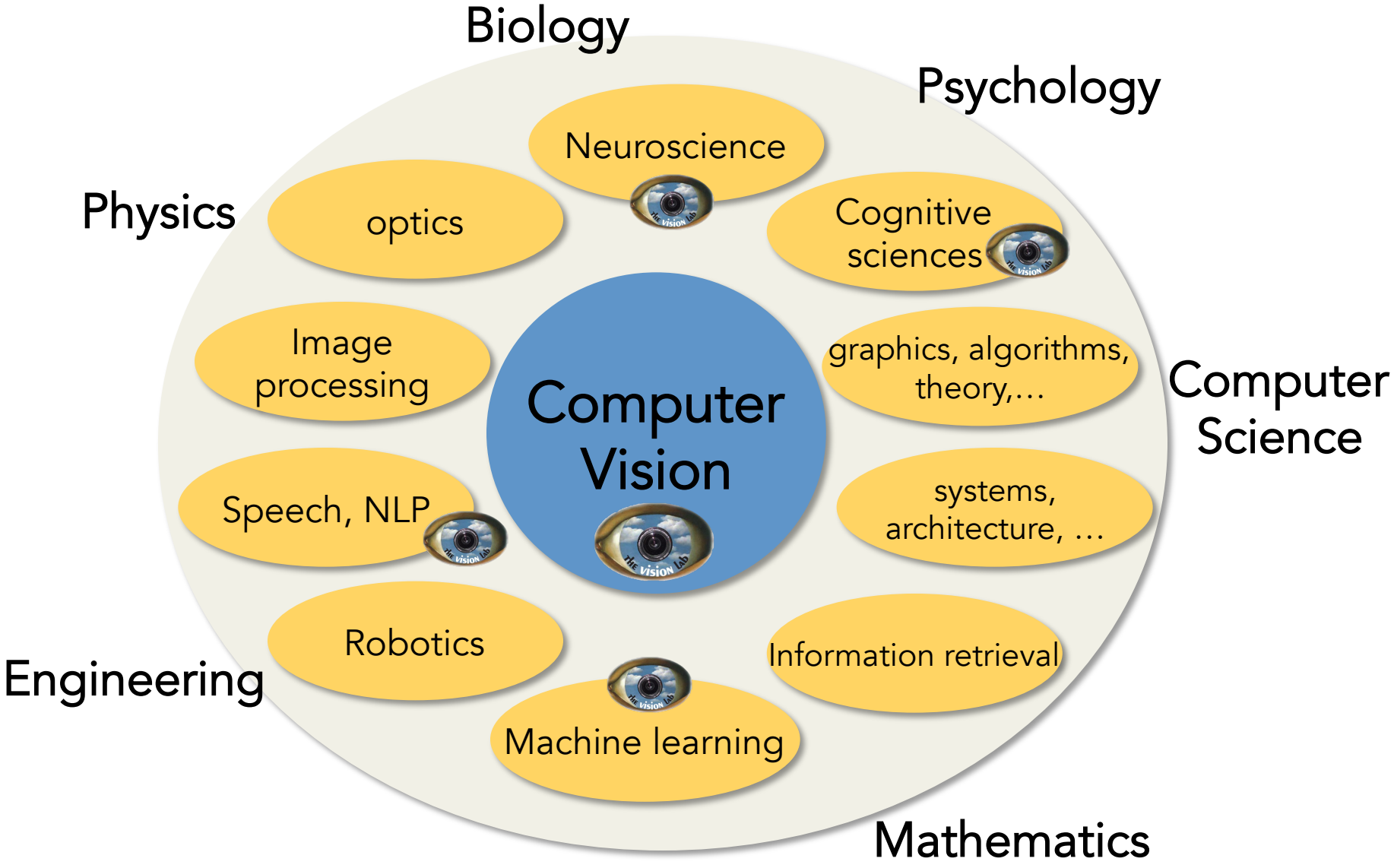









Biology
Psychology
Physics
Neuroscience
optics
Cognitive sciences
Image processing
Computer Vision
graphics, algorithms, theory,…
Computer Science
Speech, NLP
systems, architecture, …
Robotics
Information retrieval
Engineering
Machine learning
Mathematics


Biology
Psychology
Physics
Neuroscience
optics
Cognitive sciences
Image processing
Computer Vision
graphics, algorithms, theory,…
Computer Science
Speech, NLP
systems, architecture, …
Robotics
Information retrieval
Engineering
Machine learning
Mathematics

Computer Vision courses @ Stanford

- CS131 (fall, 2014, Prof. Fei-Fei Li):
  - Undergraduate introductory class
- CS231a (this term, Prof. Silvio Savarese)
  - Core computer vision class for seniors, masters, and PhDs
  - Topics include image processing, cameras, 3D reconstruction, segmentation, object recognition, scene understanding
- CS231n (this term, Prof. Fei-Fei Li & Andrej Karpathy)
  - Neural network (aka "deep learning") class on image classification
- CS231b (spring, 2015, Prof. Fei-Fei Li): Cutting Edge Computer Vision
  - Project-based advanced vision class to prepare students for CV research
- CS231m (spring, 2015, Prof. Silvio Savarese): Mobile Vision
  - Computer vision and computational photography for mobile platform (e.g. Android)
- And an assortment of CS331 and CS431 for advanced topics in computer vision

Today's agenda

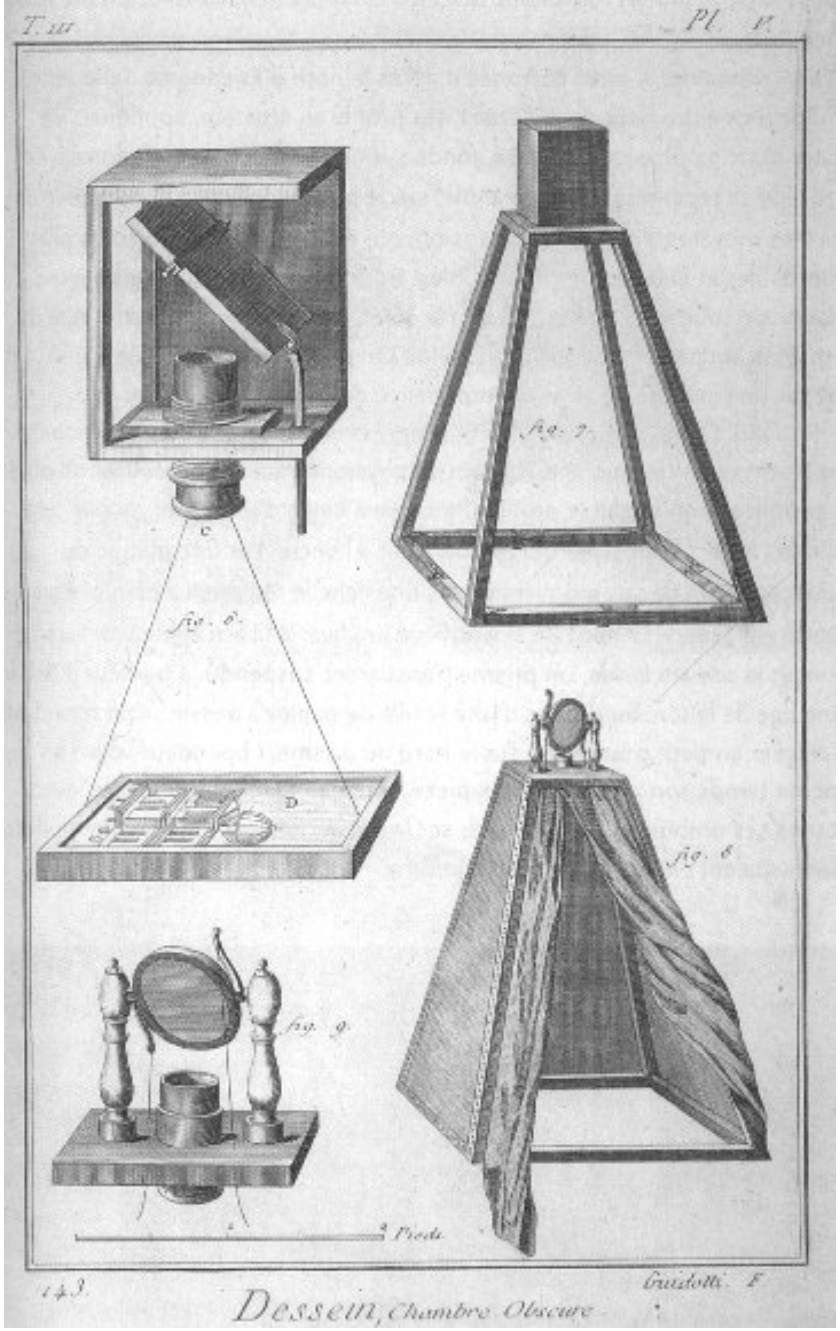
- A brief history of computer vision
- CS231n overview

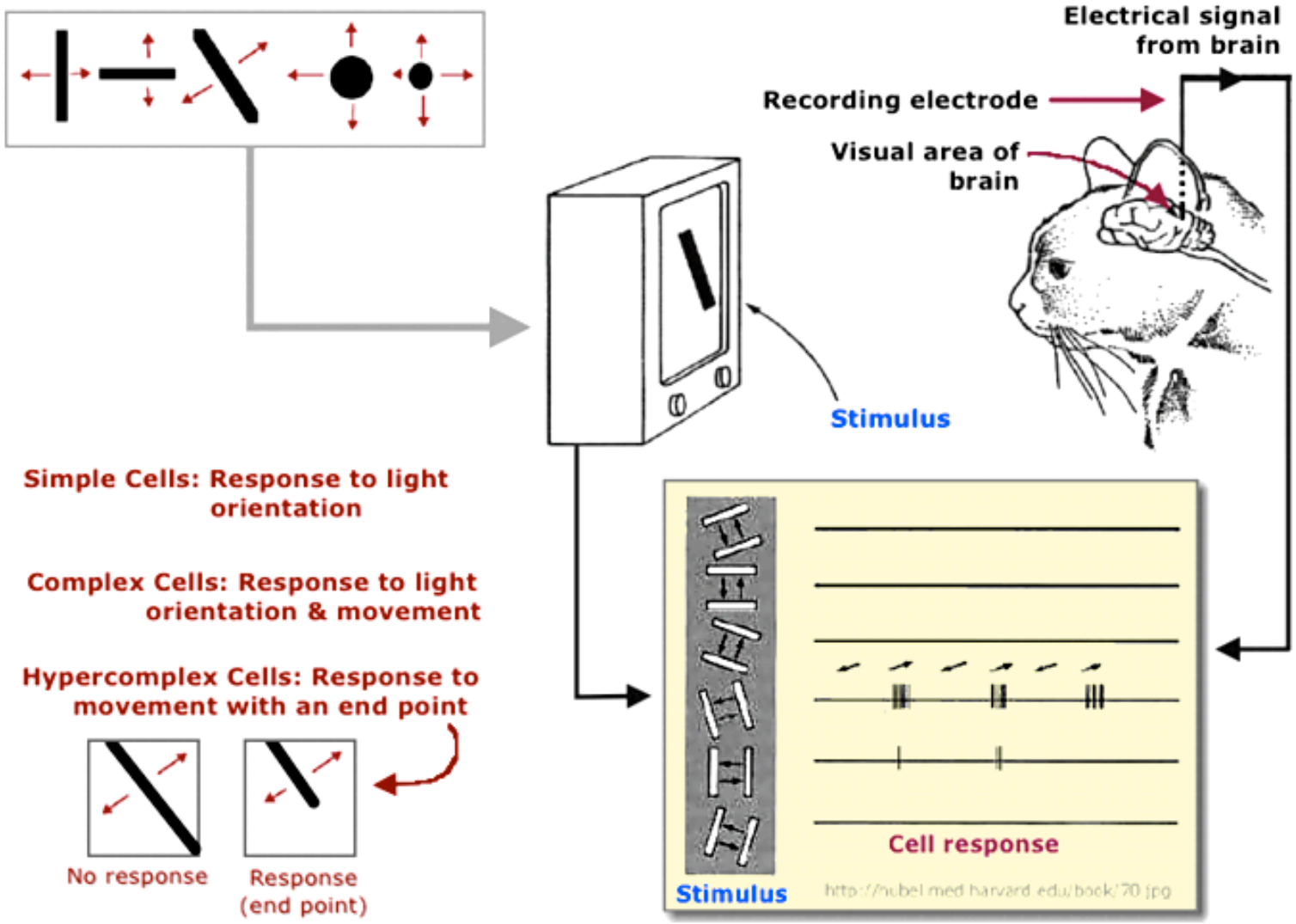


543million years, B.C.

Camera Obscura

Leonardo da Vinci
16th Century, A.D.

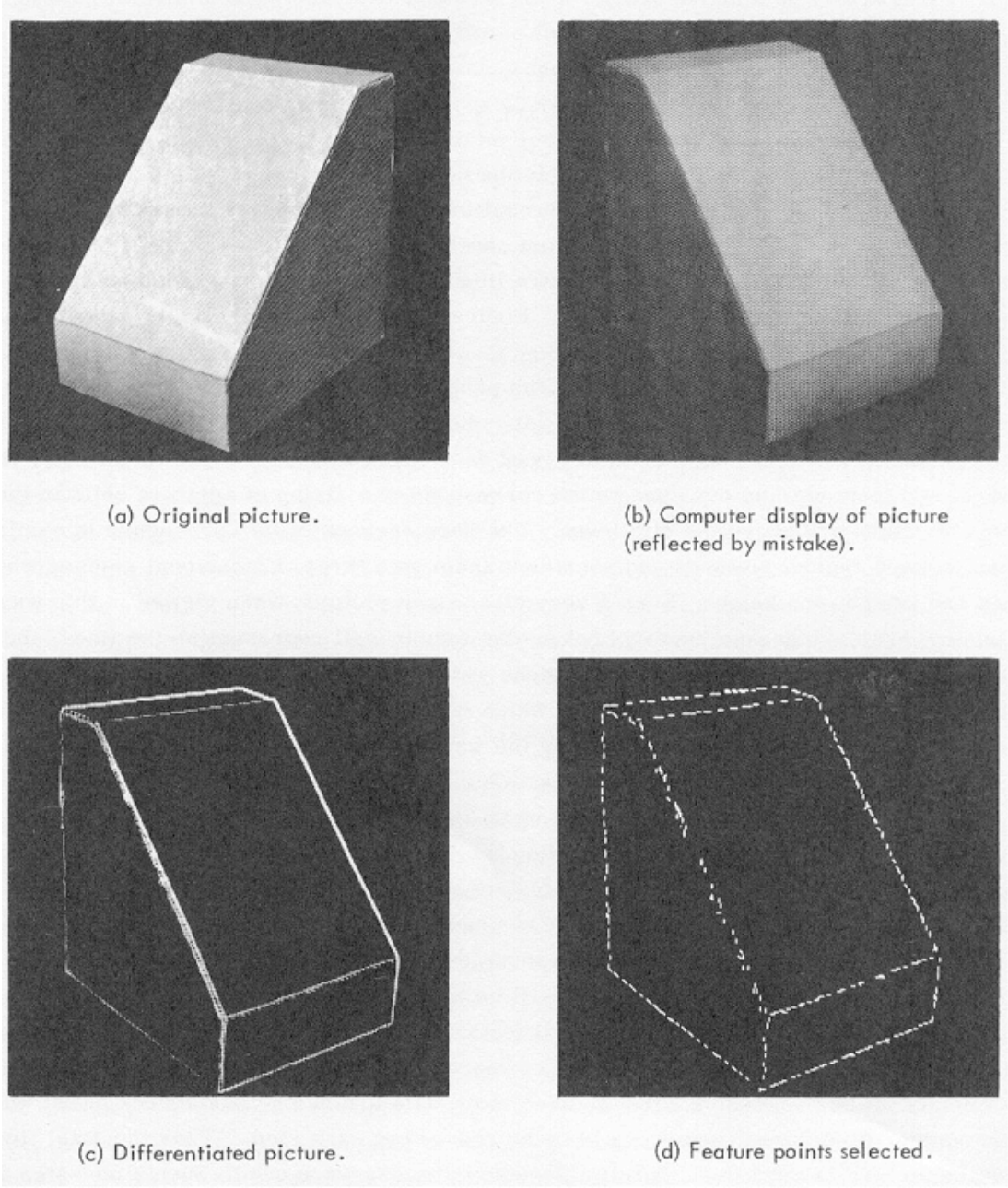




Hubel & Wiesel, 1959

Block world

Larry Roberts, 1963



(a) Original picture.

(b) Computer display of picture (reflected by mistake).

(c) Differentiated picture.

(d) Feature points selected.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

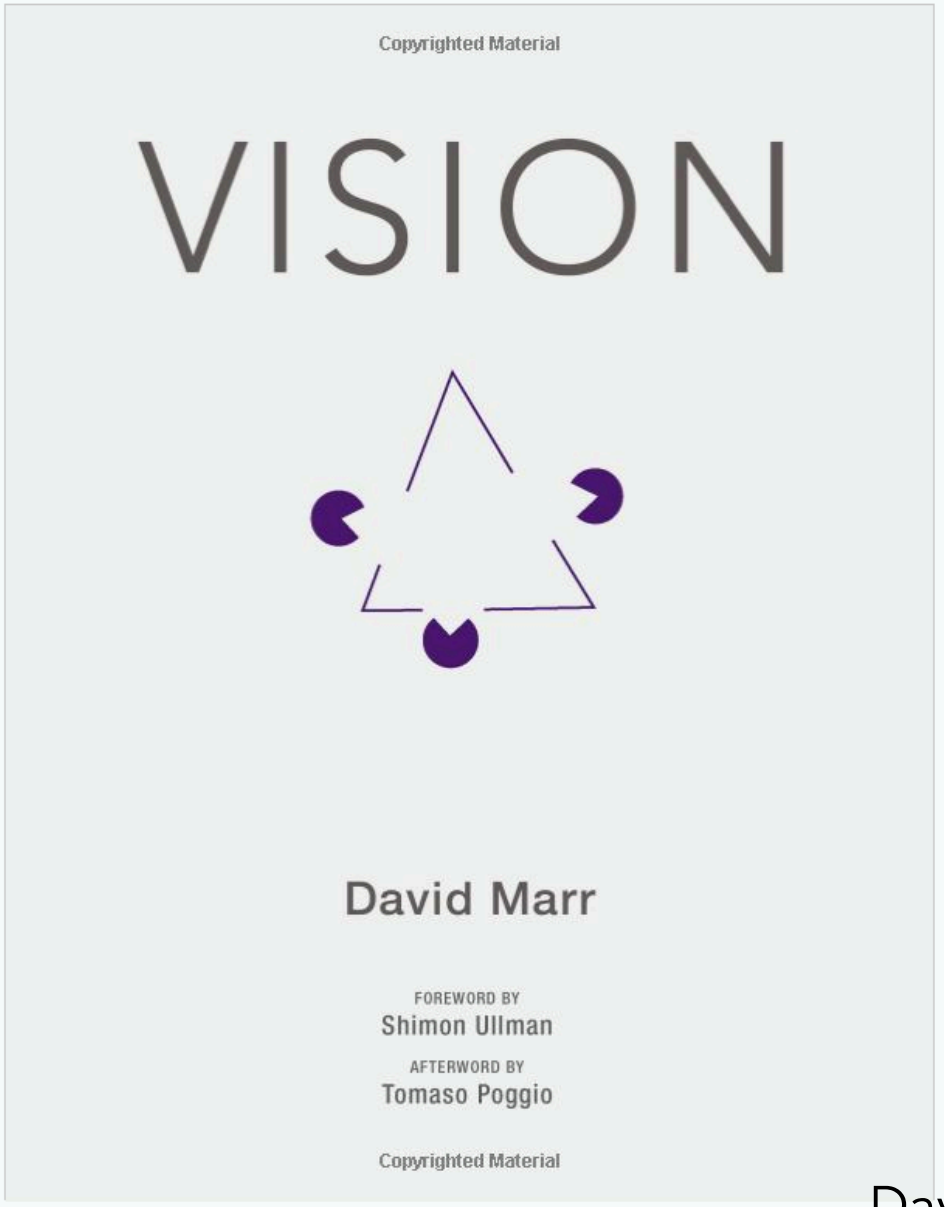
PROJECT MAC

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

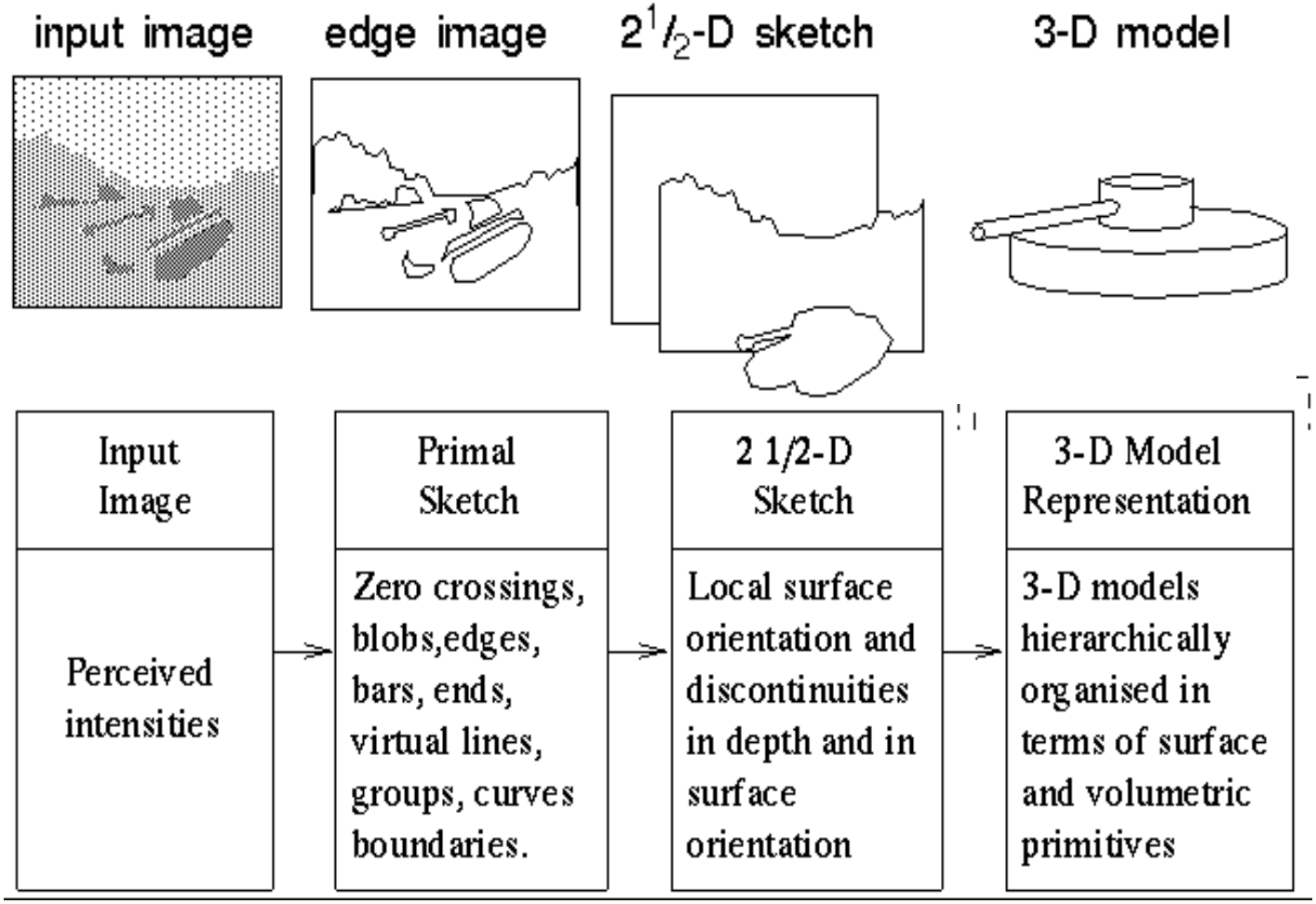
THE SUMMER VISION PROJECT

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".



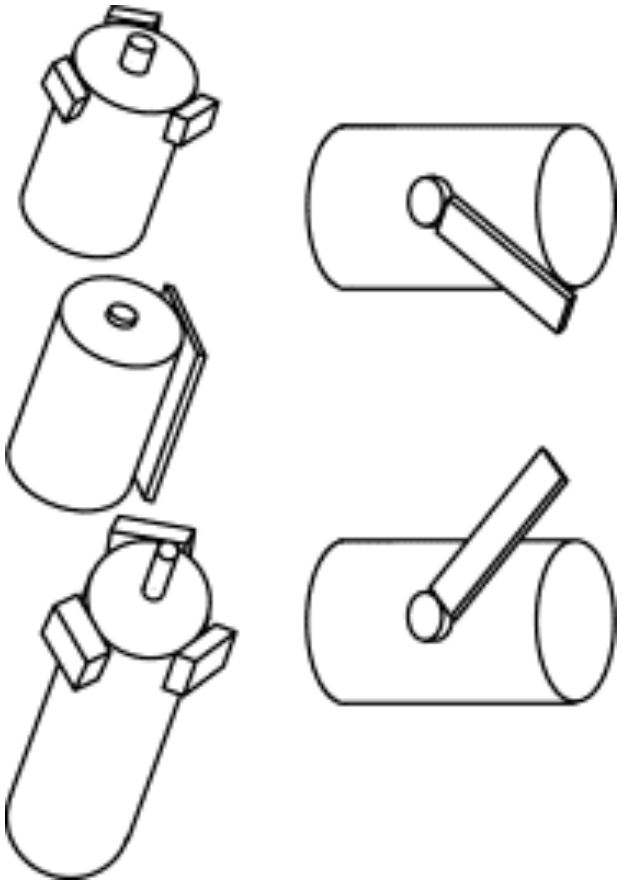
David Marr, 1970s



Stages of Visual Representation, David Marr, 1970s

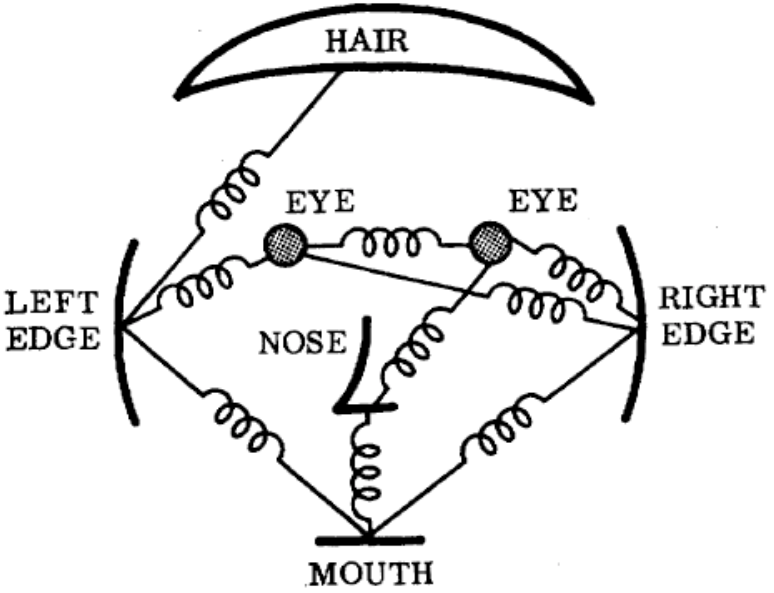
- Generalized Cylinder

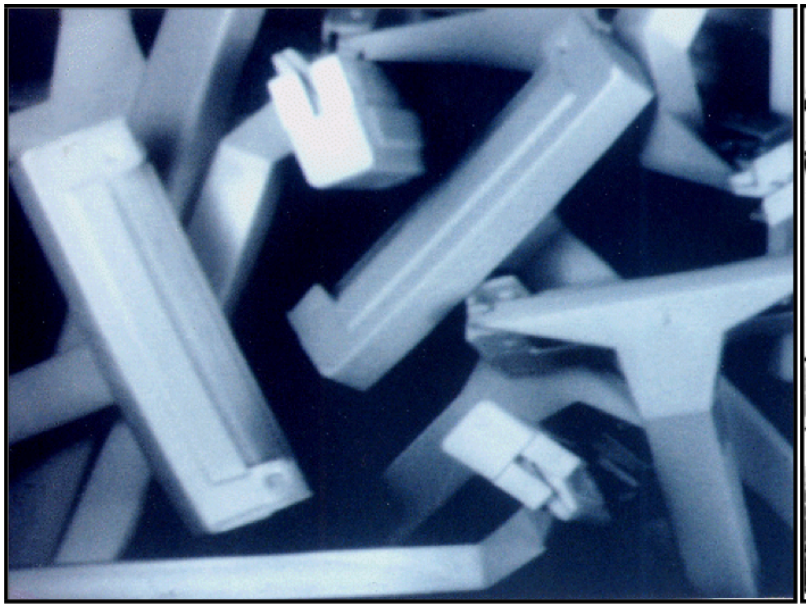
Brooks & Binford, 1979

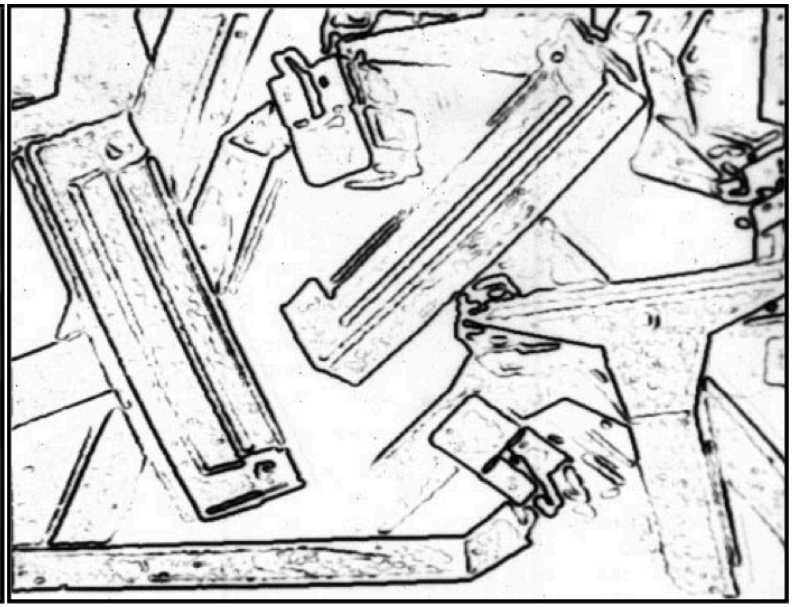


- Pictorial Structure

Fischler and Elschlager, 1973







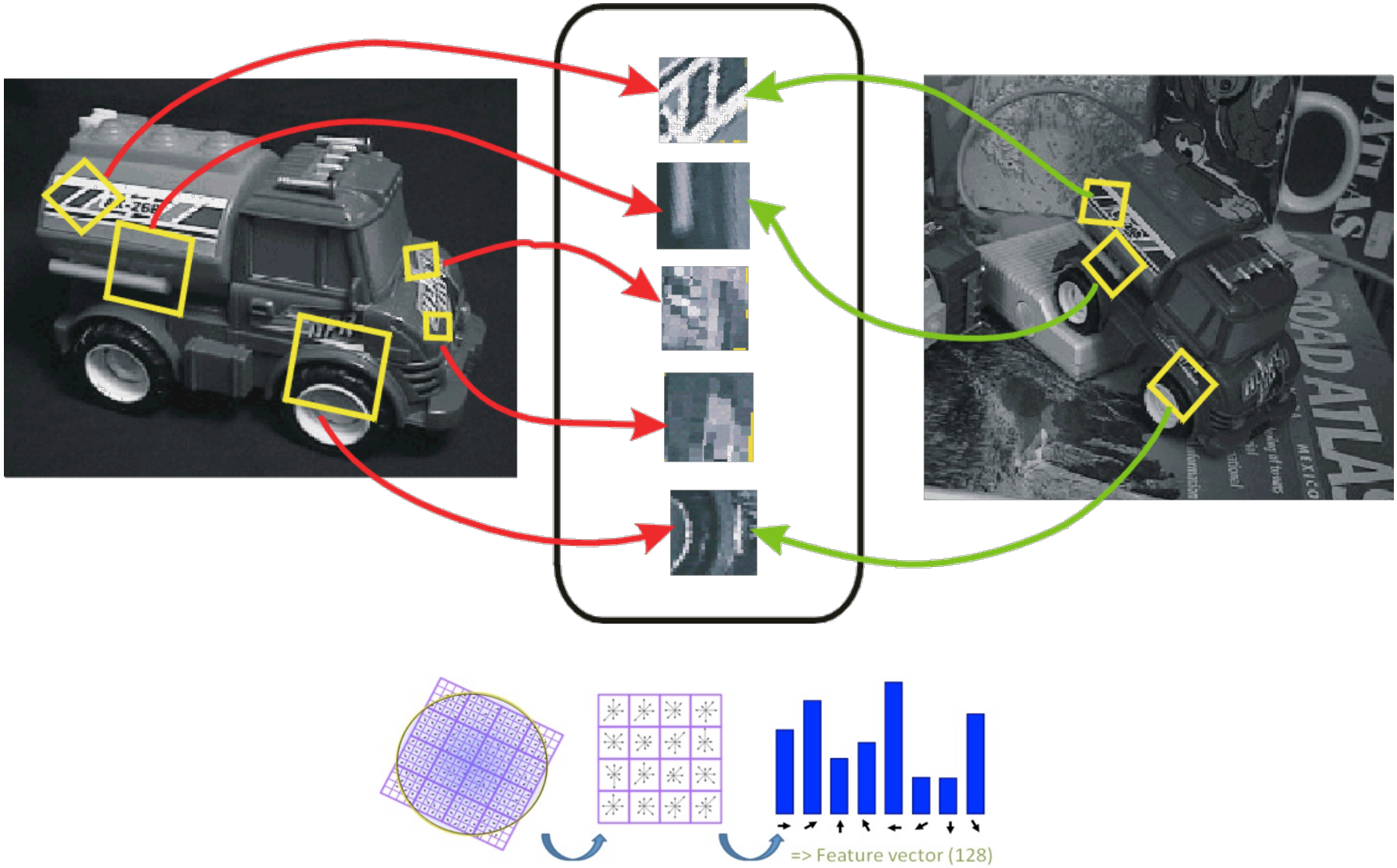
David Lowe, 1987

Normalized Cut
(Shi & Malik, 1997)

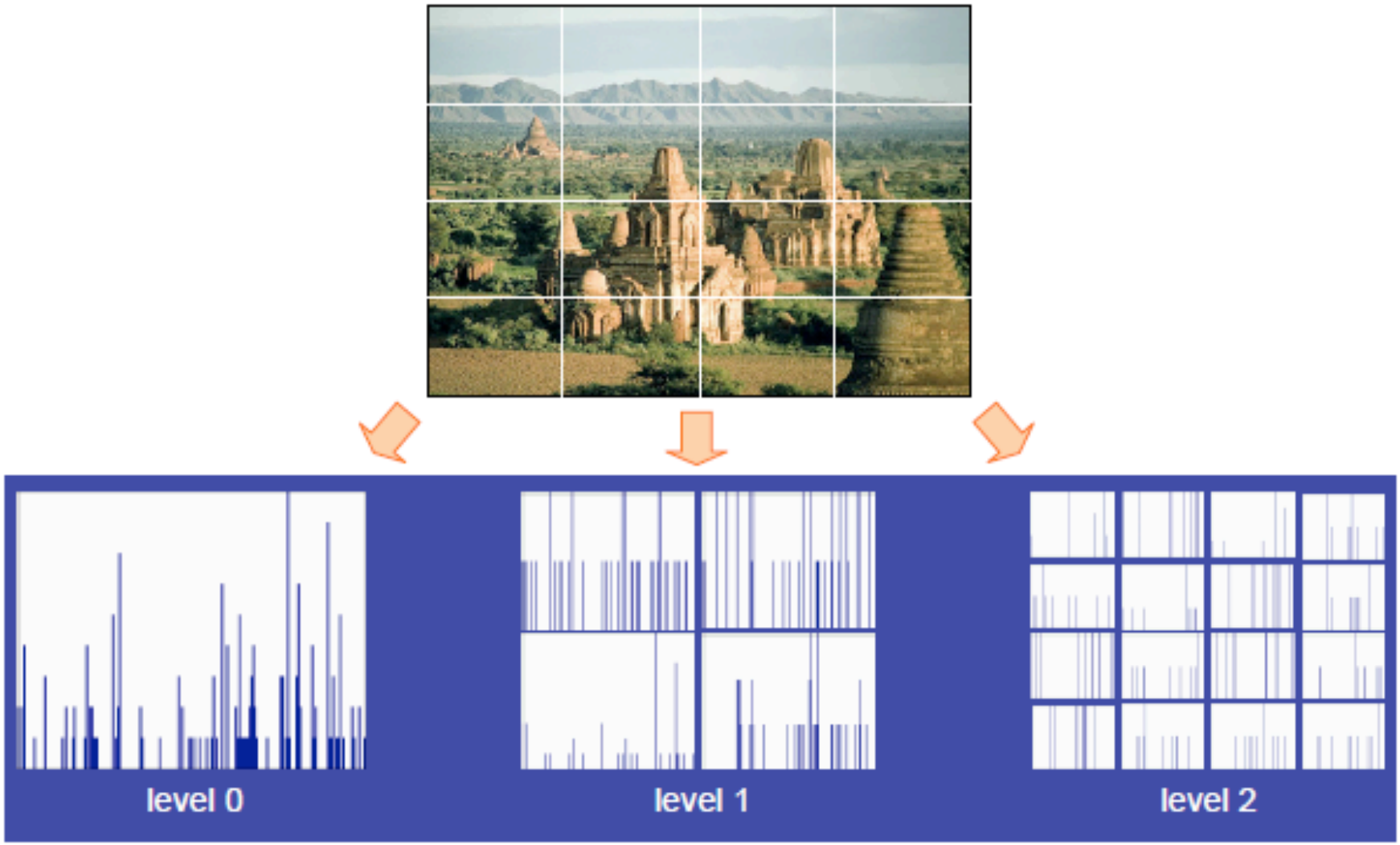




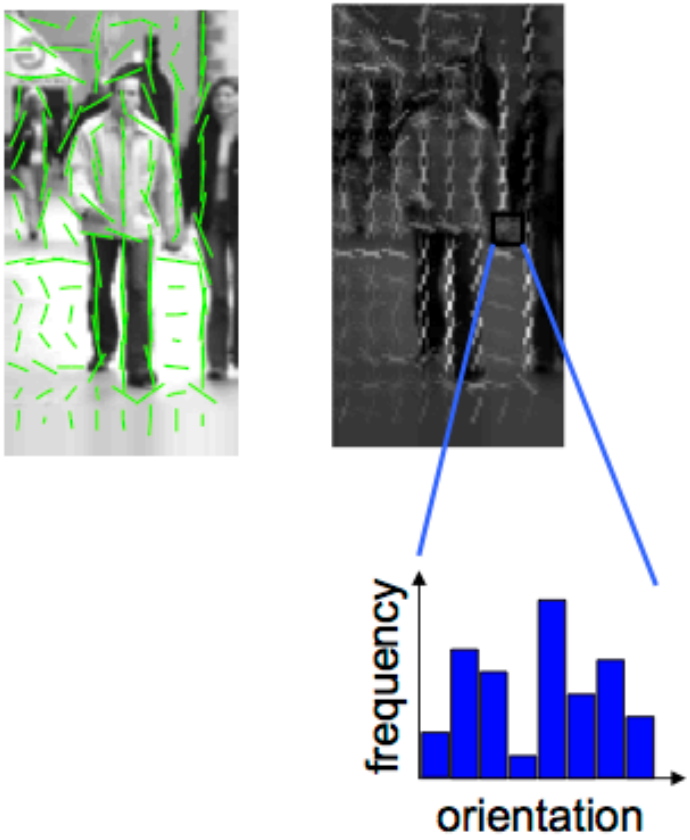
Face Detection, Viola & Jones, 2001



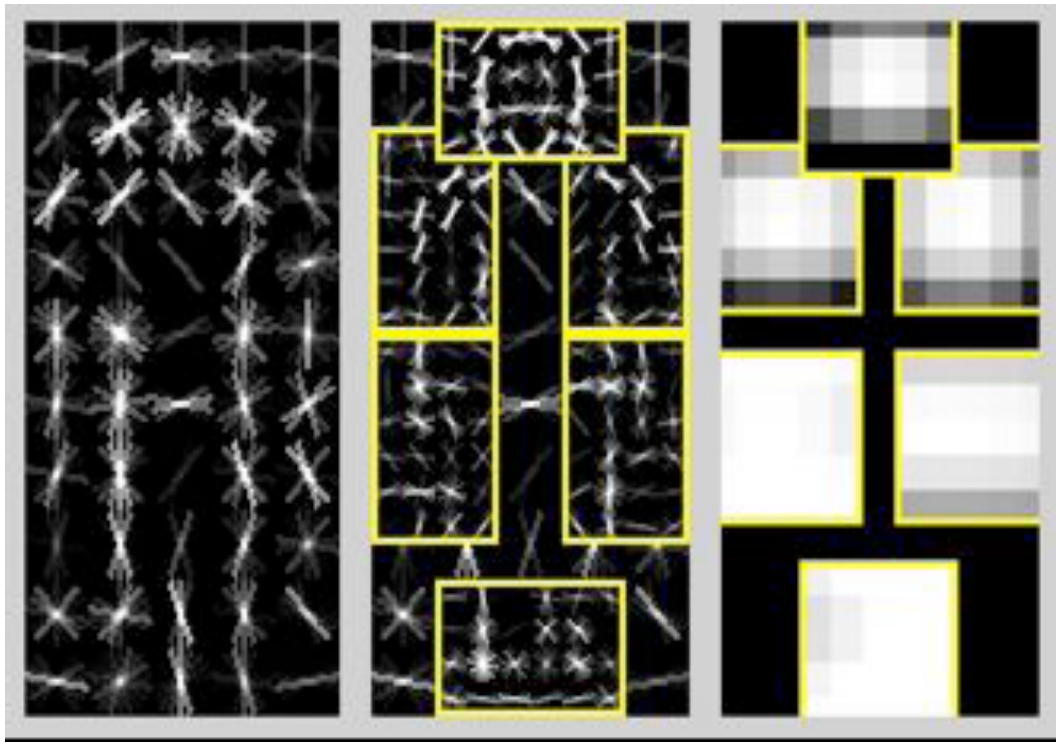
"SIFT" & Object Recognition, David Lowe, 1999



Spatial Pyramid Matching, Lazebnik, Schmid & Ponce, 2006



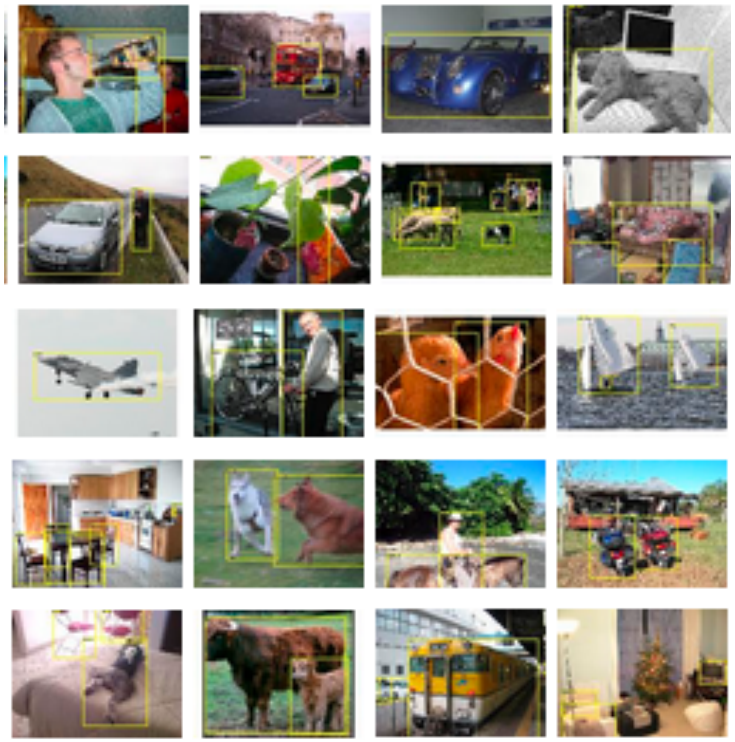
Histogram of Gradients (HoG)
Dalal & Triggs, 2005

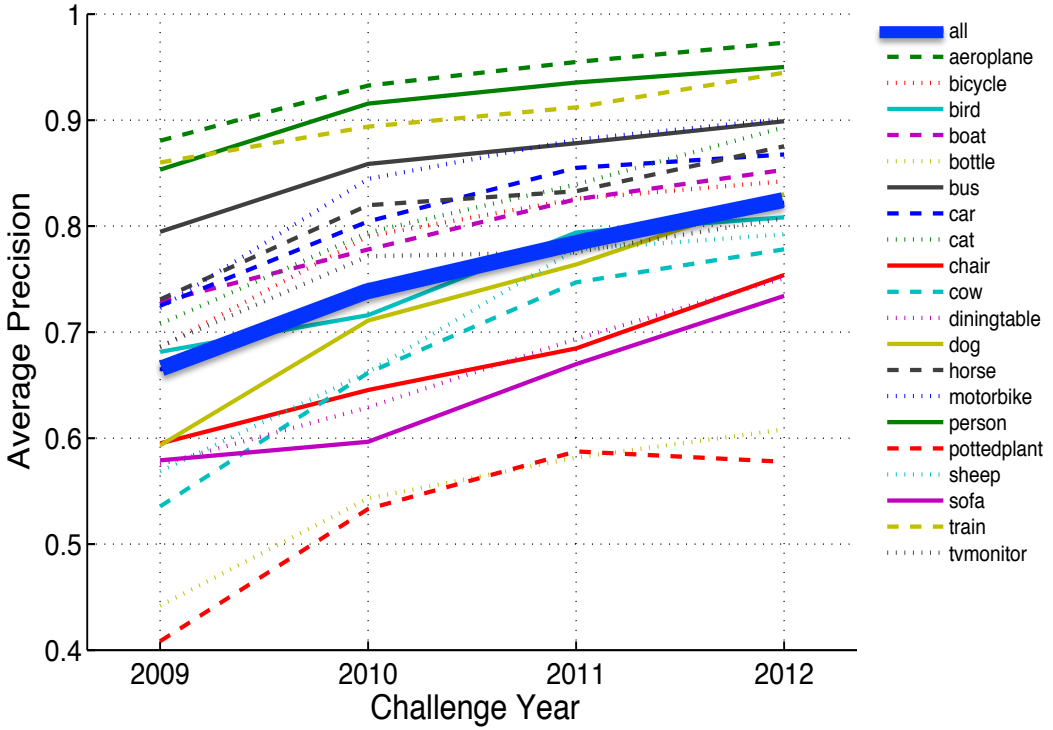


Deformable Part Model
Felzenswalb, McAllester, Ramanan, 2009

PASCAL Visual Object Challenge (20 object categories)

[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

Steel drum

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. arXiv, 2014

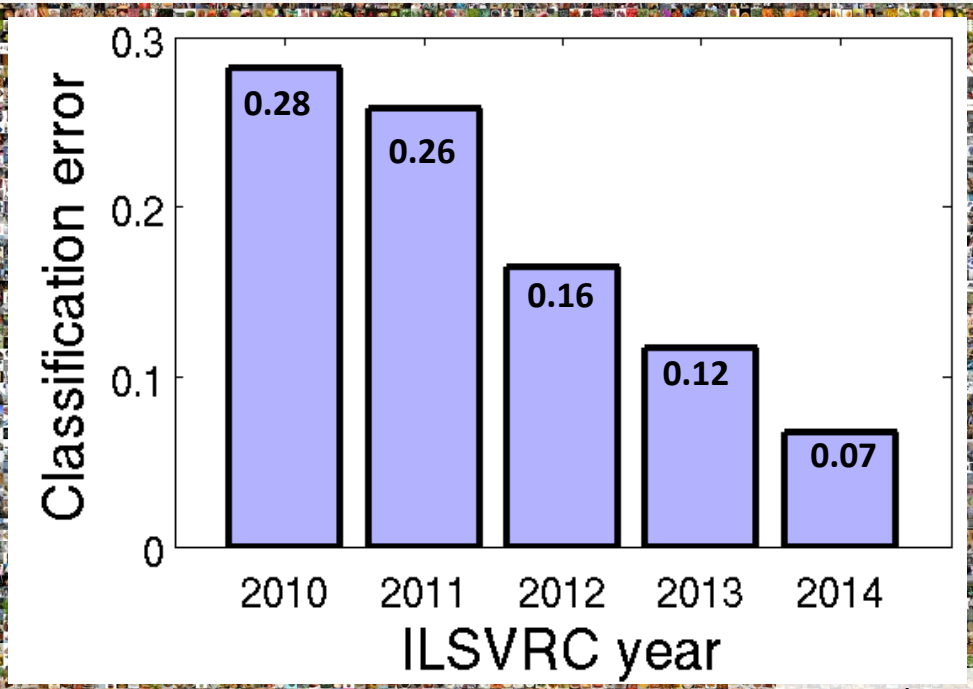
IMAGENET Large Scale Visual Recognition Challenge

Steel drum

The Image Classification Challenge:

1,000 object classes

1,431,167 images



Russakovsky et al. arXiv, 2014

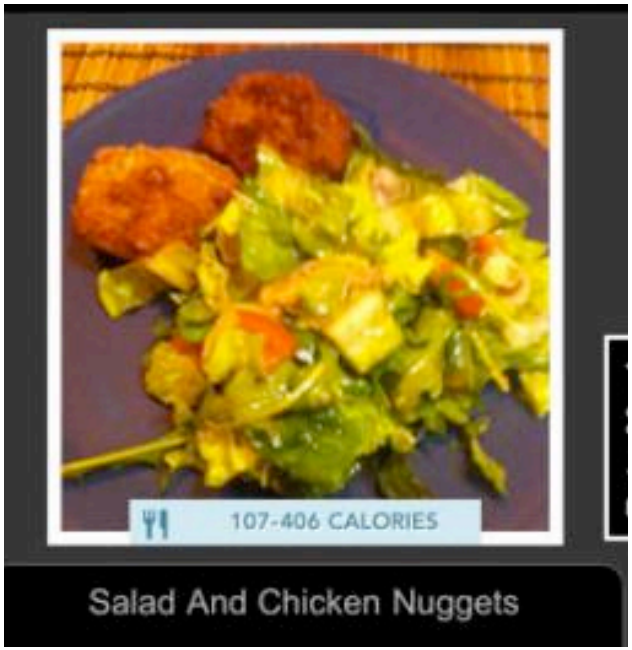
Today's agenda

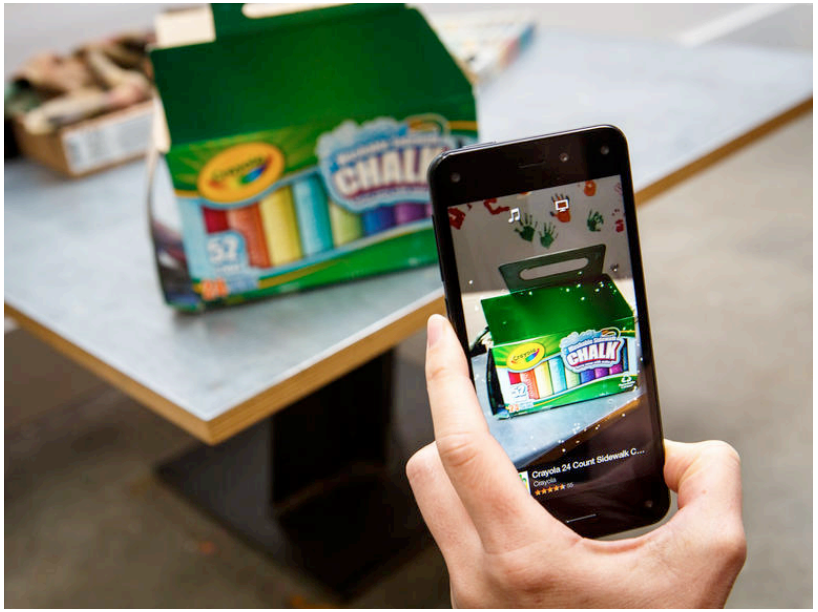
- A brief history of computer vision
- CS231n overview

CS231n focuses on one of the most important problems of visual recognition –

image classification


2:53 PM
Google goggles
Landmark
Golden Gate Bridge
Golden Gate Bridge
Web Results
Golden Gate Bridge - Wikipedia, the free encyclopedia
The Golden Gate Bridge by night, with part of downtown San Francisco ... Golden Gate Bridge is the most popular place to commit suicide in the United States ...
http://en.wikipedia.org/wiki/Golden_Gate_Bridge
Seacliff Webcam - Weather Seacliff, Golden Gate Bridge (Seacliff ...
Seacliff webcam (Golden Gate Bridge
Seacliff (United States, North Am
webcam Ocean Beach, San Fran
States ...


107-406 CALORIES
Salad And Chicken Nuggets


Crayola 24 Count Sidewalk C...

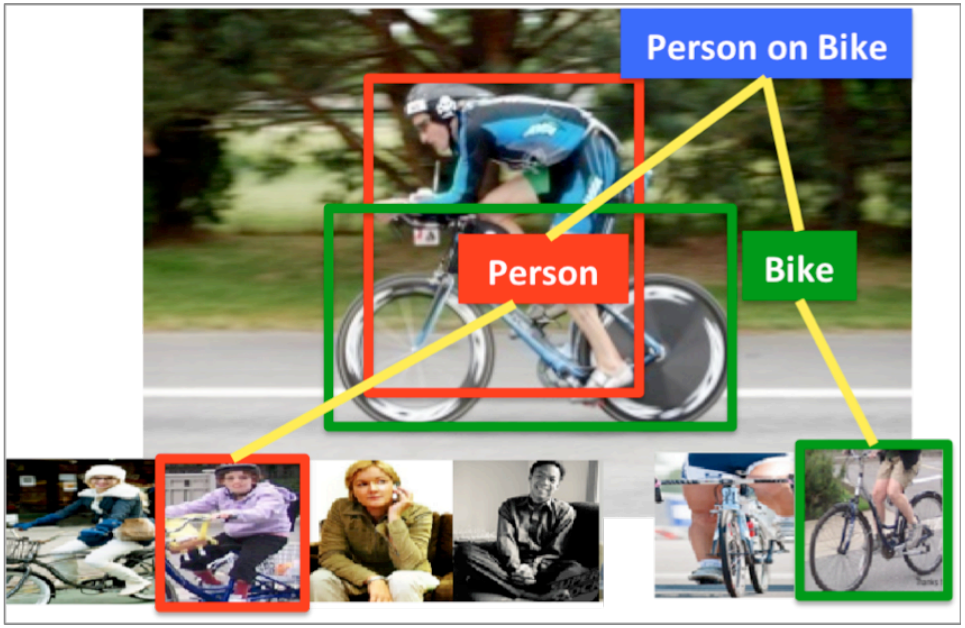

T-Mobile NL
16:06
Albums by Impala
Architecture
Time
Score

There is a number of visual recognition problems that are related to image classification, such as *object detection, image captioning*



- Object detection
- Action classification
- Image captioning
- …





Convolutional Neural Network (CNN) has become an important tool for object recognition

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC



Dense grid descriptor: HOG, LBP

Coding: local coordinate, super-vector

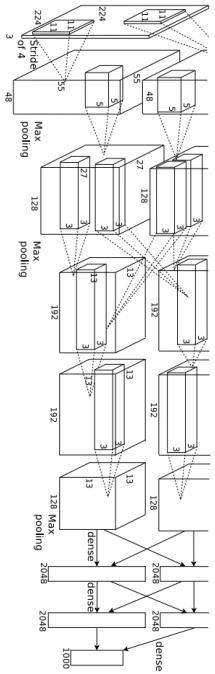
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

SuperVision



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet



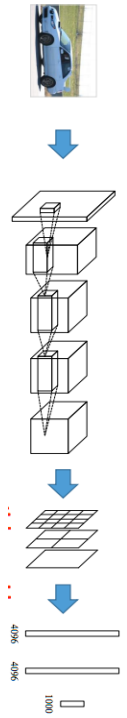
[Szegedy arxiv 2014]

VGG



[Simonyan arxiv 2014]

MSRA

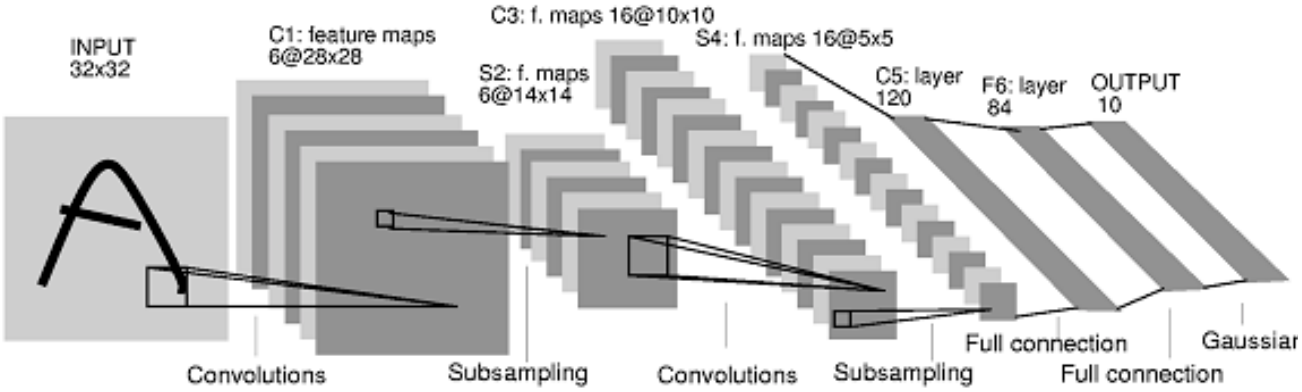


[He arxiv 2014]

Convolutional Neural Network (CNN)
is not invented overnight

1998

LeCun et al.



of transistors



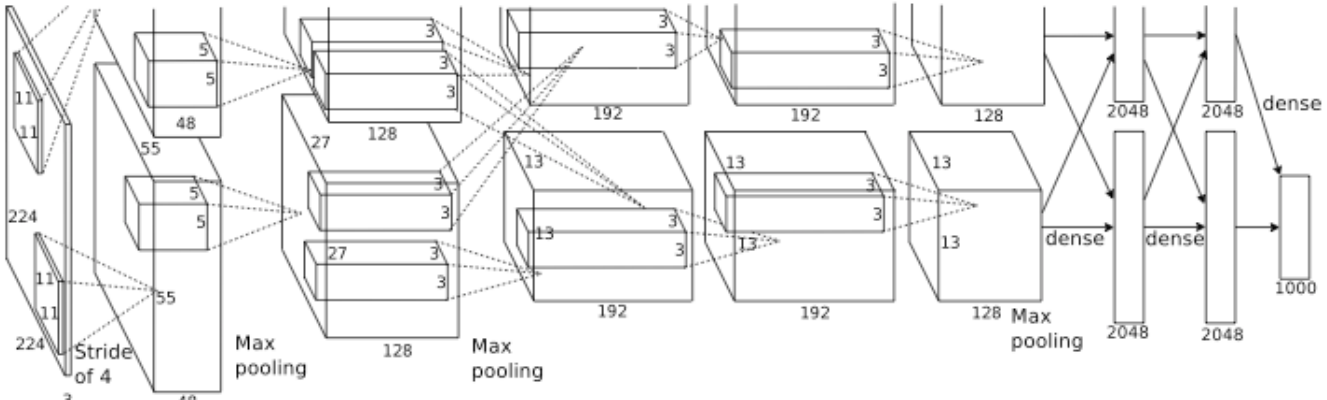
10^6

of pixels used in training

10^7 NIST

2012

Krizhevsky et al.



of transistors



10^9

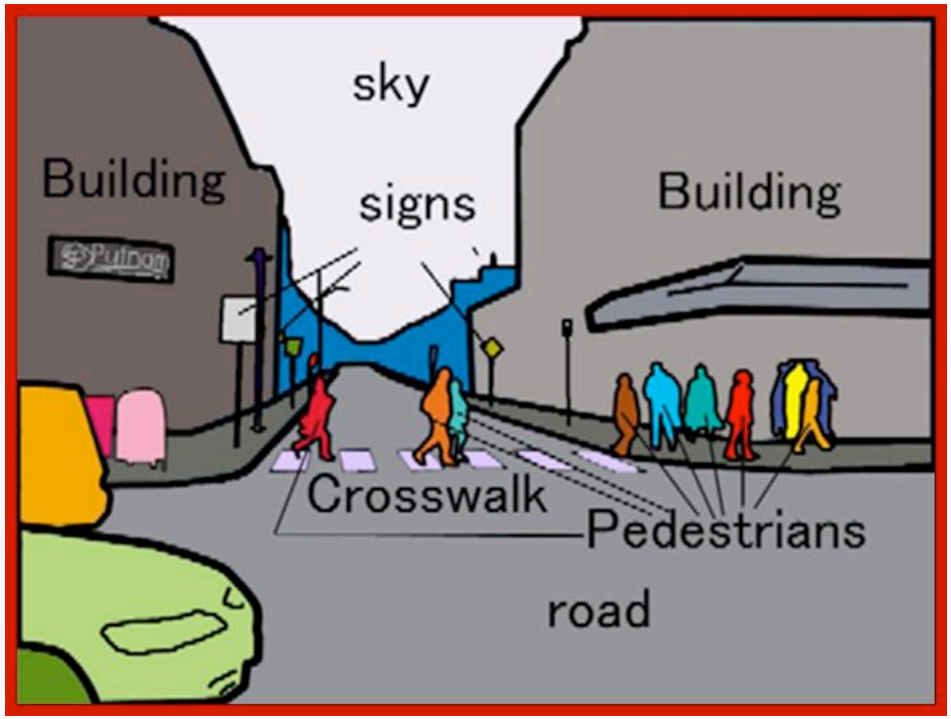
GPUs

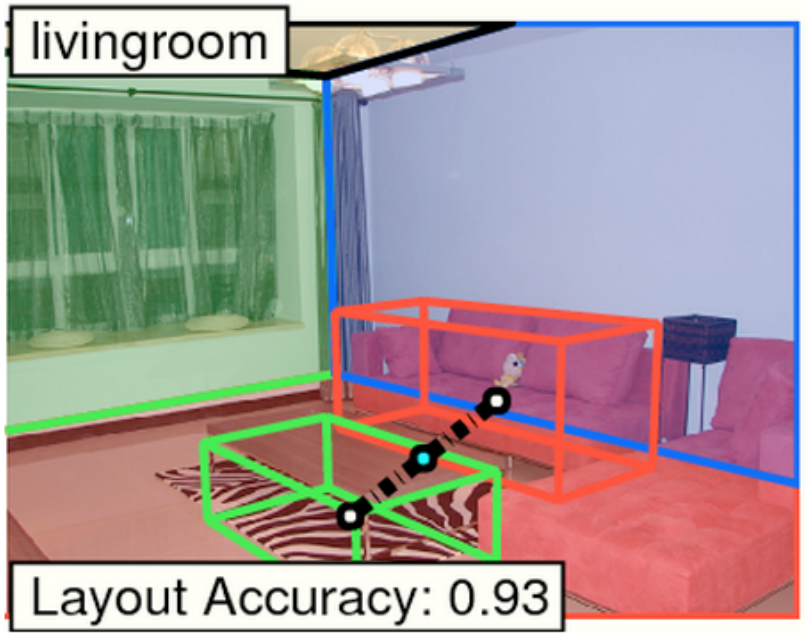


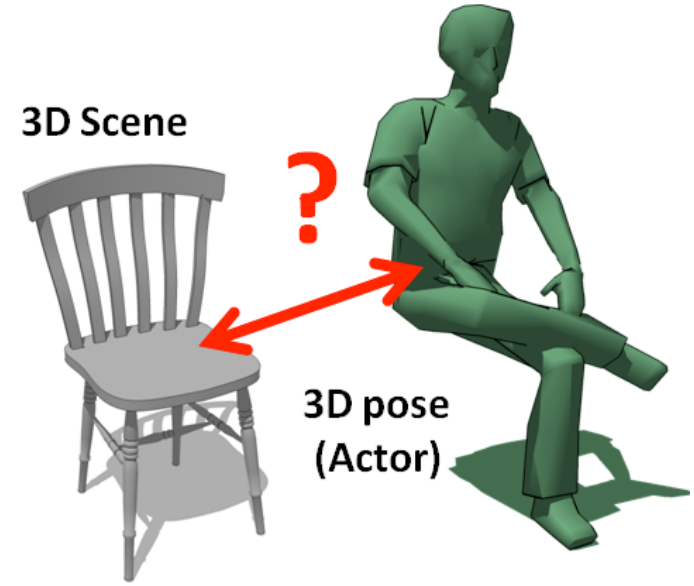
of pixels used in training

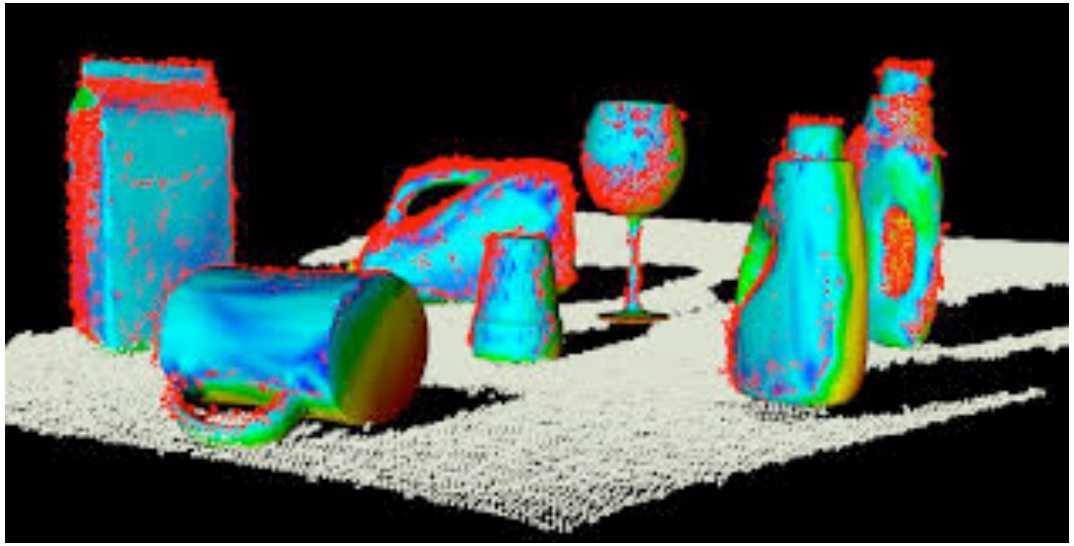
10^{14} IMAGENET

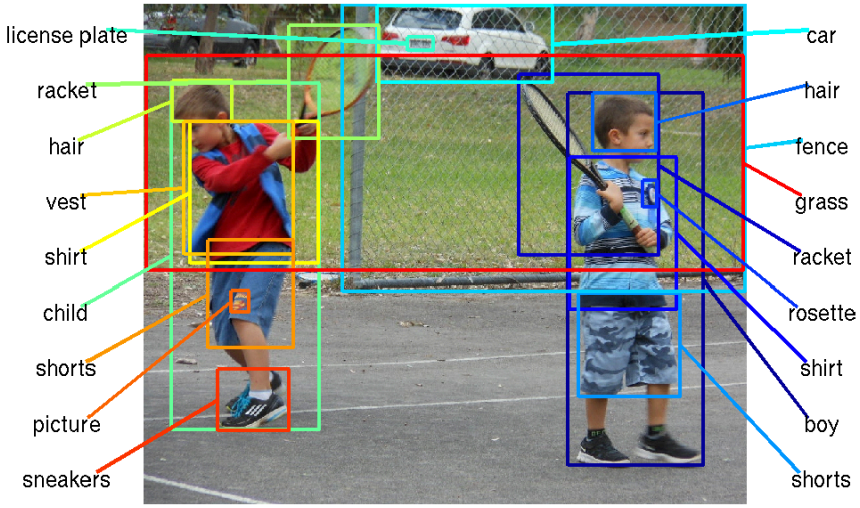
The quest for visual intelligence
goes far beyond object recognition…

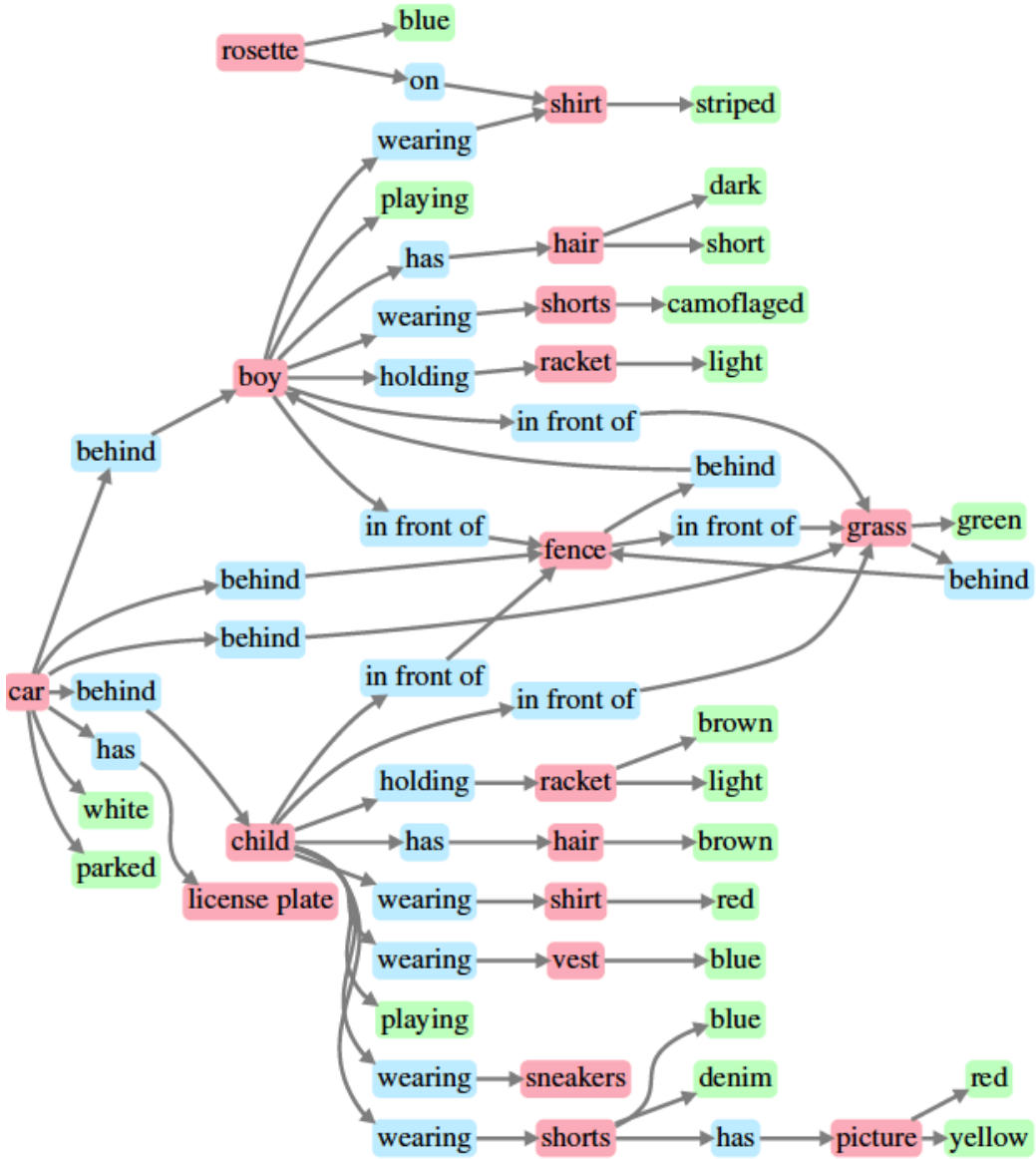

sky
Building
signs
Building
Crosswalk
Pedestrians
road


livingroom
Layout Accuracy: 0.93


3D Scene
?
3D pose
(Actor)




license plate
racket
hair
vest
shirt
child
shorts
picture
sneakers
car
hair
fence
grass
racket
rosette
shirt
boy
shorts


rosette
blue
on
shirt
striped
wearing
playing
dark
has
hair
short
wearing
shorts
camoflaged
boy
holding
racket
light
in front of
behind
behind
in front of
fence
in front of
grass
green
behind
behind
behind
car
behind
in front of
in front of
has
white
parked
child
brown
holding
racket
light
has
hair
brown
license plate
wearing
shirt
red
wearing
vest
blue
playing
blue
wearing
sneakers
denim
red
wearing
shorts
has
picture
yellow



PT = 500ms

Some kind of game or fight. Two groups of two men? The foregound pair looked like one was getting a fist in the face. Outdoors seemed like because i have an impression of grass and maybe lines on the grass? That would be why I think perhaps a game, rough game though, more like rugby than football because they pairs weren't in pads and helmets, though I did get the impression of similar clothing. maybe some trees? in the background. (Subject: SM)

Fei-Fei, Iyer, Koch, Perona, *JoV*, 2007

Andrej Karpathy blog

The state of Computer Vision and AI: we are really, really far.

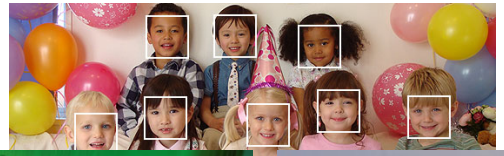
Oct 22, 2012

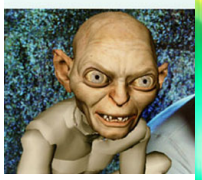


The picture above is funny.

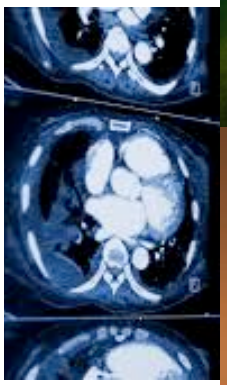
But for me it is also one of those examples that make me sad about the outlook for AI and for Computer Vision. What would it take for a computer to understand this image as you or I do? I challenge you to think explicitly of all the pieces of knowledge that have to fall in place for it to make sense. Here is my short attempt:

- You recognize it is an image of a bunch of people and you understand they are in a hallway
- You recognize that there are 3 mirrors in the scene so some of those people are "fake" replicas from different viewpoints.
- You recognize Obama from the few pixels that make up his face. It helps that he is in his suit and that he is surrounded by other people with suits.
- You recognize that there's a person standing on a scale, even though the scale occupies only very few white pixels that blend with the background. But, you've used the person's pose and knowledge of how people interact with objects to figure it out.
- You recognize that Obama has his foot positioned just slightly on top of the scale. Notice the language I'm using: It is in terms of the 3D structure of the scene, not the position of the leg in the 2D coordinate system of the image.
- You know how physics works: Obama is leaning in on the scale, which applies a force on it. Scale measures force that is applied on it, that's how it works => it will over-estimate the weight of the person standing on it.
- The person measuring his weight is not aware of Obama doing this. You derive this because you know his pose, you understand that the field of view of a person is finite, and you understand that he is not very likely to sense the slight push of Obama's foot.
- You understand that people are self-conscious about their weight. You also understand that he is reading off the scale measurement, and that shortly the over-estimated weight will confuse him because it will probably be much higher than what he expects. In other words, you reason about implications of the events that are about to unfold seconds after this photo was taken, and especially about the thoughts and how they will develop inside people's heads. You also reason about what pieces of information are available to people.
- There are people in the back who find the person's imminent confusion funny. In other words you are reasoning about state of mind of people, and their view of the state of mind of another person. That's getting frighteningly meta.
- Finally, the fact that the perpetrator here is the president makes it maybe even a little more funnier. You understand what actions are more or less likely to be undertaken by different people based on their status and identity.

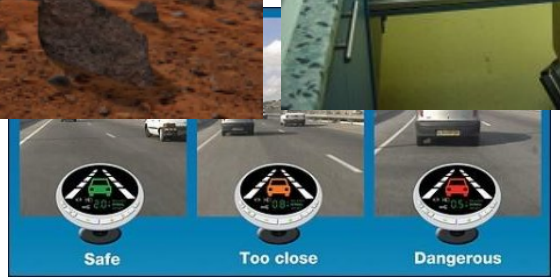



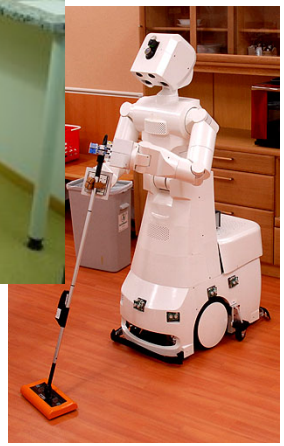













Computer Vision Technology Can Better Our Lives

Who we are

- Instructors: Prof. Fei-Fei Li & Andrej Karpathy
- Teaching Assistants
  - Justin Johnson, Ph.D. candidate, CS
  - Yuke Zhu, master candidate, CS
  - TBA
- Keeping in touch:
  - cs231n-winter1415-staff@lists.stanford.edu
  - Piazza
  - Twitter

Grading policy

- 3 Problem Sets: 15% x 3 = 45%
- Midterm Exam: 15%
- Final Course Project: 40%
  - Milestone: 5%
  - Final write-up: 35%
  - Bonus points for exceptional poster presentation
- Late policy
  - 7 free late days – use them in your ways
  - Afterwards, 25% off per day late
  - Not accepted after 3 late days per PS
  - Does not apply to Final Course Project
- Collaboration policy
  - Read the student code book, understand what is 'collaboration' and what is 'academic infraction'

Pre-requisite

- Proficiency in Python, familiarity with C/C++
  - All class assignments will be in Python (and use numpy), but some of the deep learning libraries we may look at later in the class are written in C++.
  - A Python tutorial available on course website
- College Calculus, Linear Algebra
- Equivalent knowledge of CS229 (Machine Learning)
  - We will be formulating cost functions, taking derivatives and performing optimization with gradient descent.

Syllabus

- Go to website…

http://vision.stanford.edu/teaching/cs231n/index.html

References

- Hubel, David H., and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." The Journal of physiology 160.1 (1962): 106. [PDF]
- Roberts, Lawrence Gilman. "Machine Perception of Three-dimensional Solids." Diss. Massachusetts Institute of Technology, 1963. [PDF]
- Marr, David. "Vision." The MIT Press, 1982. [PDF]
- Brooks, Rodney A., and Creiner, Russell and Binford, Thomas O. "The ACRONYM model-based vision system. " In Proceedings of the 6th International Joint Conference on Artificial Intelligence (1979): 105-113. [PDF]
- Fischler, Martin A., and Robert A. Elschlager. "The representation and matching of pictorial structures." IEEE Transactions on Computers 22.1 (1973): 67-92. [PDF]
- Lowe, David G., "Three-dimensional object recognition from single two-dimensional images," Artificial Intelligence, 31, 3 (1987), pp. 355-395. [PDF]
- Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.8 (2000): 888-905. [PDF]
- Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1. IEEE, 2001. [PDF]
- Lowe, David G. "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision 60.2 (2004): 91-110. [PDF]
- Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, 2006. [PDF]

- Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005. [PDF]
- Felzenszwalb, Pedro, David McAllester, and Deva Ramanan. "A discriminatively trained, multiscale, deformable part model." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008 [PDF]
- Everingham, Mark, et al. "The pascal visual object classes (VOC) challenge." International Journal of Computer Vision 88.2 (2010): 303-338. [PDF]
- Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009. [PDF]
- Russakovsky, Olga, et al. "Imagenet Large Scale Visual Recognition Challenge." arXiv:1409.0575. [PDF]
- Lin, Yuanqing, et al. "Large-scale image classification: fast feature extraction and SVM training." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011. [PDF]
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012. [PDF]
- Szegedy, Christian, et al. "Going deeper with convolutions." arXiv preprint arXiv:1409.4842 (2014). [PDF]
- Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014). [PDF]
- He, Kaiming, et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." arXiv preprint arXiv:1406.4729 (2014). [PDF]
- LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324. [PDF]
- Fei-Fei, Li, et al. "What do we perceive in a glance of a real-world scene?." Journal of vision 7.1 (2007): 10. [PDF]