CS6501: Deep Learning for Visual Recognition

Introduction

Today’s Class

Who am I?

What is Computer Vision?

What is Visual Recognition?

Why is Visual Recognition Hard?

Python + Numpy + Matplotlib and Manipulating Images

Questions

About Me

Vicente

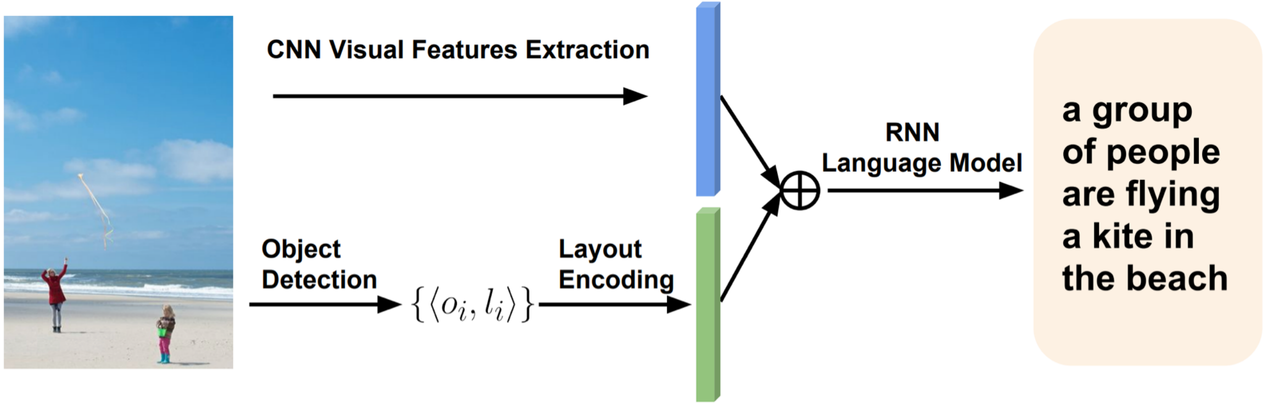
About Me

Assistant Professor, 2016 - Now	University of Virginia
Visiting Professor, 2019	Adobe Research
Visiting Researcher, 2015 - 2016	Allen Institute for Artificial Intelligence
MS, PhD in CS, 2009-2015	The University of North Carolina at Chapel Hill; Stony Brook University

... also spent time at:

Google Microsoft ebay

Describing Images with Language



NEW! Obj2Text: Generating Visually Descriptive Language from Object Layouts
Xuwang Yin, Vicente Ordonez.
Empirical Methods in Natural Language Processing. **EMNLP 2017**. Copenhagen, Denmark. September 2017.

Large Scale Retrieval and Generation of Image Descriptions
V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daume III, A.C. Berg, Y. Choi, T.L. Berg.
International Journal of Computer Vision. **IJCV 2015**. [August 2016 Issue]. [pdf] [link] [bibtex]

Im2Text: Describing Images Using 1 Million Captioned Photographs
Vicente Ordonez, Girish Kulkarni, Tamara L. Berg.
Advances in Neural Information Processing Systems. **NIPS 2011**. Granada, Spain. December 2011.

Naming Objects

Superordinates: animal, vertebrate
Basic Level: bird
Entry Level: bird
Subordinates: American robin

Superordinates: animal, vertebrate
Basic Level: bird
Entry Level: penguin
Subordinates: Chinstrap penguin

From Large Scale Image Categorization to Entry-Level Categories
Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg.
IEEE International Conference on Computer Vision. **ICCV 2013**. Sydney, Australia. December 2013.

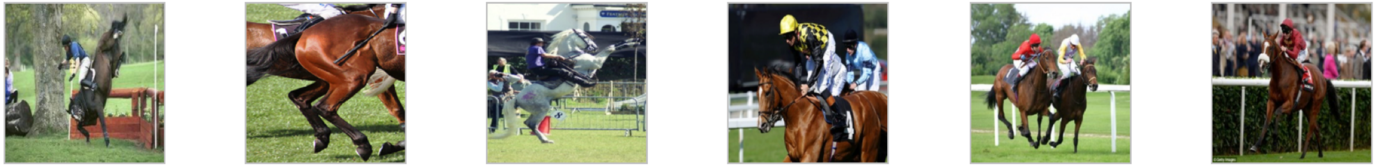
Learning to Name Objects
Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg.
Communications of the ACM. March 2016 (Vol. 59, No. 3). *(~Research Highlight)*

Predicting Entry-Level Categories
Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg.
International Journal of Computer Vision - Marr Prize Special Issue. **IJCV 2015**.

Recognizing Commonly Uncommon Situations

Query

Similar in imSitu train set



Predicted situations

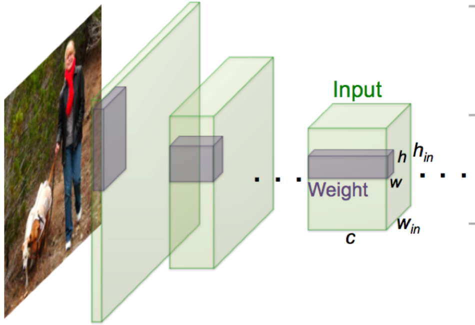
falling			
agent	source	goal	place
person	horse	land	outdoors

0.58372

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi.
Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2017**. Honolulu, Hawaii. July 2017.

http://imsitu.org/demo/

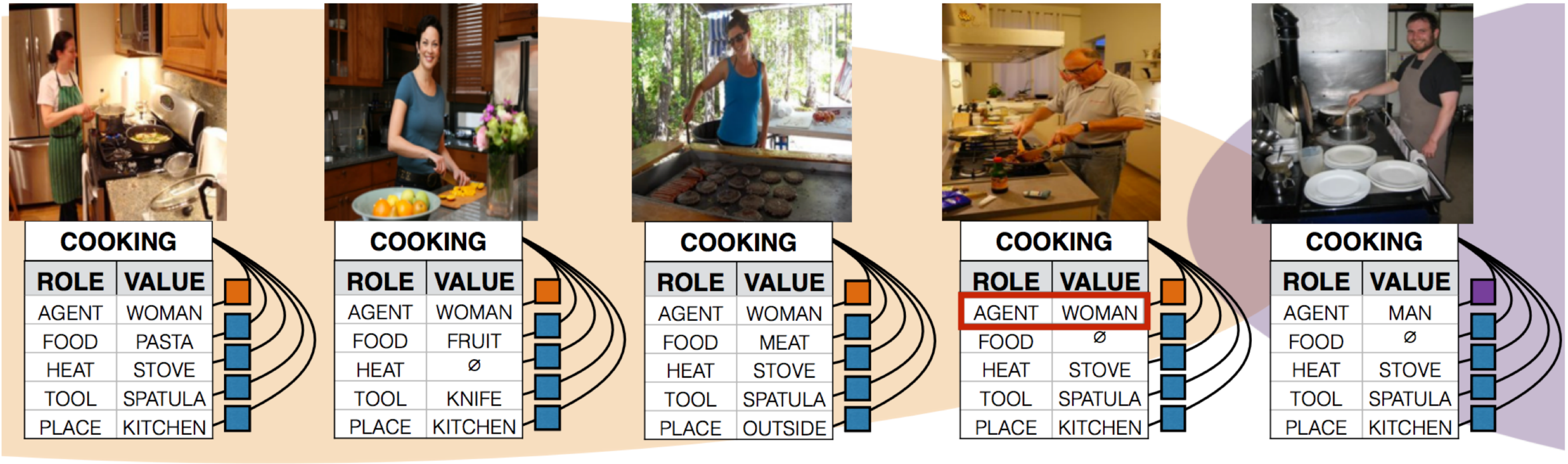
Accelerating Neural Networks: XNOR-Net



	Network Variations	Operations used in Convolution	Memory Saving (Inference)	Computation Saving (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution	Real-Value Inputs: 0.11 -0.21 ... -0.34 / -0.25 0.61 ... 0.52; Real-Value Weights: 0.12 -1.2 ... 0.41 / -0.2 0.5 ... 0.68	+ , − , ×	1x	1x	%56.7
Binary Weight	Real-Value Inputs: 0.11 -0.21 ... -0.34 / -0.25 0.61 ... 0.52; Binary Weights: 1 -1 ... 1 / -1 1 ... 1	+ , −	~32x	~2x	%56.8
BinaryWeight Binary Input (**XNOR-Net**)	Binary Inputs: 1 -1 ... -1 / -1 1 ... 1; Binary Weights: 1 -1 ... 1 / -1 1 ... 1	XNOR , bitcount	~32x	~58x	%44.2

XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks
Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, Ali Farhadi.
European Conference on Computer Vision. **ECCV 2016**. Amsterdam, The Netherlands. October 2016.

Removing Gender Bias from Situation Recognition

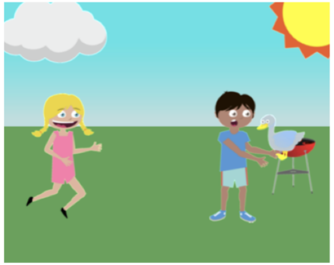


NEW! Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang.
Empirical Methods in Natural Language Processing. **EMNLP 2017**. Copenhagen, Denmark. September 2017.

Synthesizing Images from Text

Mike is surprised at the duck. The duck is standing on the grill. Jenny is running towards Mike and the duck.

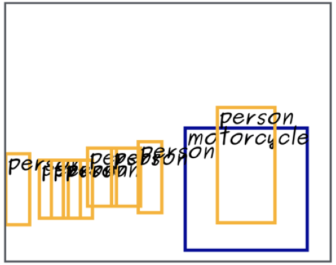
↓



Abstract Scene[1]

A guy on a motorcycle with some people watching.

↓



Object Layout[2]

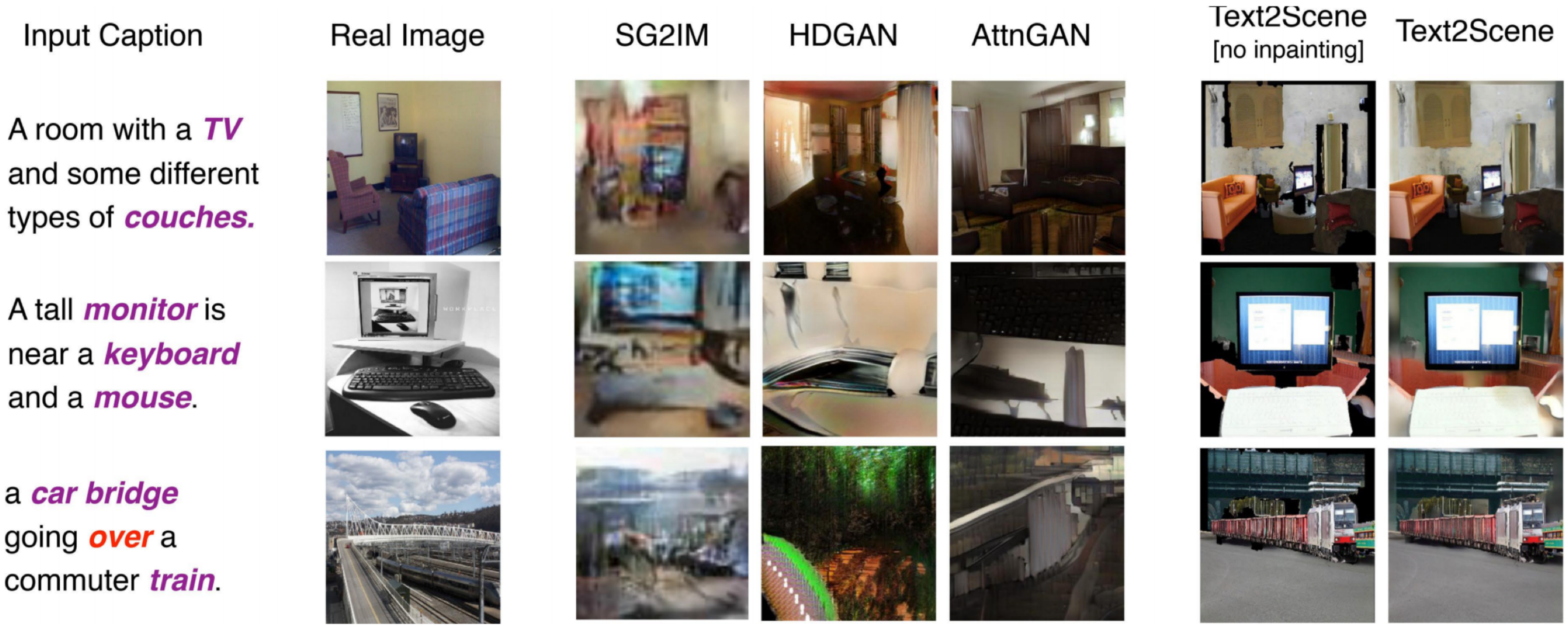
Several elephants walking together in a line near water.

↓

Synthetic Image[2]

NEW! Text2Scene: Generating Compositional Scenes from Textual Descriptions
Fuwen Tan, Song Feng, Vicente Ordonez.
Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2019**.
Long Beach, California. June 2019. [arxiv] [bibtex] *(~Oral presentation + Best Paper Finalist -- top 1% of submissions)*

Text-to-Image Synthesis: Text2Scene



What is Visual Recognition?

Make computers understand images and video

What kind of scene?

Where are the cars?

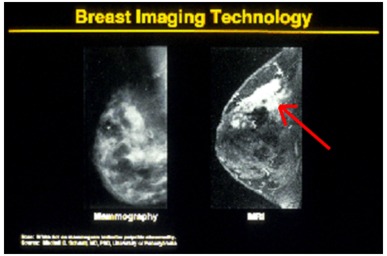
How far is the building?

…

Slide by James Hays.

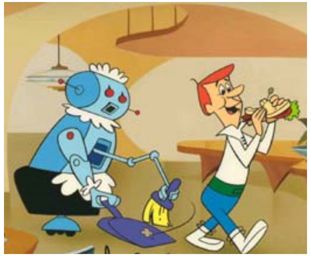
Why computer vision matters

Safety



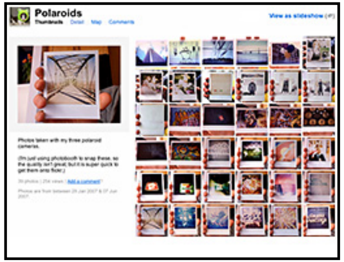
Health

Security



Comfort

Fun



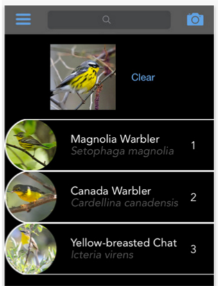
Access

Slide by James Hays.

kaggle

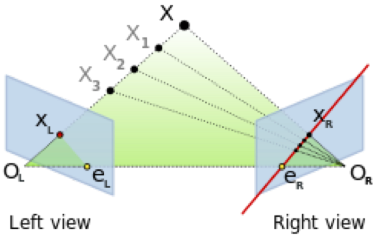
Create an algorithm to distinguish dogs from cats

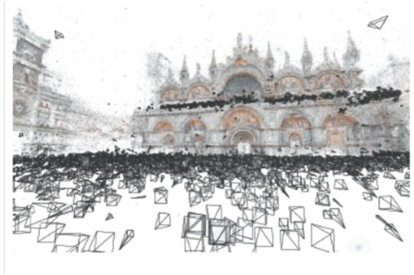
Birdsnap

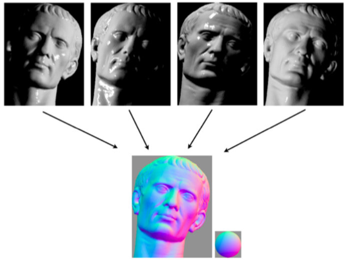


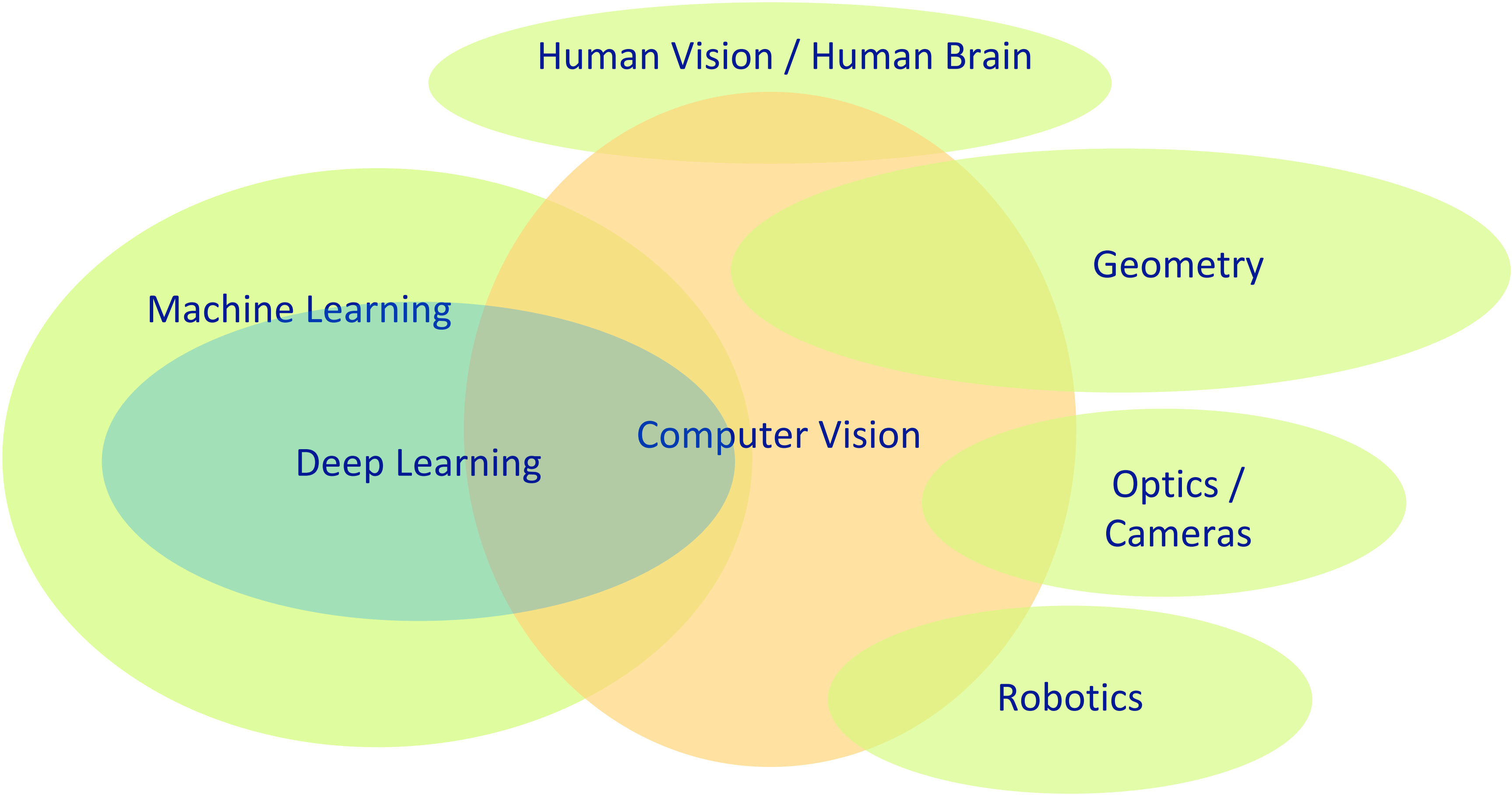
Face Detection in Cameras

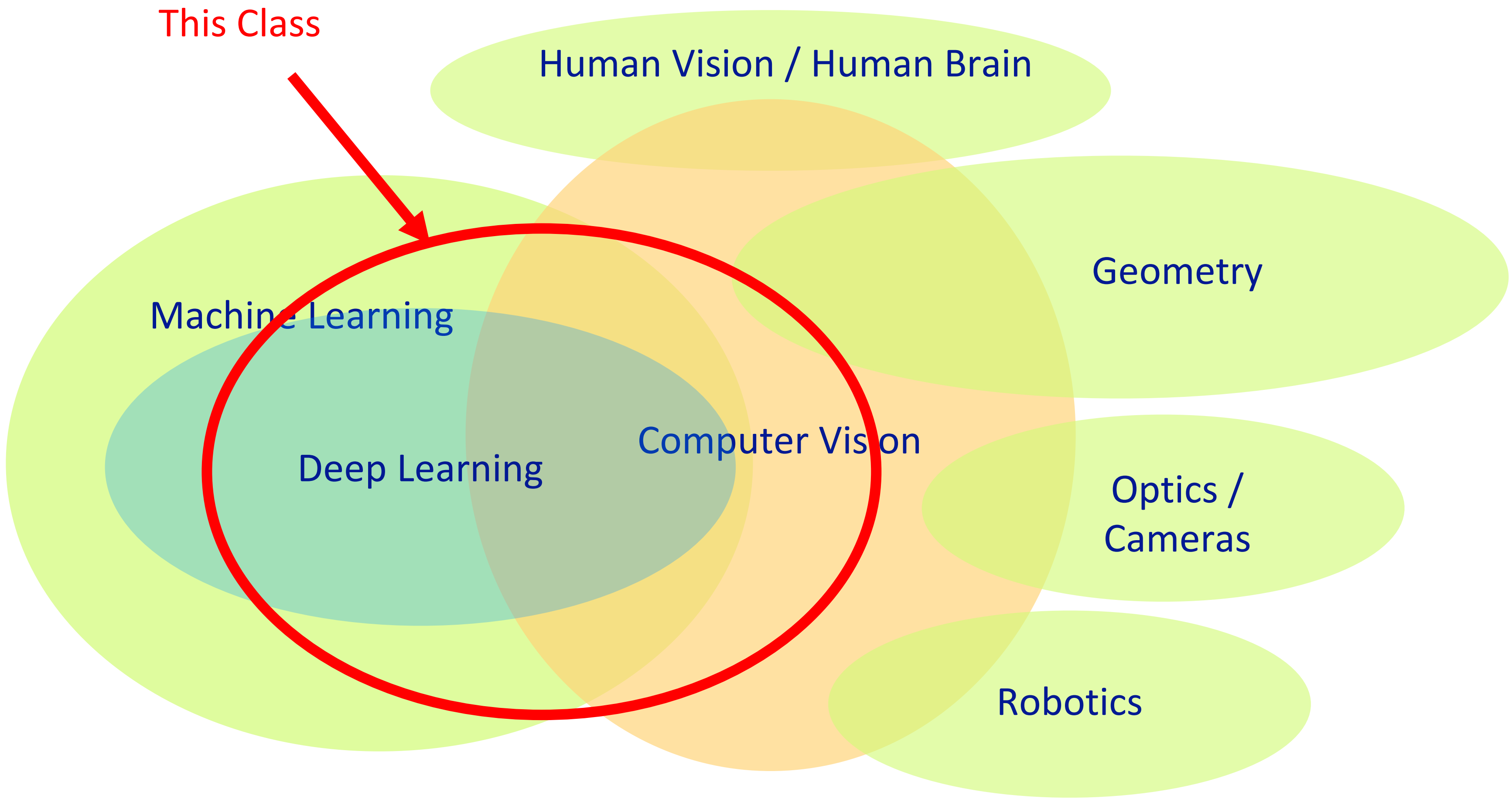
Computer Vision







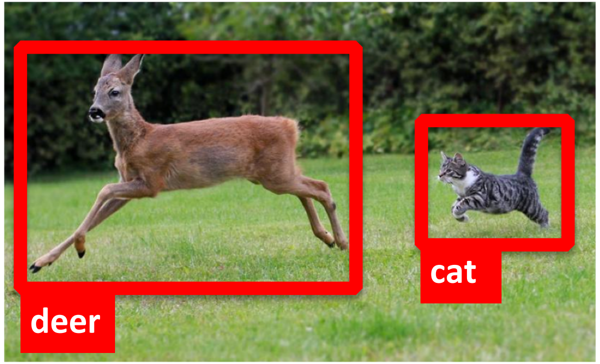

Human Vision / Human Brain
Geometry
Machine Learning
Computer Vision
Deep Learning
Optics /
Cameras
Robotics


This Class
Human Vision / Human Brain
Geometry
Machine Learning
Computer Vision
Deep Learning
Optics /
Cameras
Robotics

Relationship with Other Fields

- Computer Vision: Image → Knowledge

→



Relationship with Other Fields

- Image Processing: Image $\longrightarrow$ Image

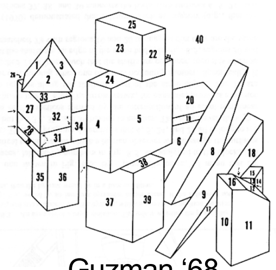
Relationship with Other Fields

- Computer Graphics: Knowledge ⟶ Image

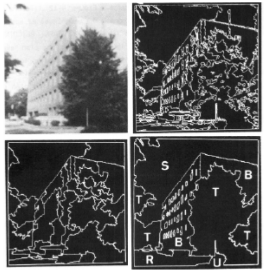
Vertices, Locations, Objects,
Shapes, Colors, Material properties,
Lighting settings, Camera settings, etc. ⟶

Ridiculously brief history of computer vision

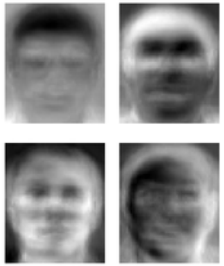
- 1966: Minsky assigns computer vision as an undergrad summer project
- 1960’s: interpretation of synthetic worlds
- 1970’s: some progress on interpreting selected images
- 1980’s: ANNs come and go; shift toward geometry and increased mathematical rigor
- 1990’s: face recognition; statistical analysis in vogue
- 2000’s: broader recognition; large annotated datasets available; video processing starts
- 2010’s: Deep learning with ConvNets
- 2030’s: ?



Guzman ‘68



Ohta Kanade ‘78



Turk and Pentland ‘91

How vision is used now

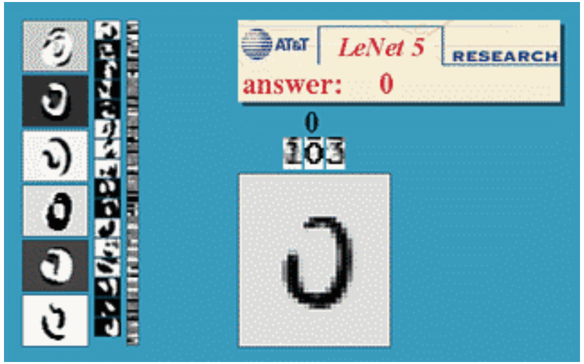
- Examples of real world applications

Some of the following slides by Steve Seitz

Optical character recognition (OCR)

Technology to convert scanned docs to text

- If you have a scanner, it probably came with OCR software



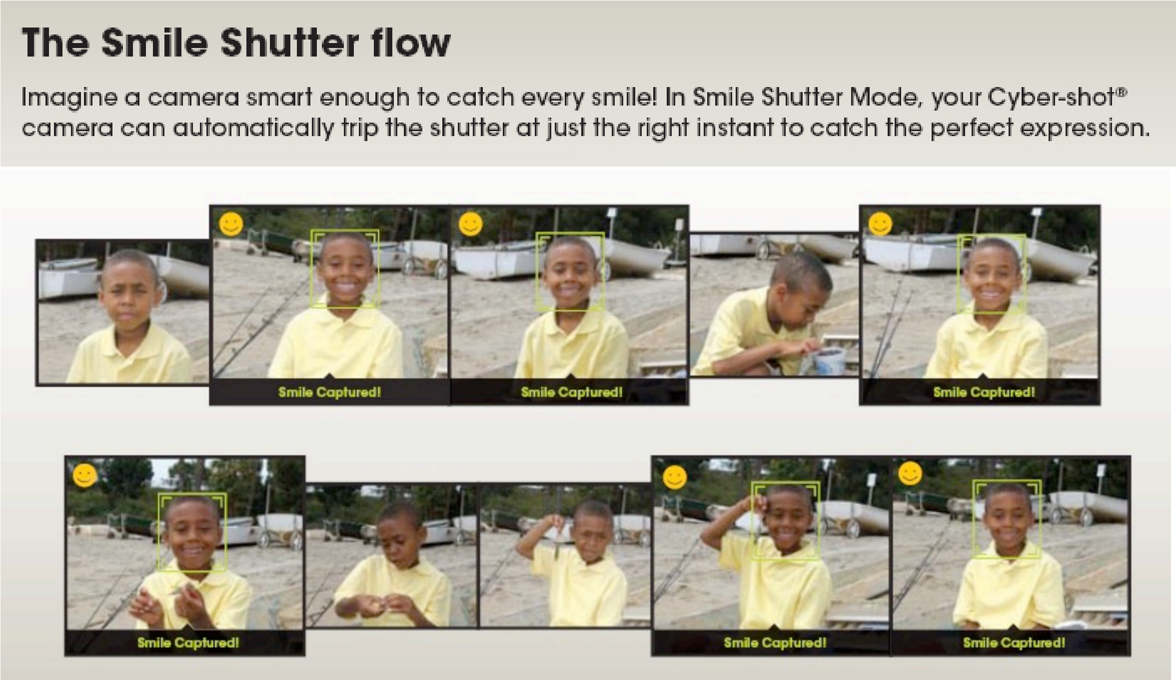
Digit recognition, AT&T labs
http://www.research.att.com/~yann/

License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

Face detection

- Digital cameras detect faces

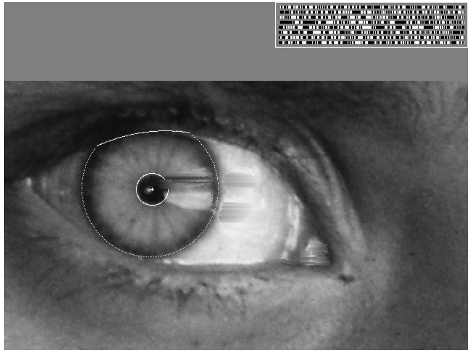
Smile detection

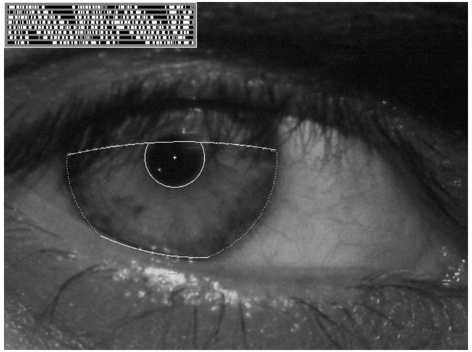


[Sony Cyber-shot® T70 Digital Still Camera](#)

Vision-based biometrics

"*How the Afghan Girl was Identified by Her Iris Patterns*" Read the story wikipedia





Login without a password…

Fingerprint scanners on
many new laptops,
other devices

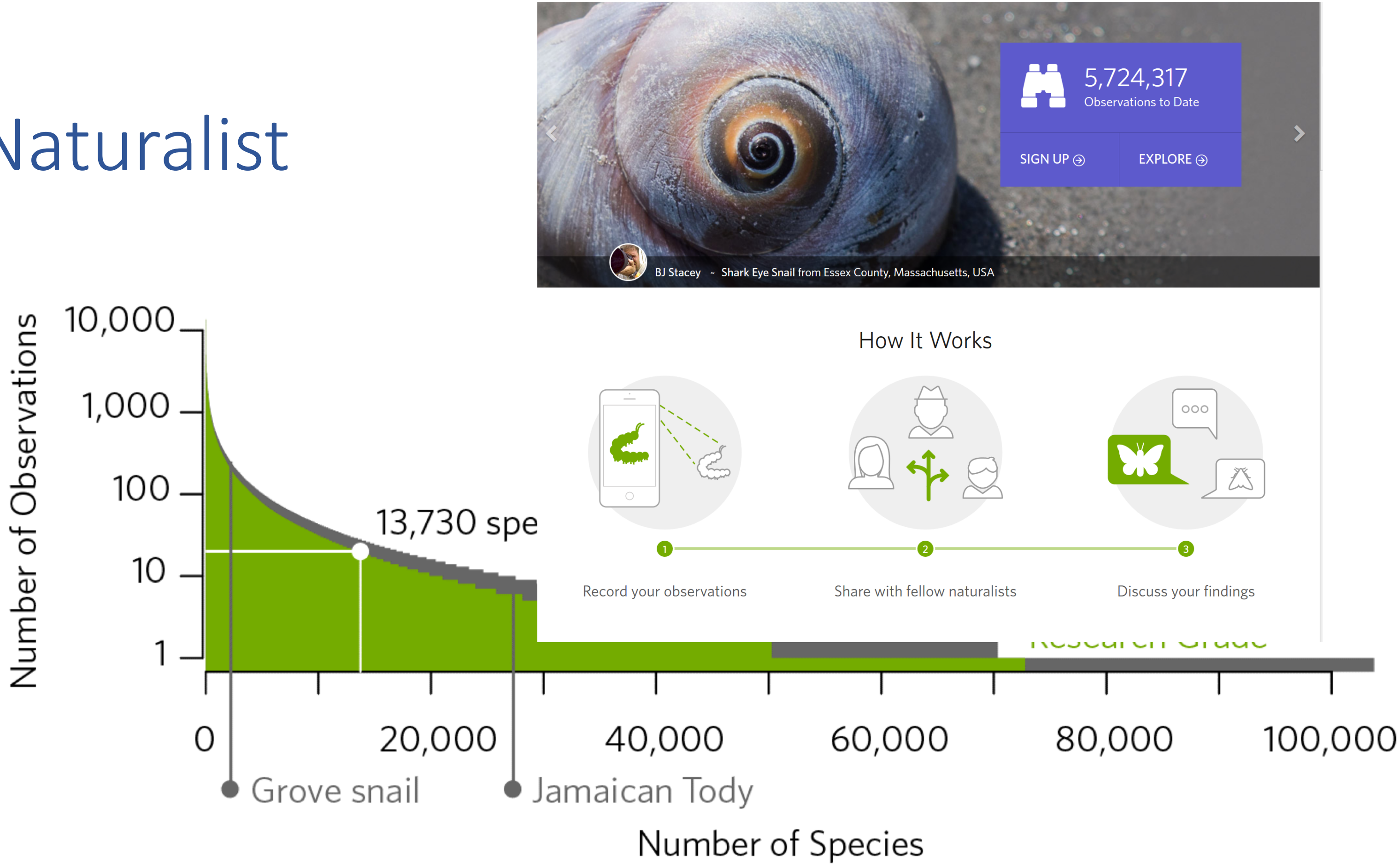
Face recognition systems now
beginning to appear more widely
http://www.sensiblevision.com/

Object recognition (in mobile phones)

[Point & Find](), [Nokia]()

[Google Goggles]()

iNaturalist



https://www.inaturalist.org/pages/computer_vi

Special effects: shape capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC

Special effects: motion capture

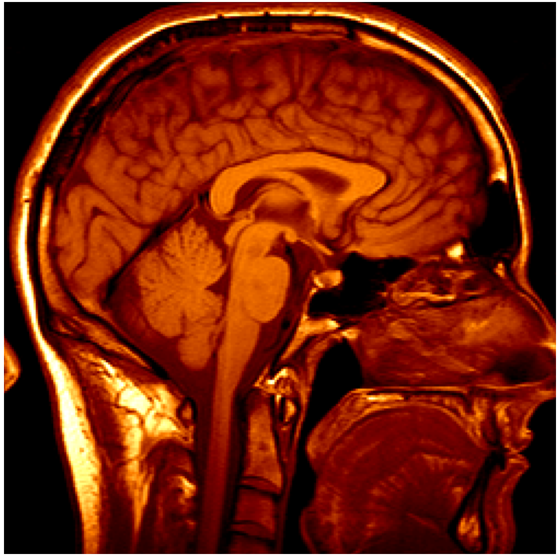
Pirates of the Carribean, Industrial Light and Magic

Sports

Sportvision first down line
Nice [explanation](#) on [www.howstuffworks.com](#)

http://www.sportvision.com/video.html

Medical Imaging



3D imaging
MRI, CT

Image guided surgery
Grimson et al., MIT

Slide content courtesy of Amnon Shashua

Smart cars

- Mobileye
  - Market Capitalization: 11 Billion dollars

Self-driving Cars e.g. Google's Waymo

Oct 9, 2010. "Google Cars Drive Themselves, in Traffic". *The New York Times*. John Markoff
June 24, 2011. "Nevada state law paves the way for driverless cars". *Financial Post*. Christine Dobby
Aug 9, 2011, "Human error blamed after Google's driverless car sparks five-vehicle crash". *The Star* (Toronto)

Ford acquires SAIPS for self-driving machine learning and computer vision tech

Posted Aug 16, 2016 by Darrell Etherington (@etherington)

Ford outlined a few of the ways it's aiming to ship driverless cars by 2021, and part of the plan involves acquisitions. CEO Mark Fields revealed at a press event in Palo Alto today that the automaker acquired SAIPS, an Israeli company focusing on machine learning and computer vision. It's also partnering exclusively with Nirenberg Neuroscience, to bring more "humanlike intelligence" to machine learning components of driverless car systems.

SAIPS' technology brings image and video processing algorithms, as well as deep learning tech focused on processing and classifying input signals, all key ingredients in the special sauce that makes up autonomous vehicle tech. This company's expertise should help with on-board interpretation of data captured by sensors on Ford's self-driving cars, and turning that data into usable info for the car's virtual driver system. SAIPS' offerings include detection of anomalies, persistent tracking of objects detected by sensors, and much more. The company's past clients include HP and Trax, but its partner group doesn't appear to have included much in the way of driving-specific applications.

CrunchBase

Ford Motor Company –

FOUNDED
1903

OVERVIEW
Ford is an automotive company that develops, manufactures, distributes, and services vehicles, parts, and accessories worldwide. It operates through two sectors: automotive and financial services. The automotive sector offers vehicles primarily under the Ford and Lincoln brand names. This sector markets cars, trucks, parts, and accessories through retail dealers in North America and distributors ...

LOCATION
Dearborn, MI

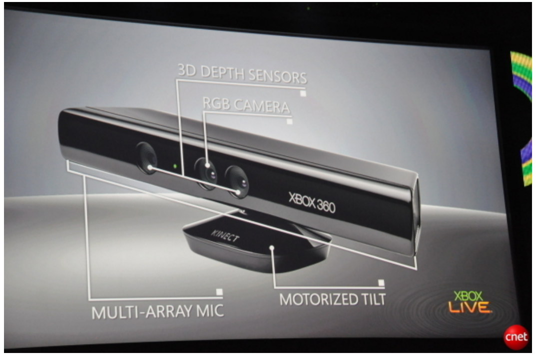
CATEGORIES
Automotive

WEBSITE
http://www.ford.com/

Full profile for Ford Motor Company

Interactive Games: Kinect – (Maybe)

- Object Recognition: http://www.youtube.com/watch?feature=iv&v=fQ59dXOo63o
- Mario: http://www.youtube.com/watch?v=8CTJL5lUjHg
- 3D: http://www.youtube.com/watch?v=7QrnwoO1-8A
- Robot: http://www.youtube.com/watch?v=w8BmgtMKFbY



Industrial robots

Vision-guided robots position nut runners on wheels

Vision in space

NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

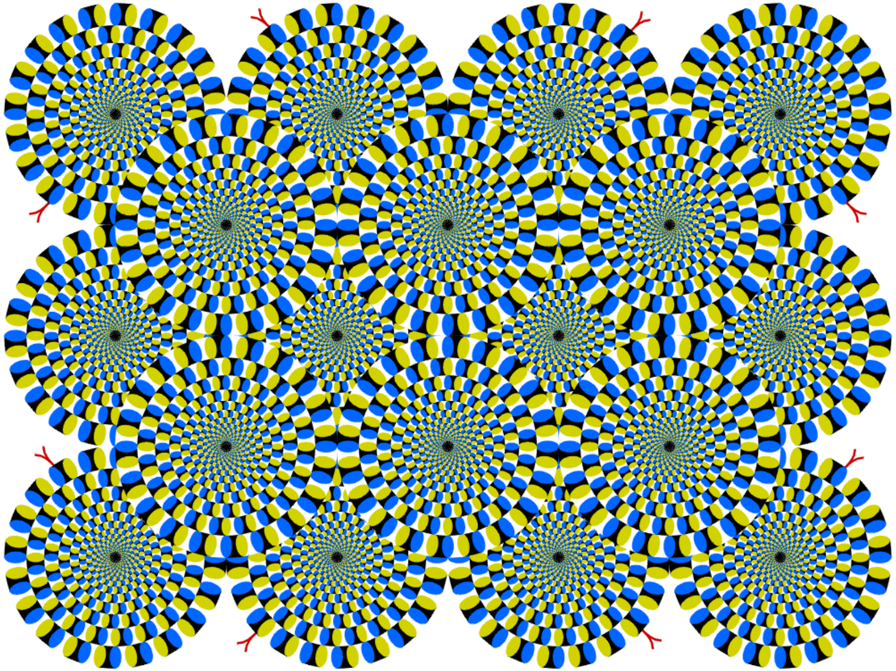
Vision systems (JPL) used for several tasks

- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read “Computer Vision on Mars” by Matthies et al.

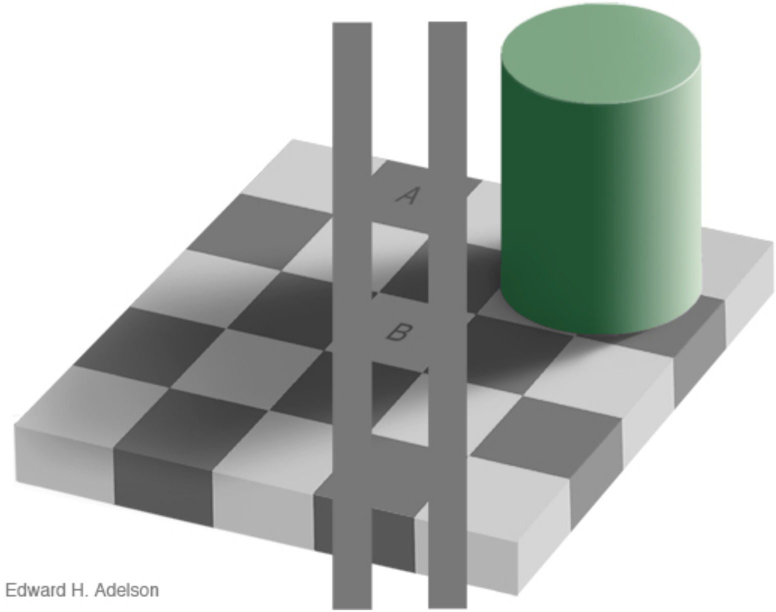
Augmented Reality and Virtual Reality

Magic Leap, Oculus, Hololens, etc.

Is seeing trivial?

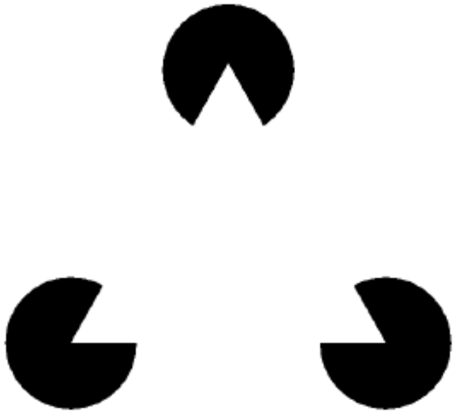


Is seeing trivial?



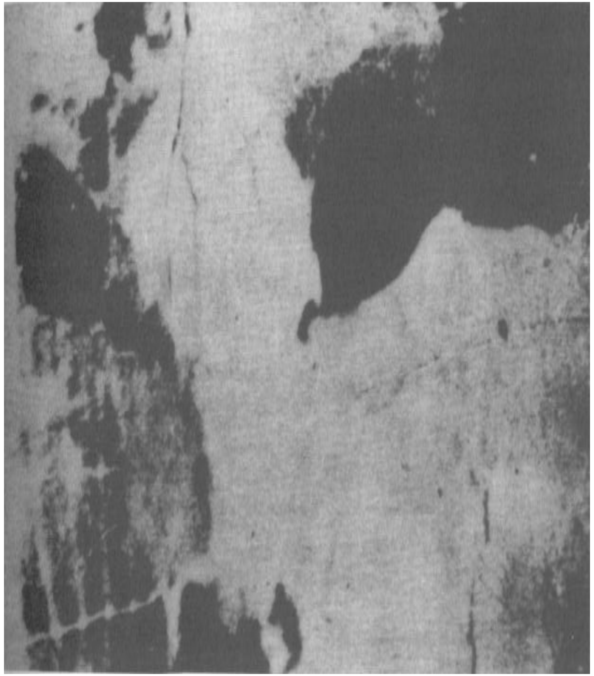
http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html

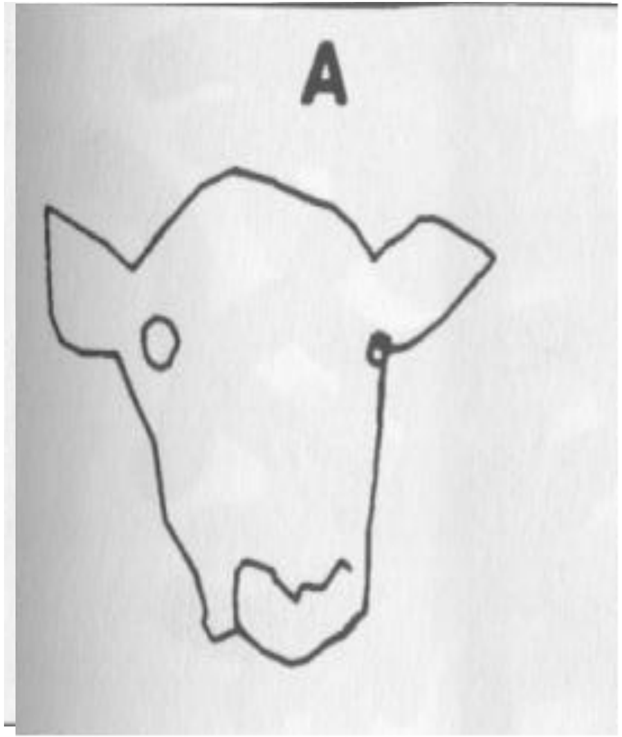
Is seeing trivial?



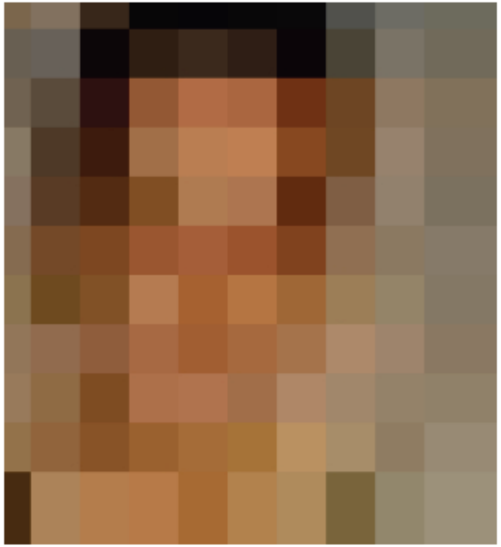
Is seeing trivial?

Is seeing trivial?

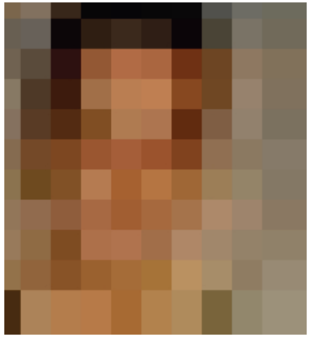




Face or non-face?

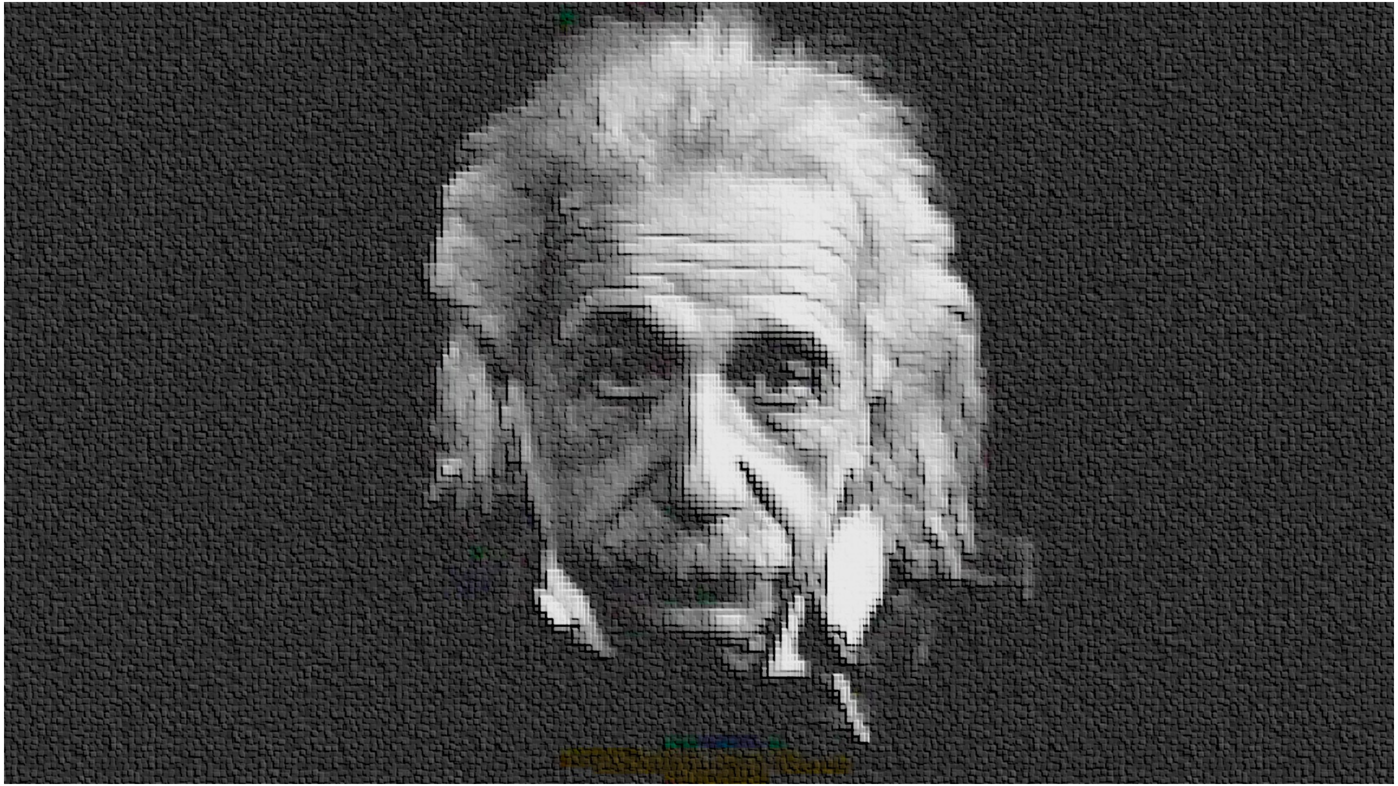


Face or non-face?

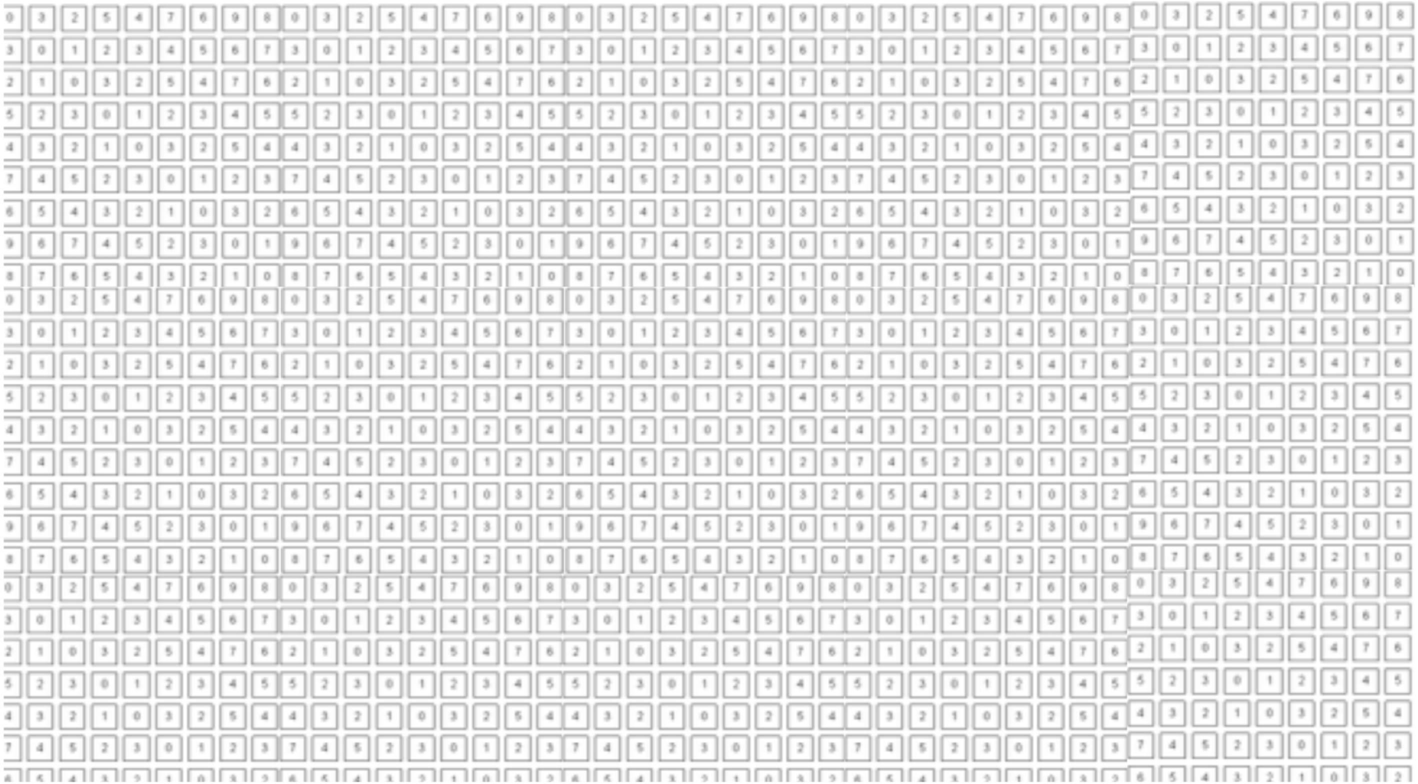


Why is vision (and recognition) so hard?

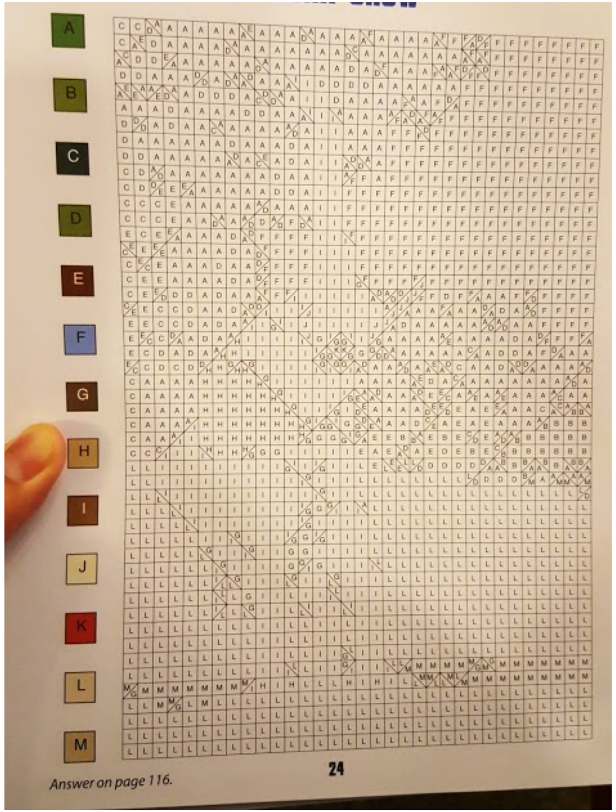
This is an image to us:

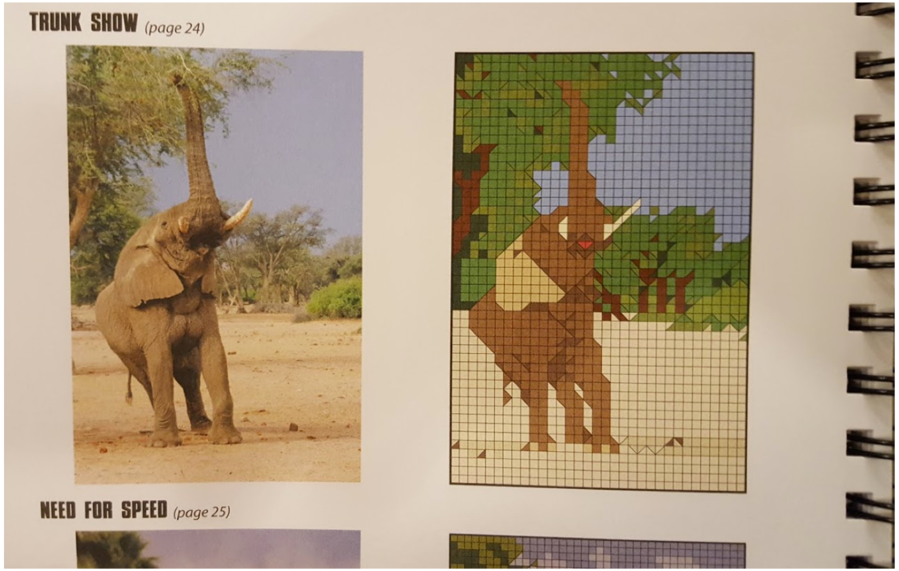


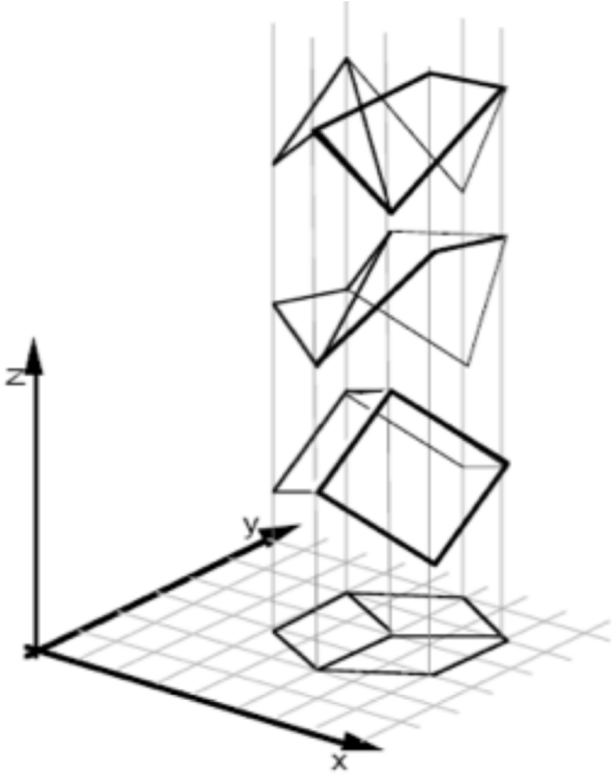
This is an image to a computer:



Vision is Hard







[Sinha and Adelson 1993]

View Points

Michelangelo 1475-1564

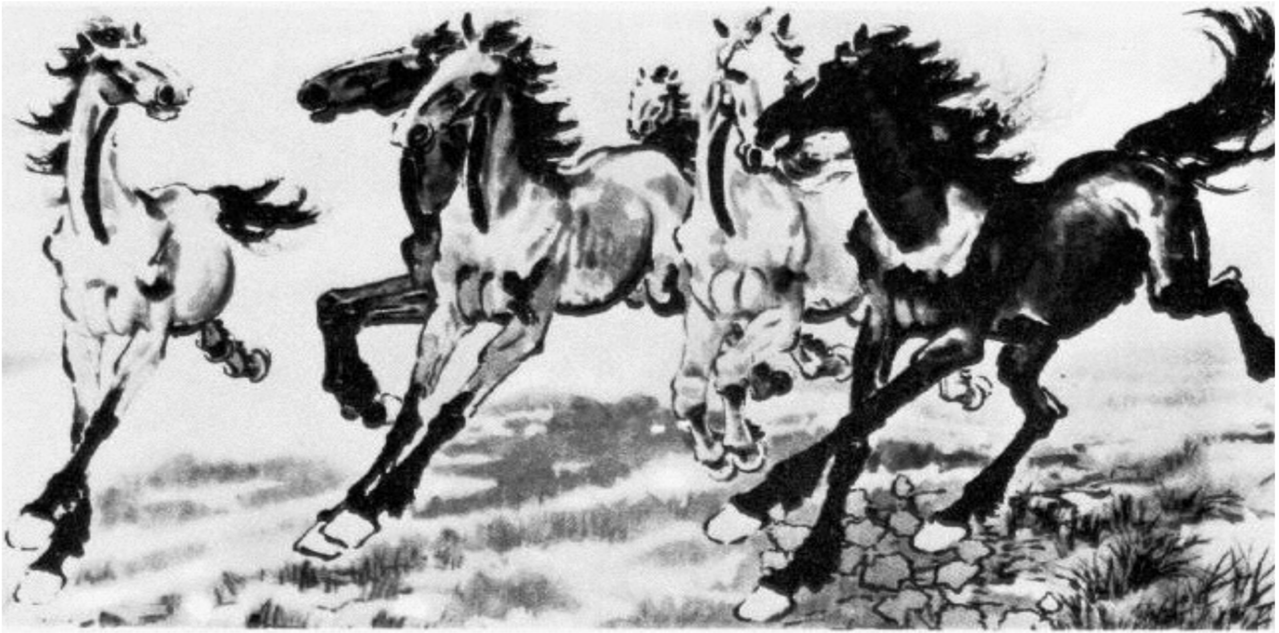
slide by Fei Fei, Fergus & Torralba

Illumination

slide credit: S. Ullman

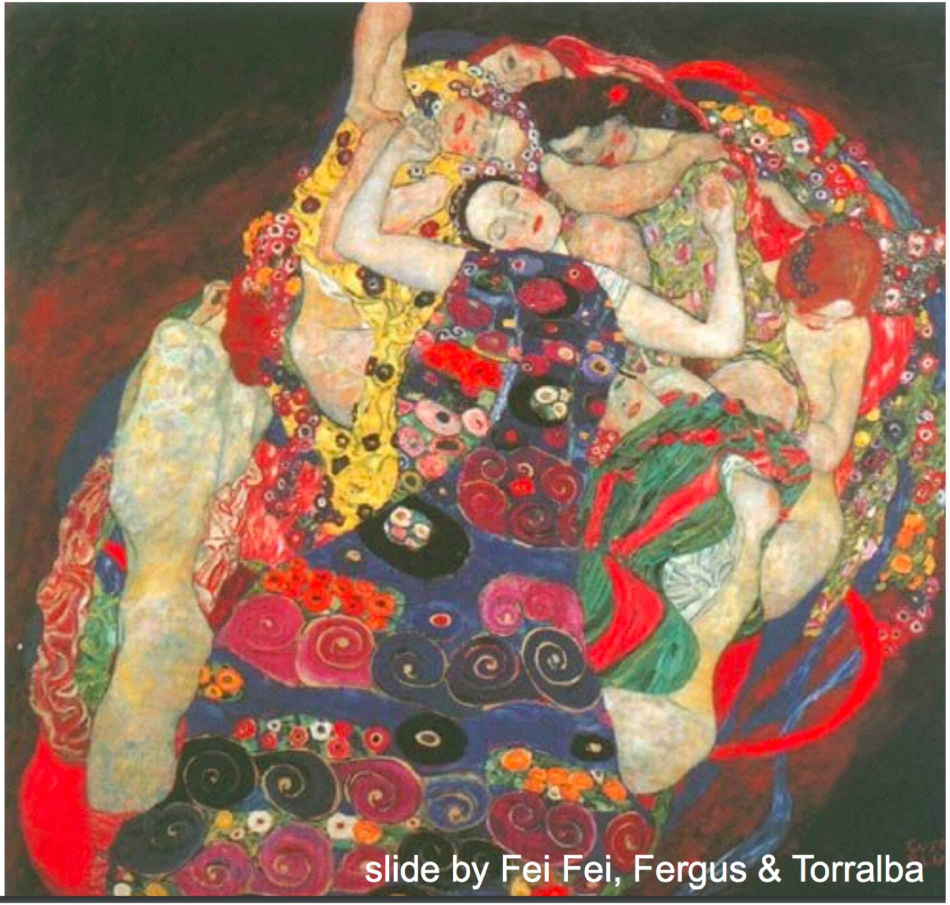
Occlusions

Deformation



Xu, Beihong 1943

Background clutter



Klimt, 1913

Intra-class variation

slide by Fei-Fei, Fergus & Torralba

What is the state of the art today?

- Given enough training data Computer Vision systems are surprisingly robust to the previously outlined challenges e.g. illumination changes, intra-class variation.
- Still not at the same level as humans, despite the hype.
- Still many open challenges, such as few-shot learning, transfer learning, and unsupervised learning.

Deep Learning and Vision

- Deep Learning has been a great disruption into the field of Computer Vision. Has made a lot of new things work!

- Many deep learning methods being applied to vision these days.

- This is not a pure deep learning course but a lot of topics will be covered in the context of visual recognition modles. We will briefly review some pre-deep learning methods, and then mostly deep learning.

Objectives

(a) Develop intuitions between aspects in human vision and computer vision,

(b) Understanding foundational concepts for representation learning using neural networks

(c) Becoming familiar with state-of-the-art models for tasks such as image classification, object detection, image segmentation, scene recognition, etc.

(d) Obtain practical experience in the implementation of visual recognition models using deep learning.

About the Course

CS6501-003: Deep Learning for Visual Recognition

- Instructor: Vicente Ordóñez
- Email: vicente@virginia.edu
- Website: http://vicenteordonez.com/deeplearning/
- Class Location: **Olsson Hall 005**
- Class Times: **Monday-Wednesday 3:30pm and 4:45pm**
- Piazza: https://piazza.com/virginia/spring2020/cs6501003/home
- Office Hours: TBD

Teaching Assistants

Ziyan Yang

(tw8cb@virginia.edu)

Hours: TBD

Paola Cascante-Bonilla

(pc9za@virginia.edu)

Hours: TBD

Pre-requisites

- Python programming skills
- Calculus / Linear Algebra / Probability

Grading

- Assignments: 400pts (4 assignments)
  (100pts + 100pts + 100pts + 100pts)

- Course Project: 400pts
  Groups of up to 3 students (more only if justified)

- Paper Reading Summaries: 100pts
- Class Paper Presentation: 100pts (groups of mostly 2 students)

Textbook

- *No textbook required.*

Suggested Reading for Next Class

- Szeliski Book, Chapter 3: Image Processing.
- [What the Frog's Eye Tells the Frog's Brain]

Also…

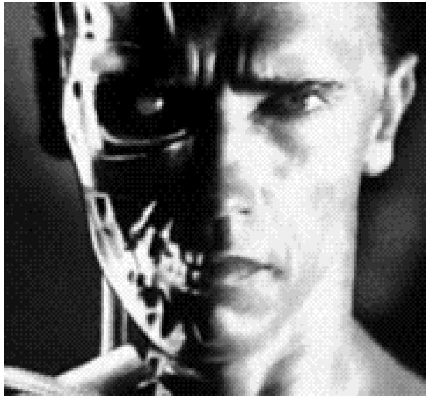
- Assignment 1 will be released on course website
- In the meantime complete, the pytorch/jupyter/Google Colaboratory tutorial + Numpy + Matplotlib Image Processing Primer

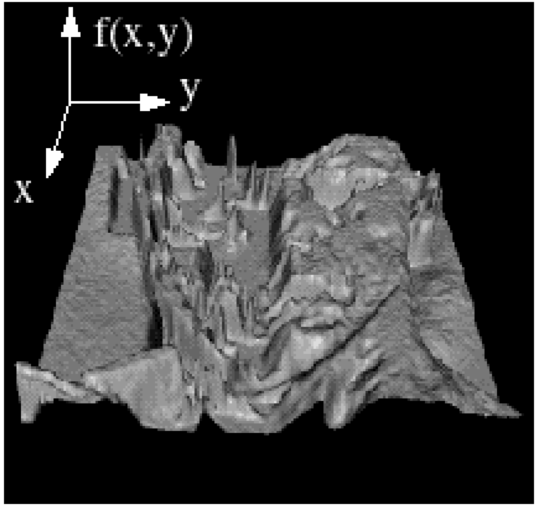
Reminder of what is an image for a computer.

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

Images as Functions

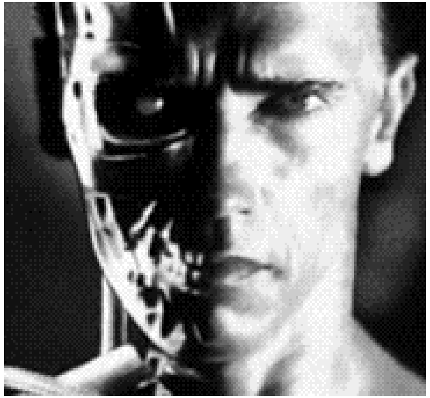
$$z = f(x, y)$$

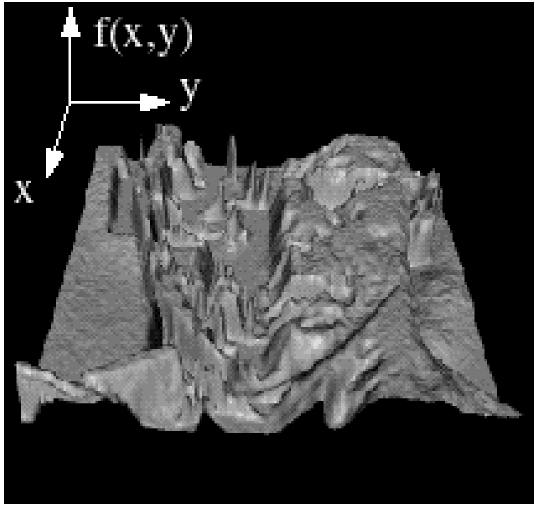




Images as Functions

$$z = f(x, y)$$





- The domain of x and y is [0, img-width) and [0 and img-height)
- x, and y are discretized into integer values.

Light

- What determines the color of a pixel?

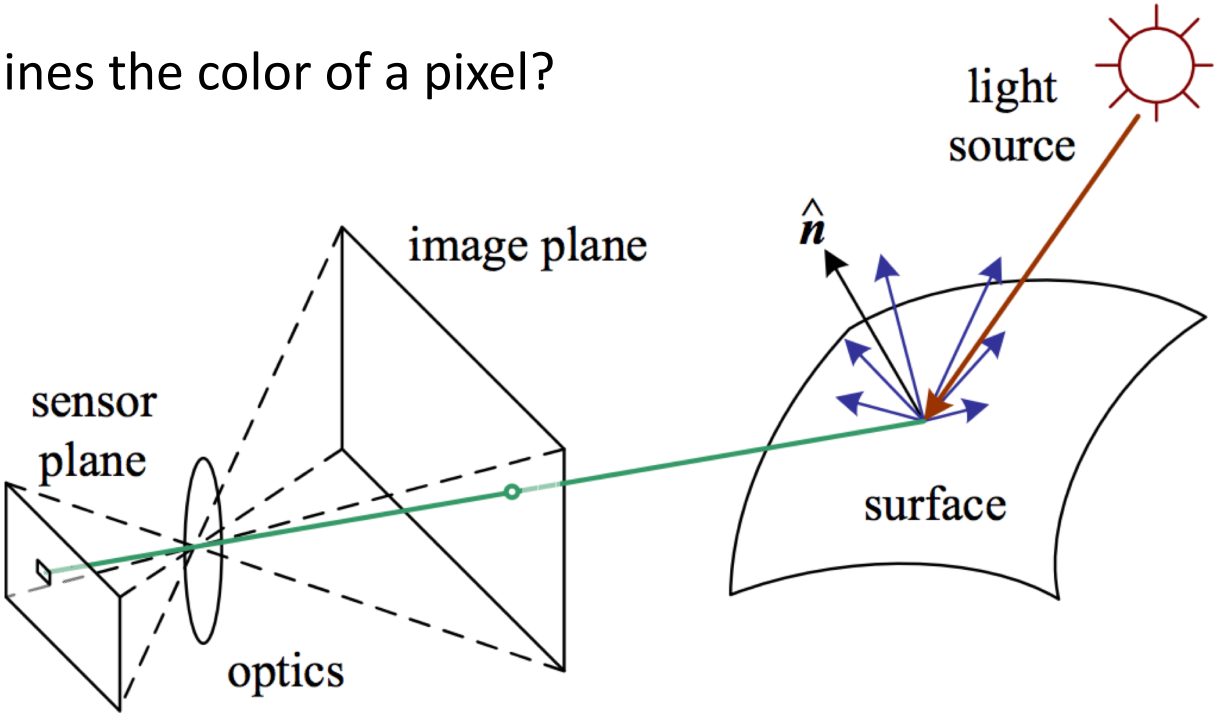
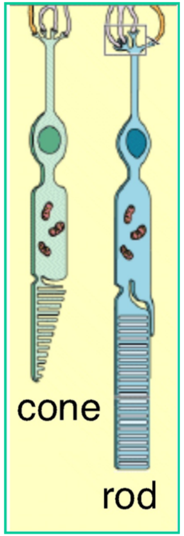


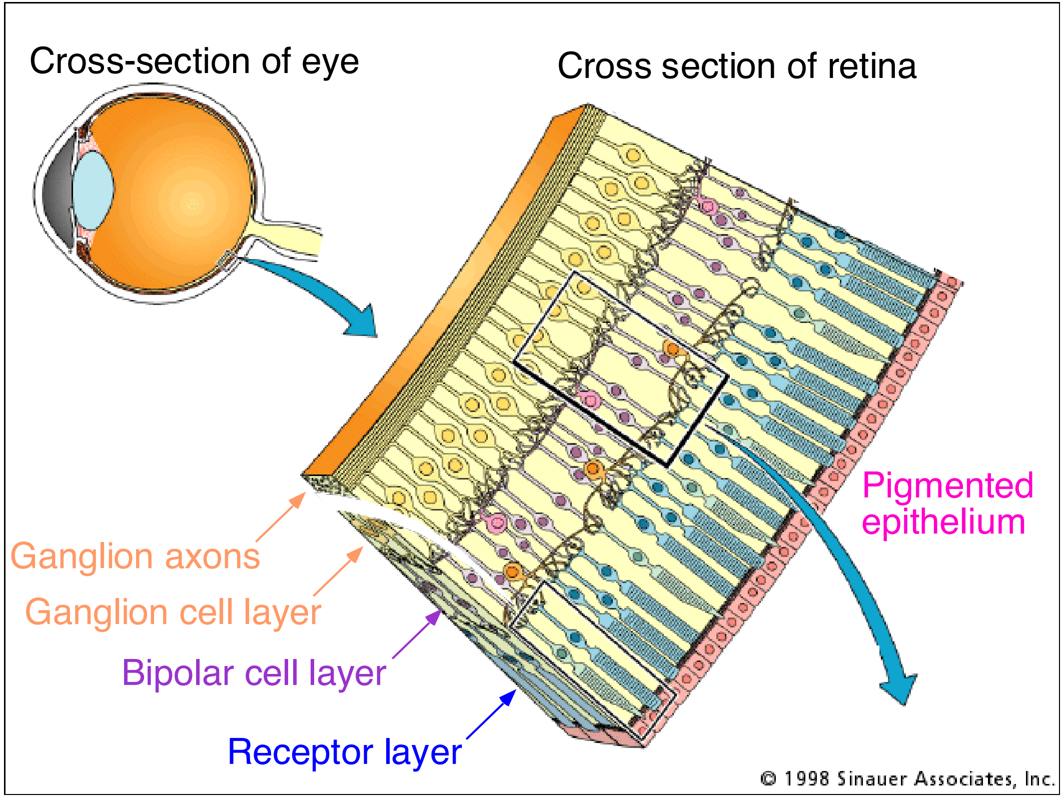
Figure from Szeliski

The Retina

Cones
- cone-shaped
- less sensitive
- operate in high light
- color vision

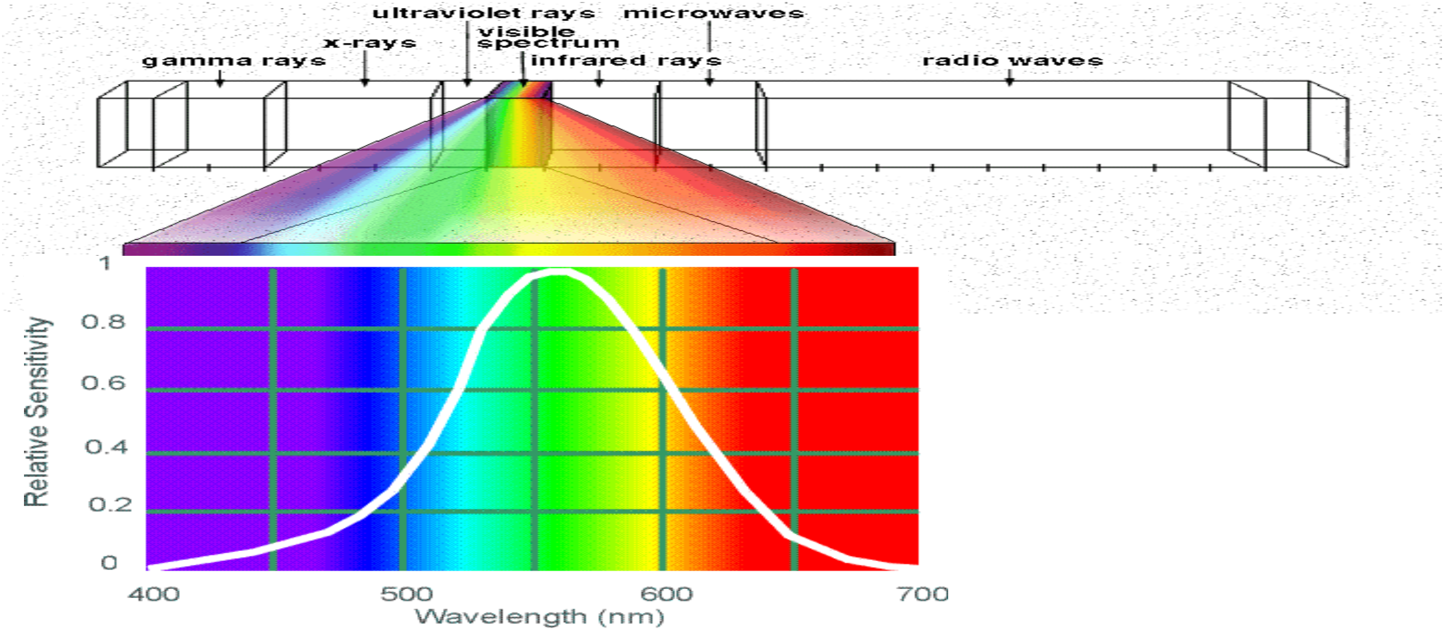
Rods
- rod-shaped
- highly sensitive
- operate at night
- gray-scale vision





[What the Frog's Eye Tells the Frog's Brain]

Electromagnetic Spectrum

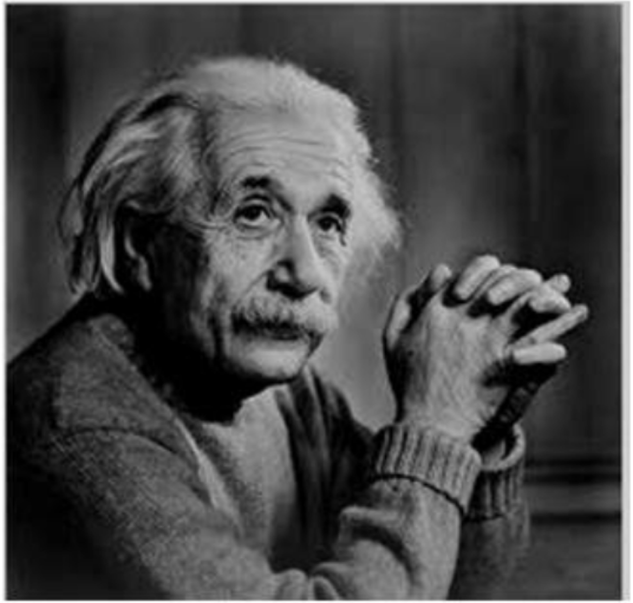


Human Luminance Sensitivity Function

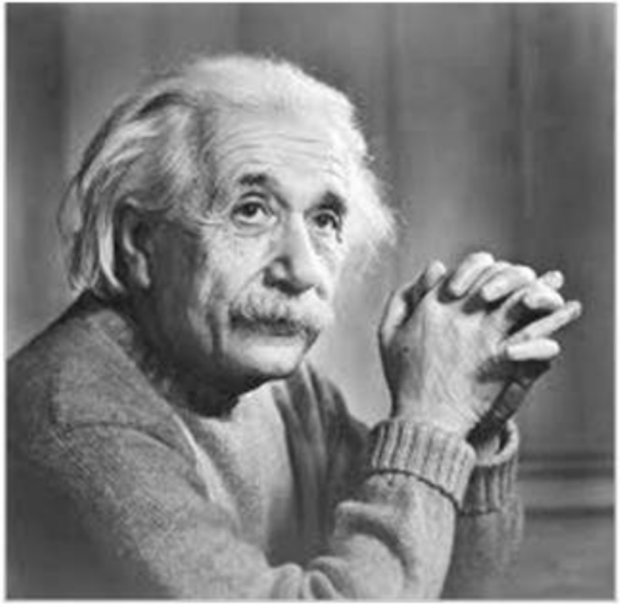
http://www.yorku.ca/eye/photopik.htm

Basic Image Processing

I



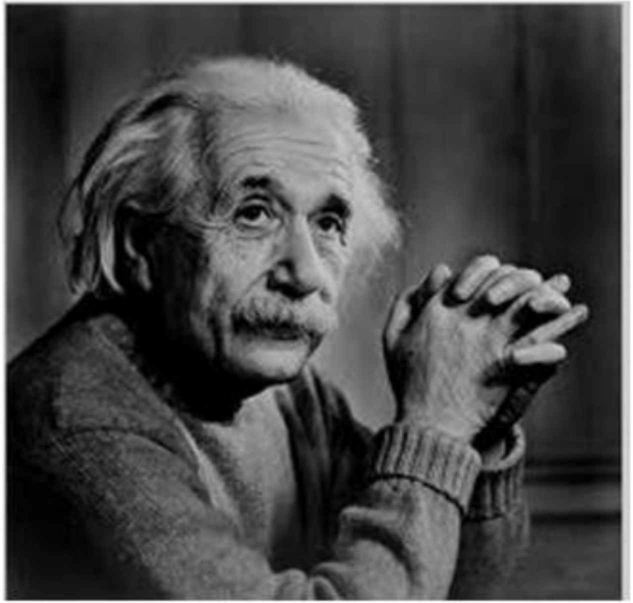
αI



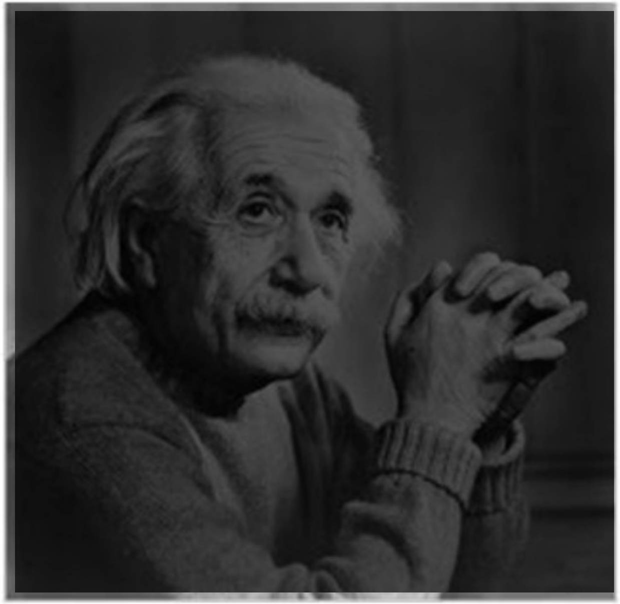
$\alpha > 1$

Basic Image Processing

I

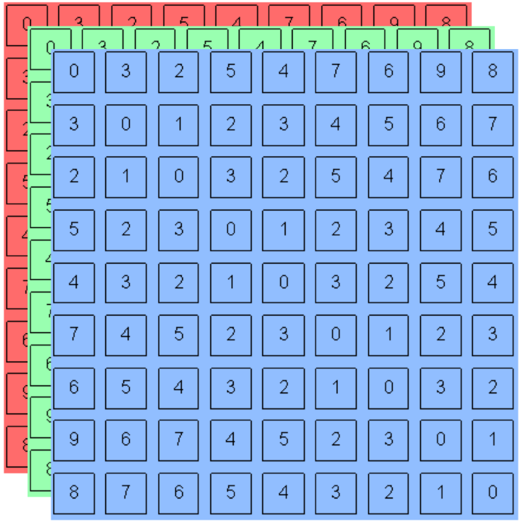


αI



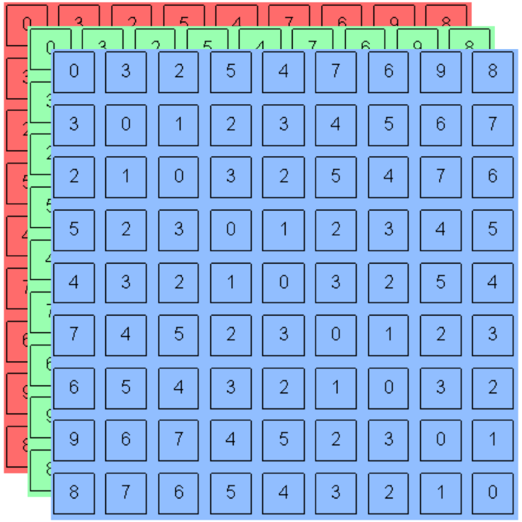
$0 < \alpha < 1$

Color Images as Tensors



$channel\ x\ height\ x\ width$

Color Images as Tensors



$channel\ x\ height\ x\ width$

Channels are usually RGB: Red, Green, and Blue

Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc

Questions?