Computer Vision AIML* ZG525

2025-26 Third Semester, M.Tech (AIML)

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Session 1:

Introduction to Computer Vision

Dhruba Adhikary
dhruba.a@wilp.bits-pilani.ac.in

Acknowledgement: Slide Materials adopted from - Intro to Computer Vision (Cornell Tech); Noah Snavely and prepared by the lead instructor Prof S P Vimal

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Course Objectives

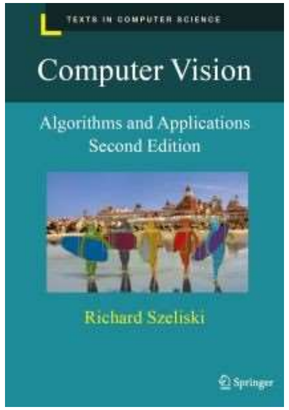
CO1: Students should understand the fundamentals of a camera producing an image, including camera calibration, optical distortions, perspective corrections etc.

CO2: Students should be familiar with various building block algorithms in Computer Vision, including Image processing and Deep Learning with emphasis on the algorithm building blocks.

CO3: Students should create at least one end-user application.

Textbook-1:

Image Processing, Analysis, and Machine Vision: Milan Sonka, Vaclav Hlavac, Roger Boyle, Fourth edition, Cengage Learning

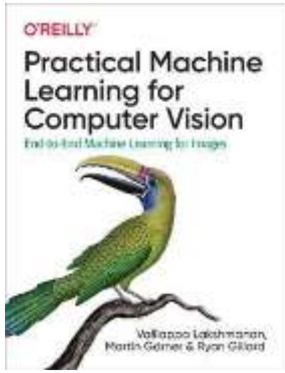


Reference-1:

Rick Szeliski, *Computer Vision: Algorithms and Applications*
Online at: http://szeliski.org/Book/

Reference-2:

Practical Machine Learning for Computer Vision:
End-to-End Machine Learning for Images, O'Rielly, 2021



Evaluation Plan

- Evaluation

  Two Quizzes for 5% each; Best one towards final grading; No Makeup; → 5%

  Two Assignments – x % + y % = 25%

  Mid Term Exam - 30%

  Comprehensive Exam - 40%

- Webinars/Tutorials

  4 tutorials : 2 before mid-sem & 2 after mid-sem

- Teaching Assistant: TBD

What will you learn in this course?

1. Computer Vision Fundamentals; (~2 Sessions)
2. Selected Topics in Low Level Vision & Mid Level Vision; (~3 Sessions)
3. Review of Deep Learning Approaches for Computer Vision (~ 1 Session) [As Webinar]
4. Image Classification – Problem, Deep Learning Architectures, Metrics, Use cases. (~2 Sessions)
5. Classic & Modern Approaches, Applications for

   (a) Object Detection, Recognition (~2 Sessions)

   (b) Face Detection, Recognition (~1 Session) [As Webinar]

   (c) Object Tracking (~2 Sessions)

   (d) Object Segmentation (~2 Sessions)

   (e) OCR (~ 1 Session) [As Webinar]
6. Visual Bag of Words & Semantic Hierarchy (~1 Session)
7. Edge Devices for Computer Vision (~1 Session) [As Webinar]

What will you learn in this course?

- On your learning
  - Reading materials given at the end of each class
  - Most of the topics needs to balance depth and breadth. We will use research articles and other online materials quite frequently.
  - Python demonstrations will be part of regular classes as required □ TA's will join the lectures sessions to support with demonstration.
  - Wherever necessary, stress will be given to related classic topics in Computer Vision.
  - Assignment problems [to be solved in a group of maximum of 4 members] will be given on natural images and remote sensing images & you can make a choice.

- Plagiarism Policy

All submissions for graded components must be the result of your original effort. It is strictly prohibited to copy and paste verbatim from any sources, whether online or from your peers. The use of unauthorized sources or materials, as well as collusion or unauthorized collaboration to gain an unfair advantage, is also strictly prohibited.

Please note that we will not distinguish between the person sharing their resources and the one receiving them for plagiarism, and the consequences will apply to both parties equally.

In cases where suspicious circumstances arise, such as identical verbatim answers or a significant overlap of unreasonable similarities in a set of submissions, will be investigated, and severe punishments will be imposed on all those found guilty of plagiarism.

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Agenda

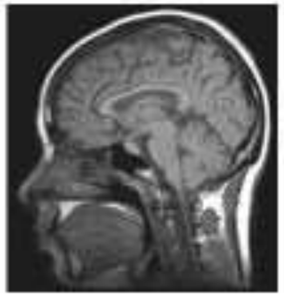
Introduction to the course:

- What is Computer Vision?
- Why Computer Vision is hard?
- Applications of Computer Vision
- Image representation and image analysis tasks

What is Computer Vision?

- **Computer vision** is the science and technology of machines that see.
- Concerned with the theory for building artificial systems that obtain information from images.
- The image data can take many forms, such as a video sequence, depth images, views from multiple cameras, or multi-dimensional data from a medical scanner





Computer Vision

Make computers understand images and videos.

What kind of scene?

Where are the cars?

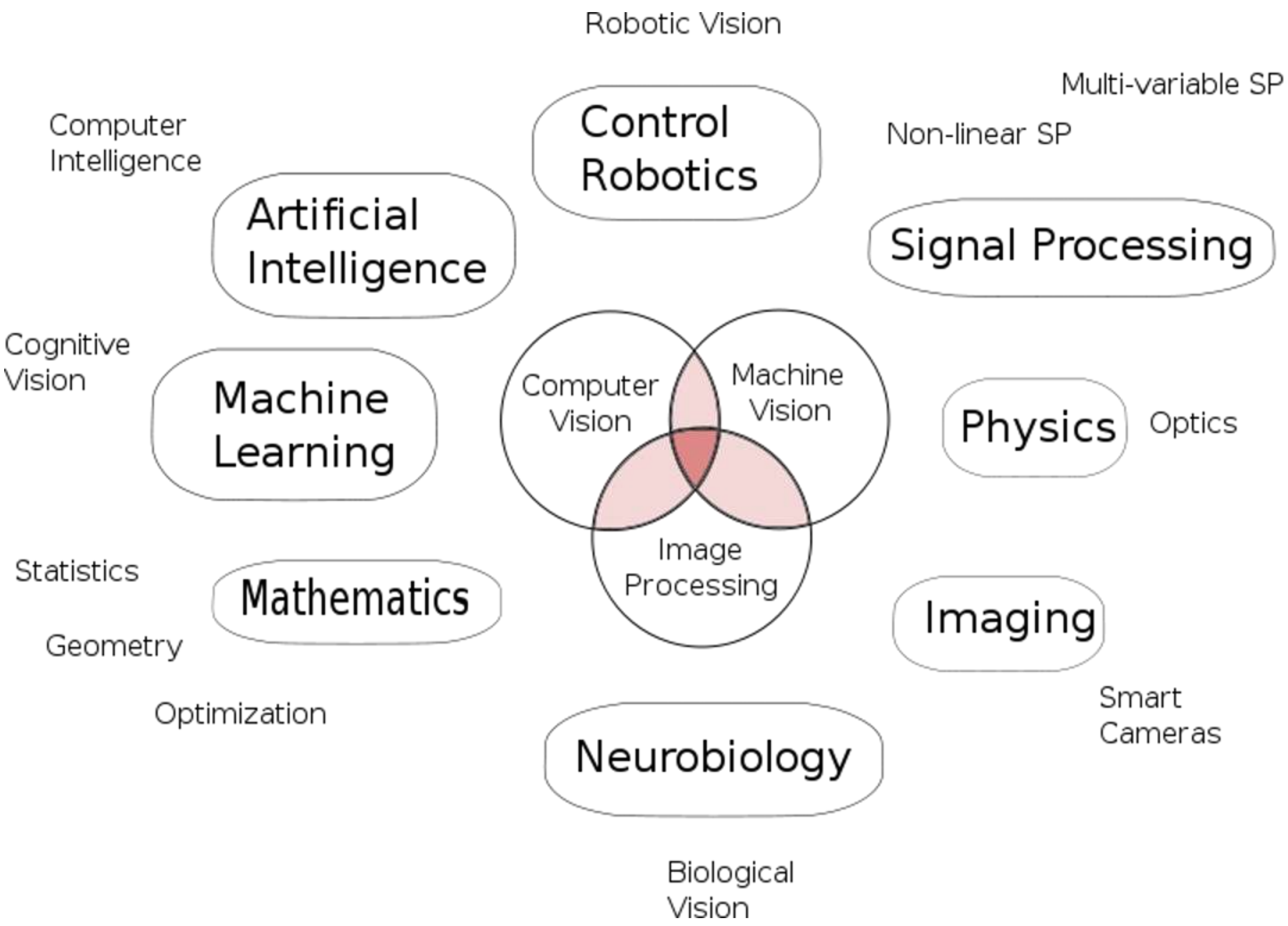
How far is the building?

…

Vision is really hard

- Vision is an amazing feat of natural intelligence
  - Visual cortex occupies about 50% of Macaque brain
  - More human brain devoted to vision than anything else

Vision is Multi-disciplinary



Every image tells a story

- Goal of computer vision: perceive the "story" behind the picture
- Compute properties of the world
  – 3D shape
  – Names of people or objects
  – What happened?

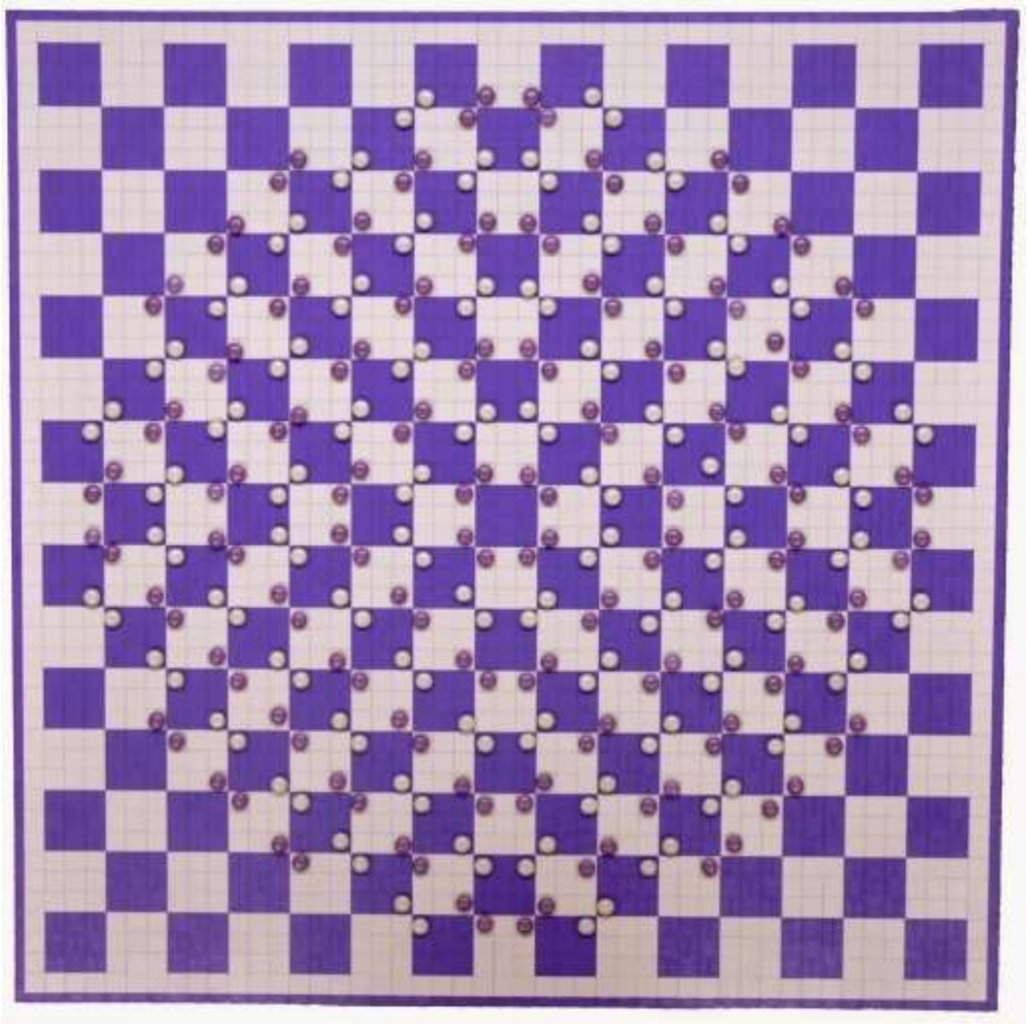
The goal of computer vision

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

Can computers match human perception?

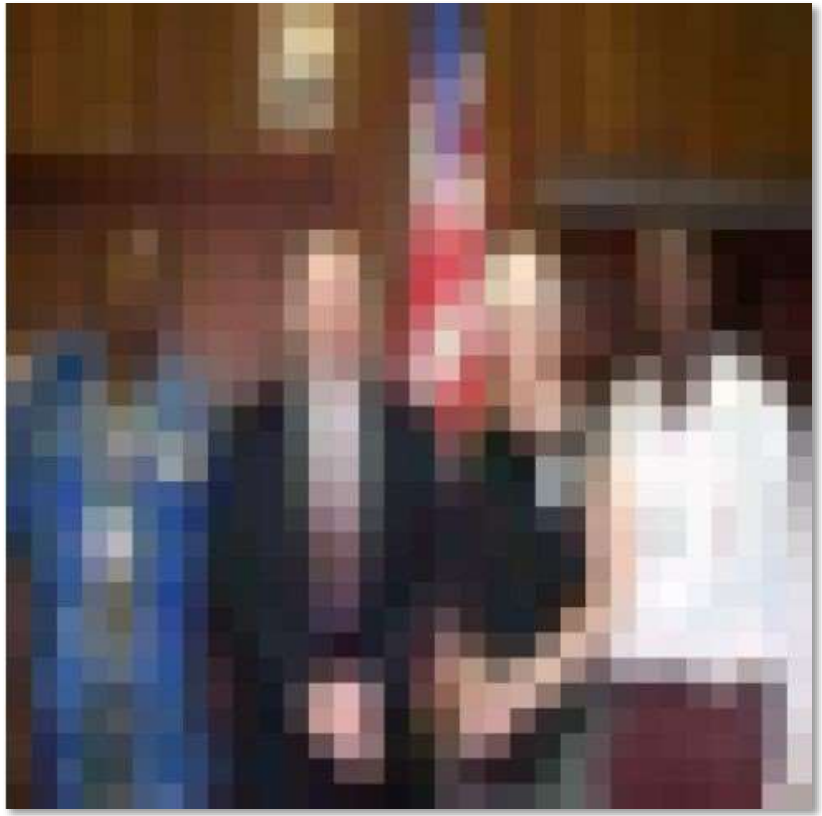
- Yes and no (mainly no)
  - computers can be better at "easy" things
  - humans are better at "hard" things
- But huge progress
  - Accelerating in the last five years due to deep learning
  - What is considered "hard" keeps changing

Human perception has its shortcomings



https://twitter.com/pickover/status/1460275132958662657/

But humans can tell a lot about a scene from a little information…



Source: “80 million tiny images” by Torralba, et al.

What is in here?

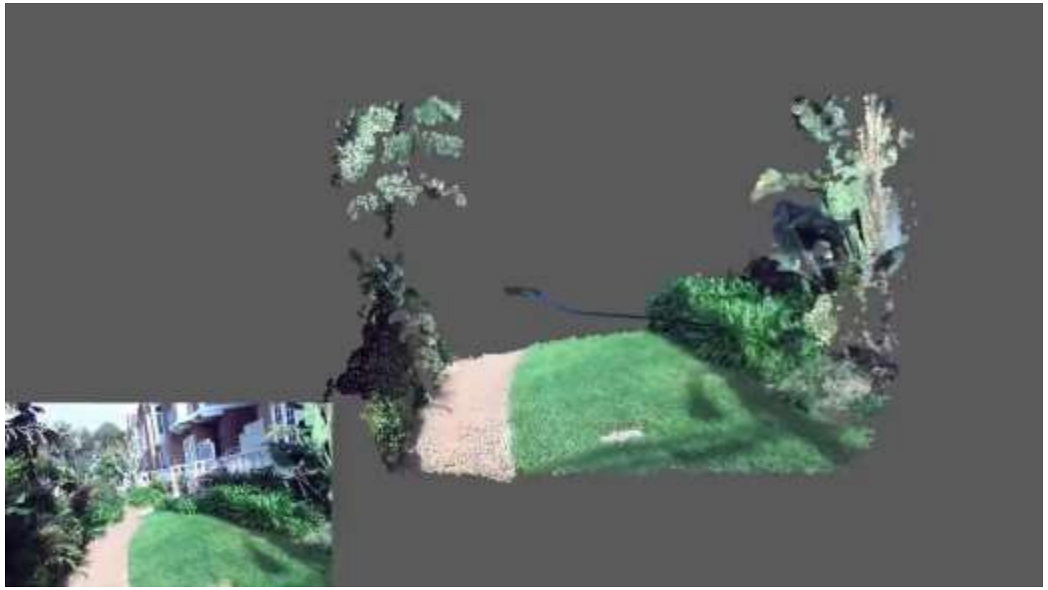
What is in here?

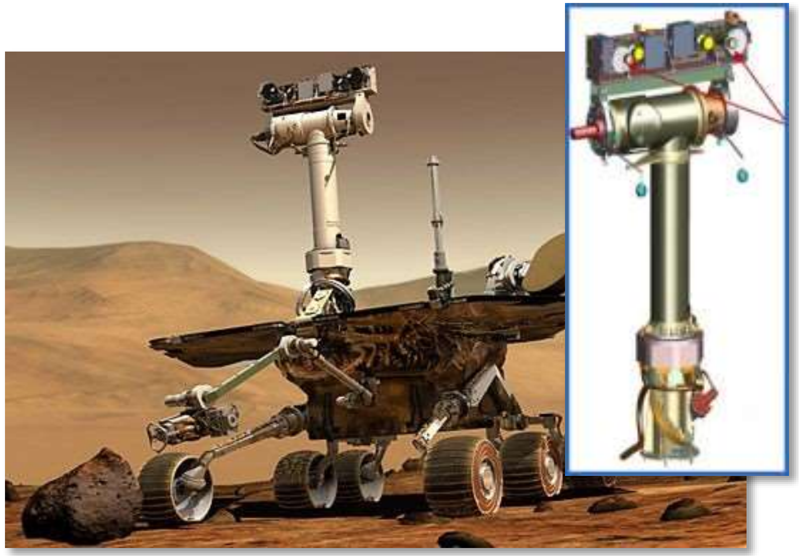
The goal of computer vision

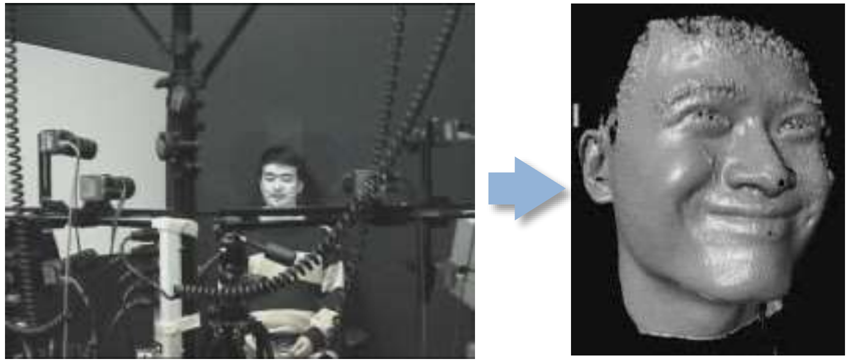
The goal of computer vision

- Compute the 3D shape of the world

ZED 2i Camera

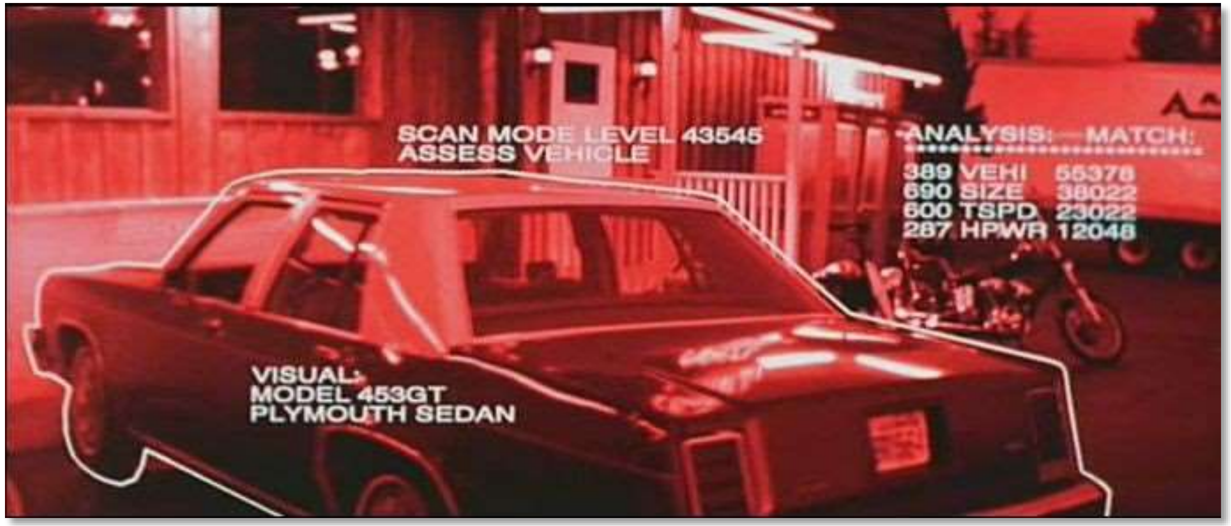






The goal of computer vision

- Recognize objects and people



Terminator 2,
1991


中华人民
和国万岁
世界
民大团结万岁


sky
building
flag
face
banner
wall
street lamp
bus
bus
cars
slide credit: Fei-Fei, Fergus & Torralba

The goal of computer vision

- “Enhance” images

The goal of computer vision

- Improve photos ("Computational Photography")

Super-resolution (source: 2d3)

Low-light photography
(credit: [Hasinoff et al., SIGGRAPH ASIA 2016])

Depth of field on cell phone camera
(source: [Google Research Blog])

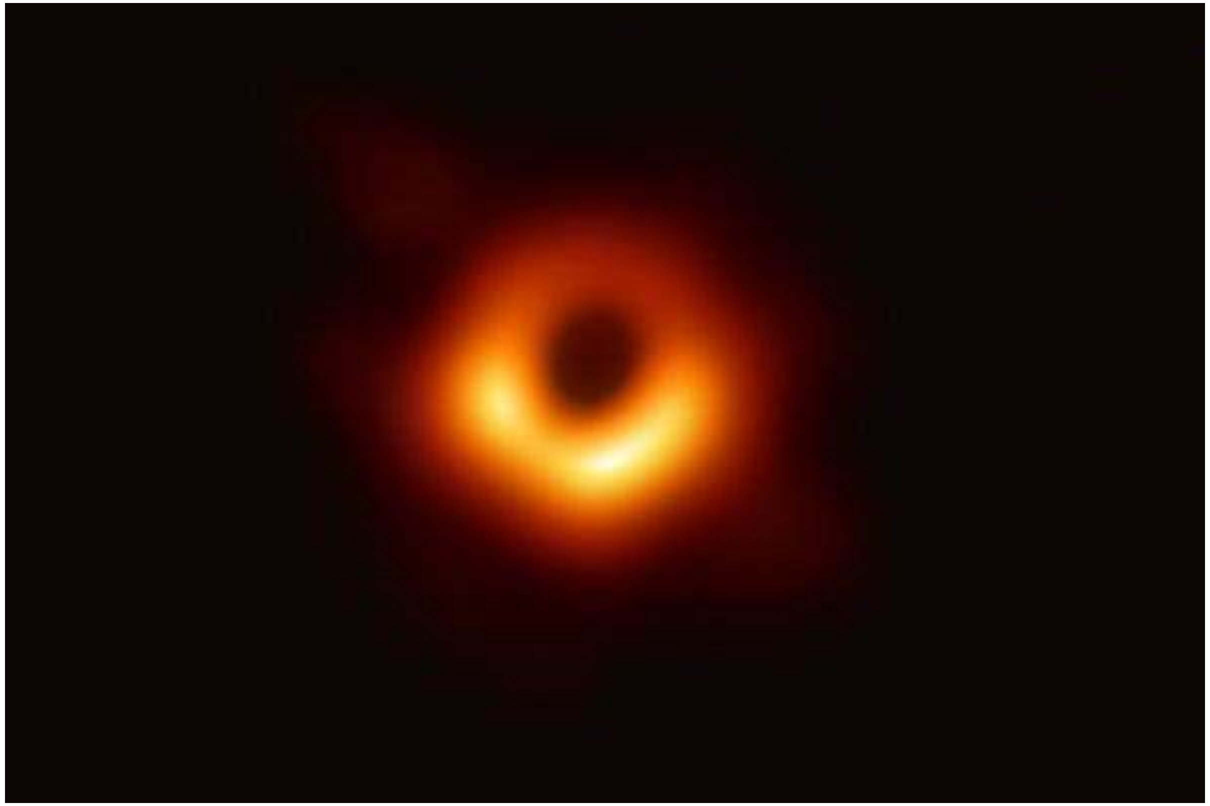
Removing objects
([Google Magic Eraser])

The New York Times

Darkness Visible, Finally: Astronomers Capture First Ever Image of a Black Hole

Astronomers at last have captured a picture of one of the most secretive entities in the cosmos.

April 10, 2019



Why study computer vision?

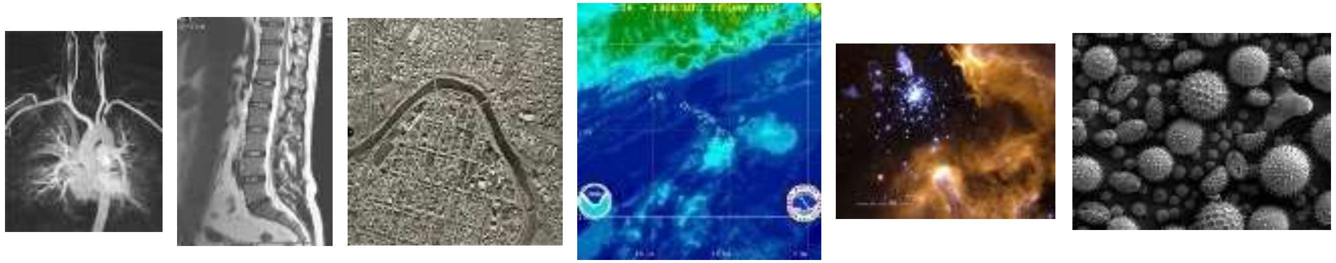
- Billions of images/videos captured per day

flickr

Google Photos

YouTube

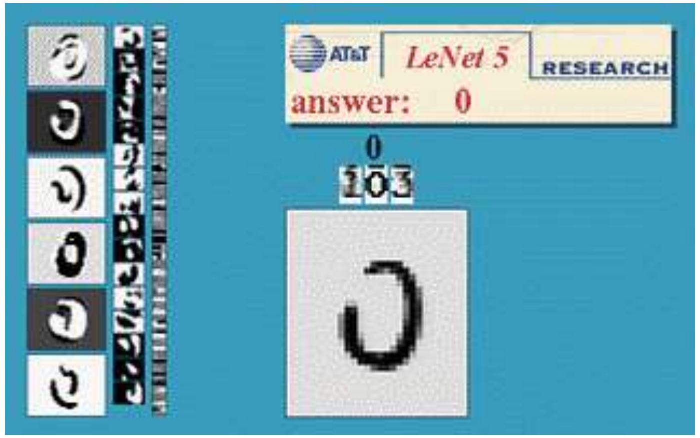
TikTok



- Huge number of potential applications

Optical character recognition (OCR)

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



Digit recognition, AT&T labs (1990's)
http://yann.lecun.com/exdb/len

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

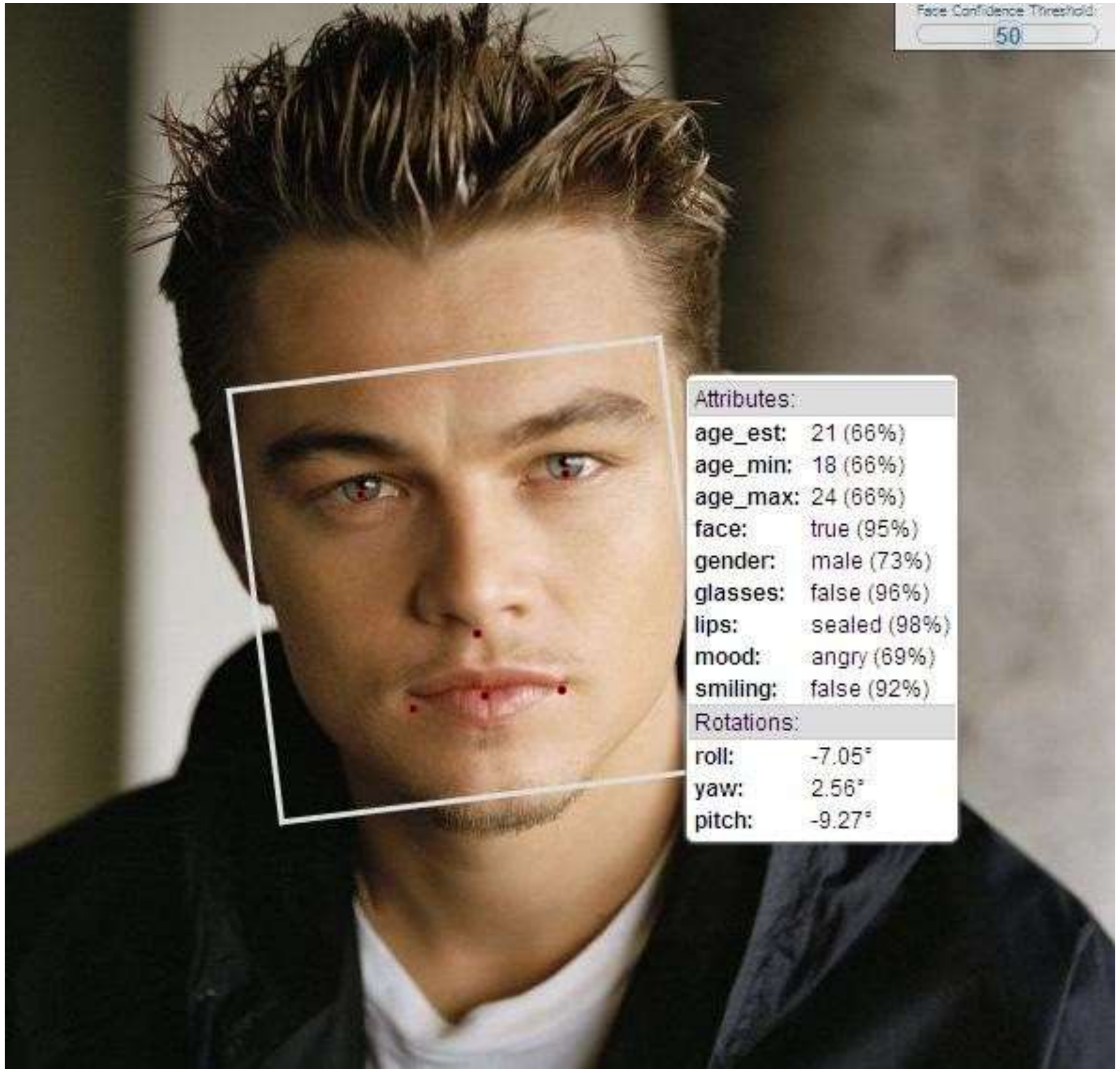
Automatic check

Sudoku grabber
http://sudokugrab.blogspot.com/

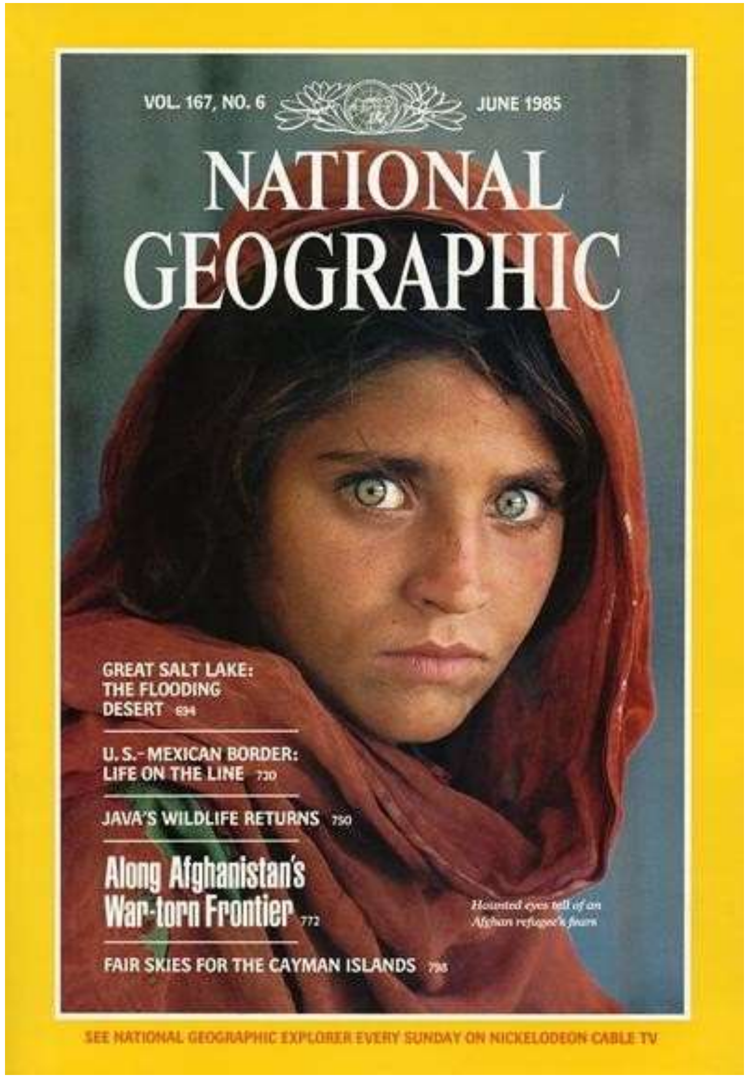
Face detection

- Nearly all cameras detect faces in real time
  - (Why?)

Face analysis and recognition



Vision-based biometrics

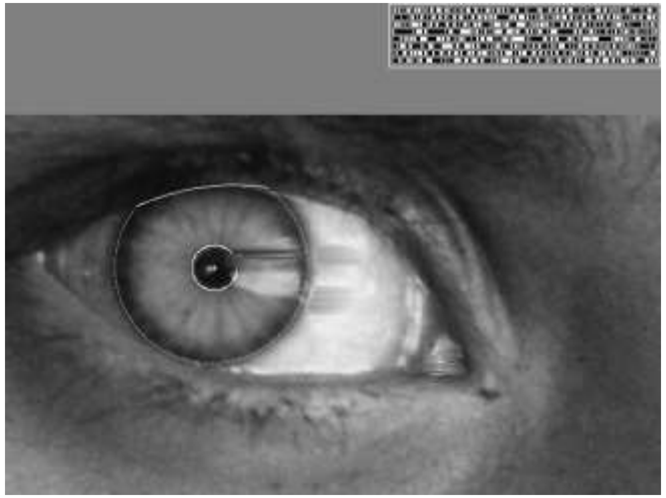


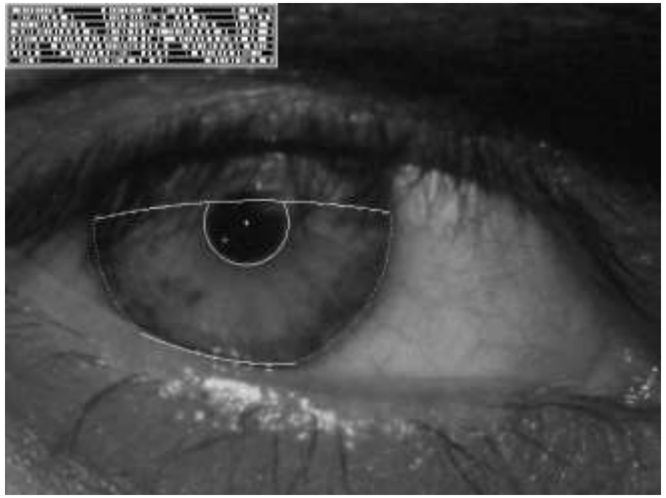
Who is she?

Source: S. Seitz

Vision-based biometrics

"How the Afghan Girl was Identified by Her Iris Patterns" [Story , youtube link]





Source: S. Seitz

Login without a password

Fingerprint scanners on many new smartphones and other devices

Face unlock on Apple iPhone X
See also
http://www.sensiblevision.com/

Bird identification

Merlin Bird ID (based on Cornell Tech technology!)

Special effects: shape capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC

Source: S. Seitz

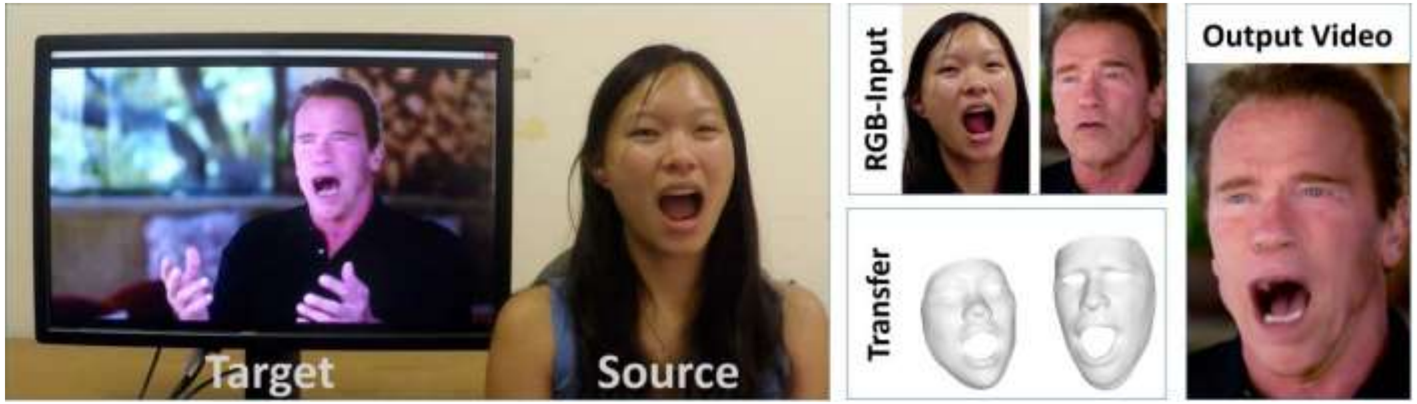
Special effects: motion capture

Pirates of the Carribean, Industrial Light and Magic

Source: S. Seitz

3D face tracking w/ consumer cameras

Snapchat Lenses



Face2Face system (Thies et al.)

Image synthesis

Karras, et al., *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, ICLR 2018

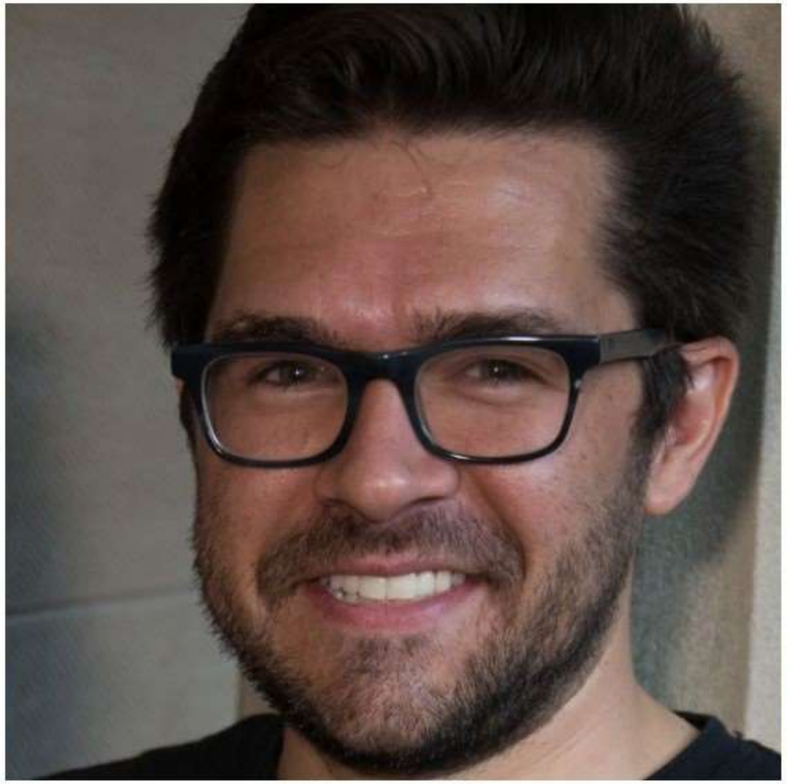
Image synthesis

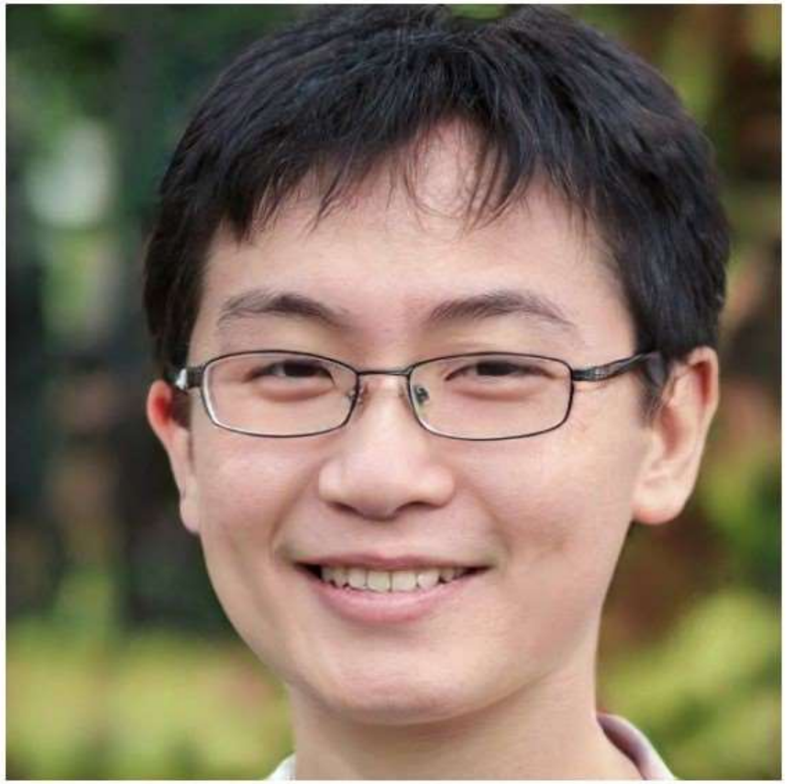
"An astronaut riding a horse in a photorealistic style" – DALL-E 2

"A photo of a Corgi dog riding a bike in Times Square. It is wearing sunglasses and a beach hat" – Imagen

Which face is real?

Click on the person who is real.





https://www.whichfaceisreal.com/

Smart cars

- [Mobileye]
- Tesla Autopilot
- Safety features in many cars

Robotics

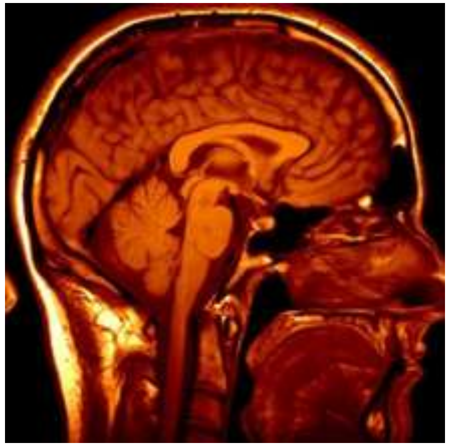
NASA's Mars Curiosity Rover
https://en.wikipedia.org/wiki/Curiosity_(rover)

Amazon Picking Challenge
http://www.robocup2016.org/en/events/amazon-picking-challenge/

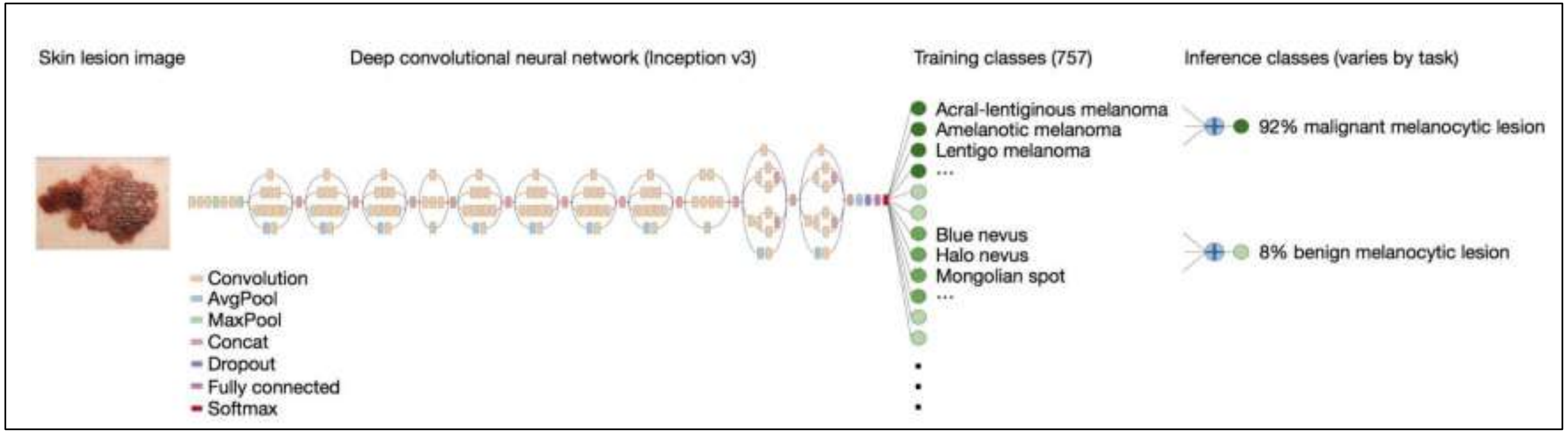
Amazon Prime Air

Amazon Scout

Medical imaging



3D imaging
(MRI, CT)



Skin cancer classification with deep learning
https://cs.stanford.edu/people/esteva/nature/

Current state of the art

- You just saw many examples of current systems.
  - Many of these are less than 5 years old
- Computer vision is an active research area, and rapidly changing
  - Many new apps in the next 5 years
  - Deep learning powering many modern applications
- Many startups across a dizzying array of areas
  - Deep learning, robotics, autonomous vehicles, medical imaging, construction, inspection, VR/AR, …

Why is computer vision difficult?

Viewpoint
variation

Illumination

Scale

Why is computer vision difficult?

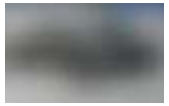
Intra-class variation

Motion (Source: S. Lazebnik)

Background clutter

Occlusion

Challenges: local ambiguity



slide credit: Fei-Fei, Fergus & Torralba

But there are lots of visual cues we can use…

Source: S. Lazebnik

Bottom line

- Perception is an inherently ambiguous problem
  - Many different 3D scenes could have given rise to a given 2D image

Artist Julian Beever with his anamorphic Coke bottle

  - We often must use prior knowledge about the world's structure

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.

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

Oct 22, 2012

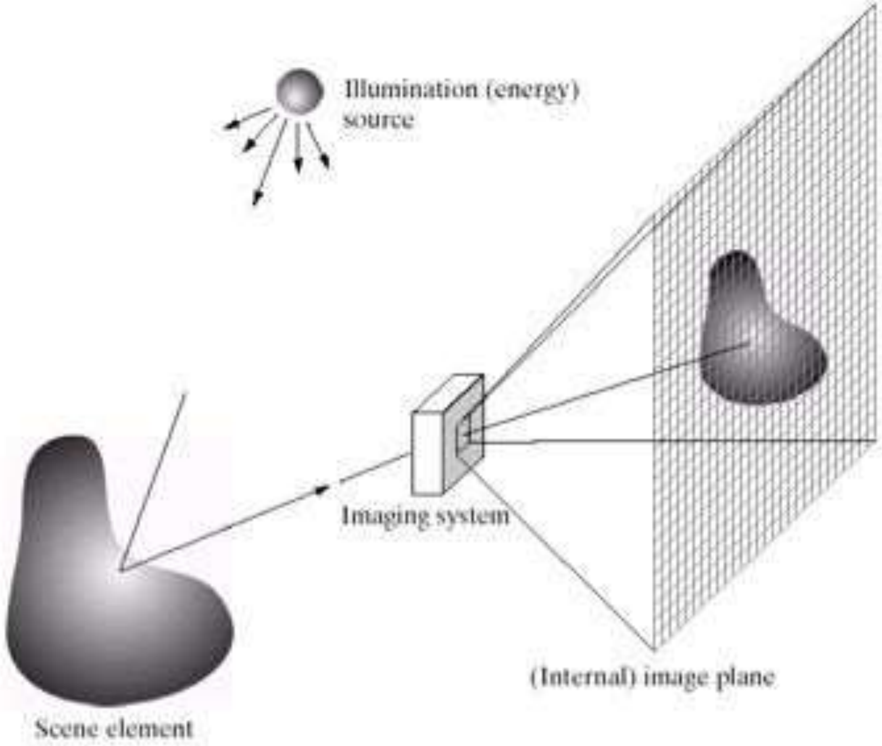
The picture above is funny.

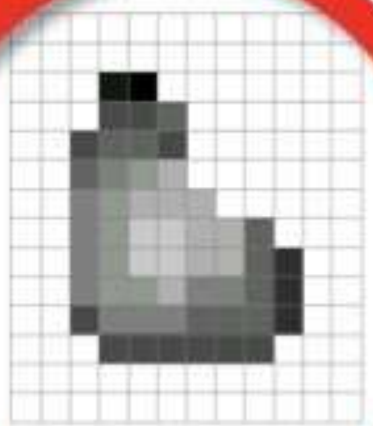
But for me it is also one of those examples that make me sad about the outlook for AI and for Computer Vision. What would it take for a computer to understand this image as you or I do? I challenge you to think explicitly of all the pieces of knowledge that have to fall in place for it to make sense.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

What is an image?

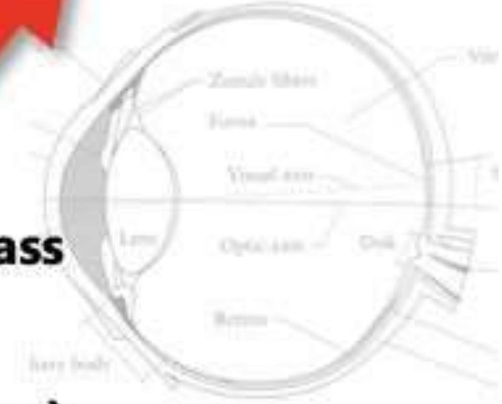
What is an image?





Digital Camer

We'll focus on these in this class



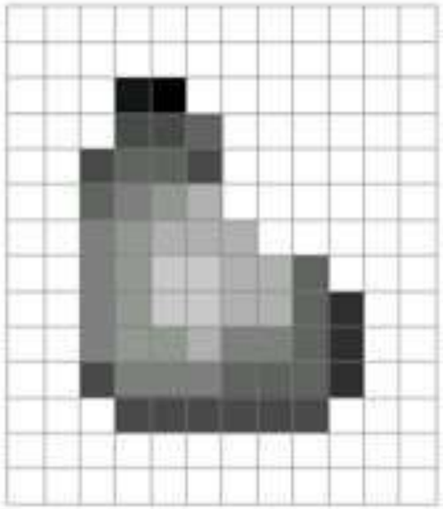
(More on this process later)

The Eye

Source: A. Efros

What is an image?

- A grid (matrix) of intensity values



=

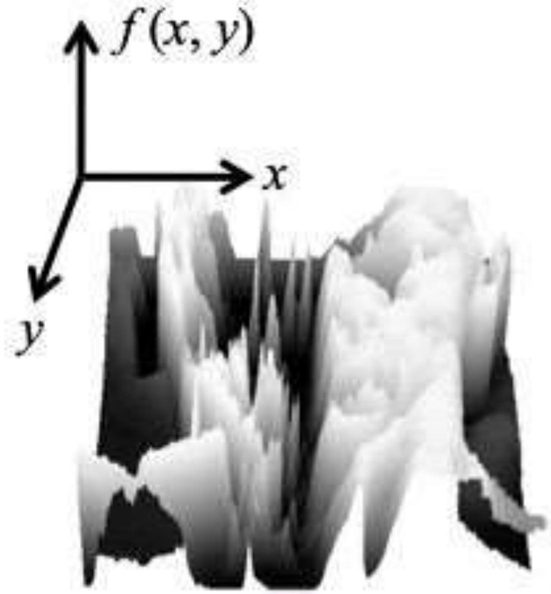
255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	20	0	255	255	255	255	255	255	255
255	255	255	75	75	75	255	255	255	255	255	255
255	255	75	95	95	75	255	255	255	255	255	255
255	255	96	127	145	175	255	255	255	255	255	255
255	255	127	145	175	175	175	255	255	255	255	255
255	255	127	145	200	200	175	175	95	255	255	255
255	255	127	145	200	200	175	175	95	47	255	255
255	255	127	145	145	175	127	127	95	47	255	255
255	255	74	127	127	127	95	95	95	47	255	255
255	255	255	74	74	74	74	74	74	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255

(common to use one byte per value: 0 = black, 255 = white)

What is an image?

- Can think of a (grayscale) image as a **function** f from R^2 to R:
  - $f(x,y)$ gives the **intensity** at position (x,y)

snoop



3D view

  - A **digital** image is a discrete (**sampled**, **quantized**) version of this function

Image representation for analysis

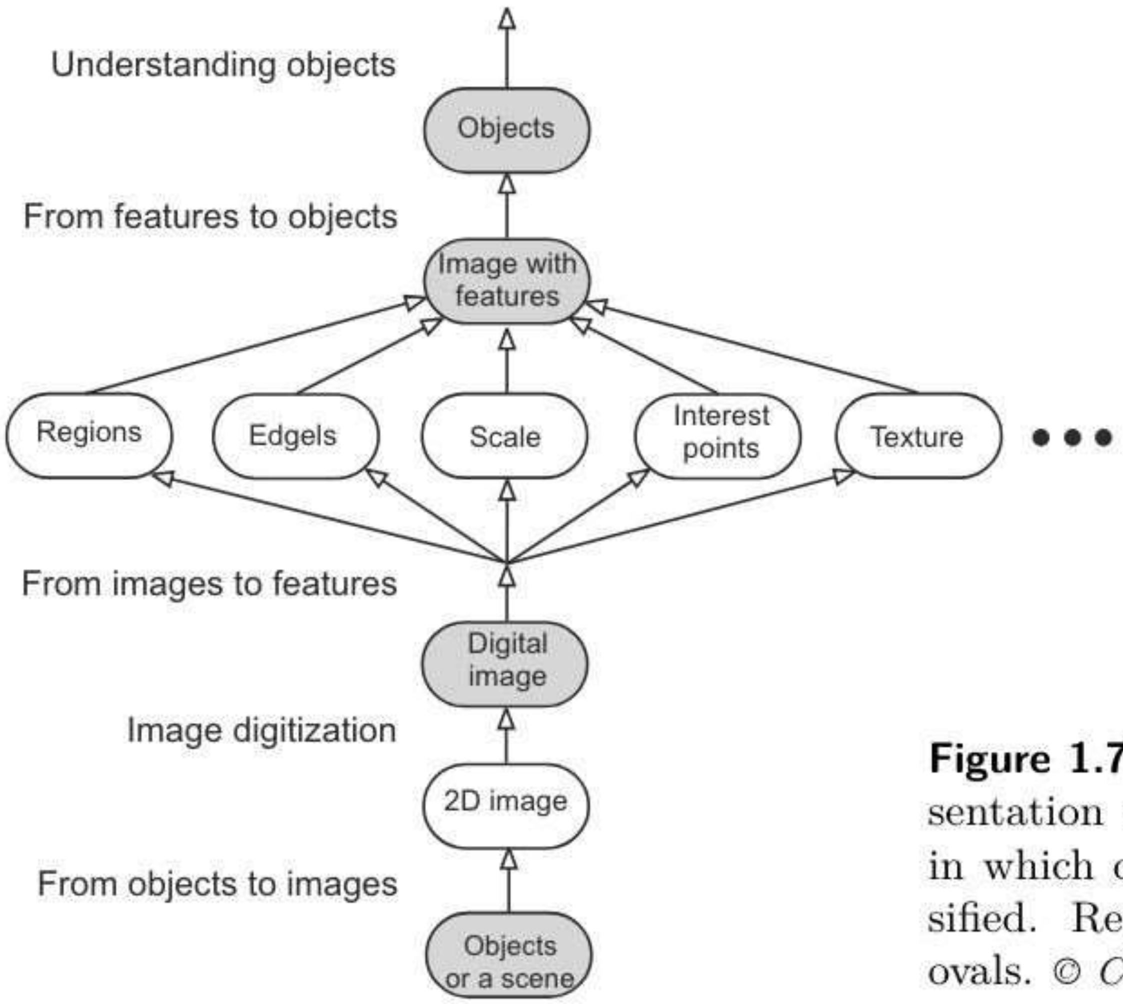


Figure 1.7: Four possible levels of image representation suitable for image analysis problems in which objects have to be detected and classified. Representations are depicted as shaded ovals. © *Cengage Learning 2015.*

Representations

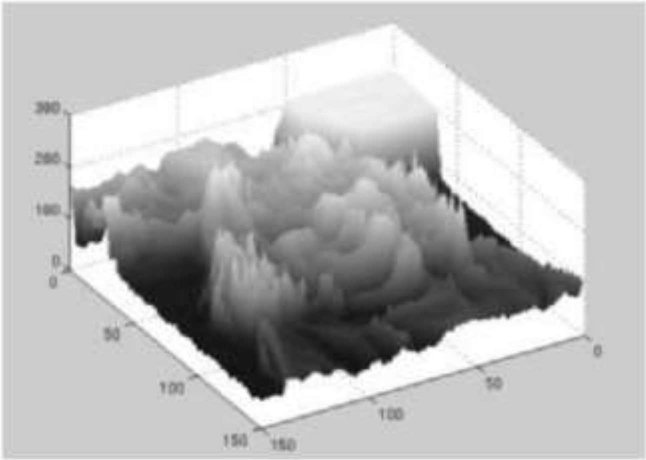
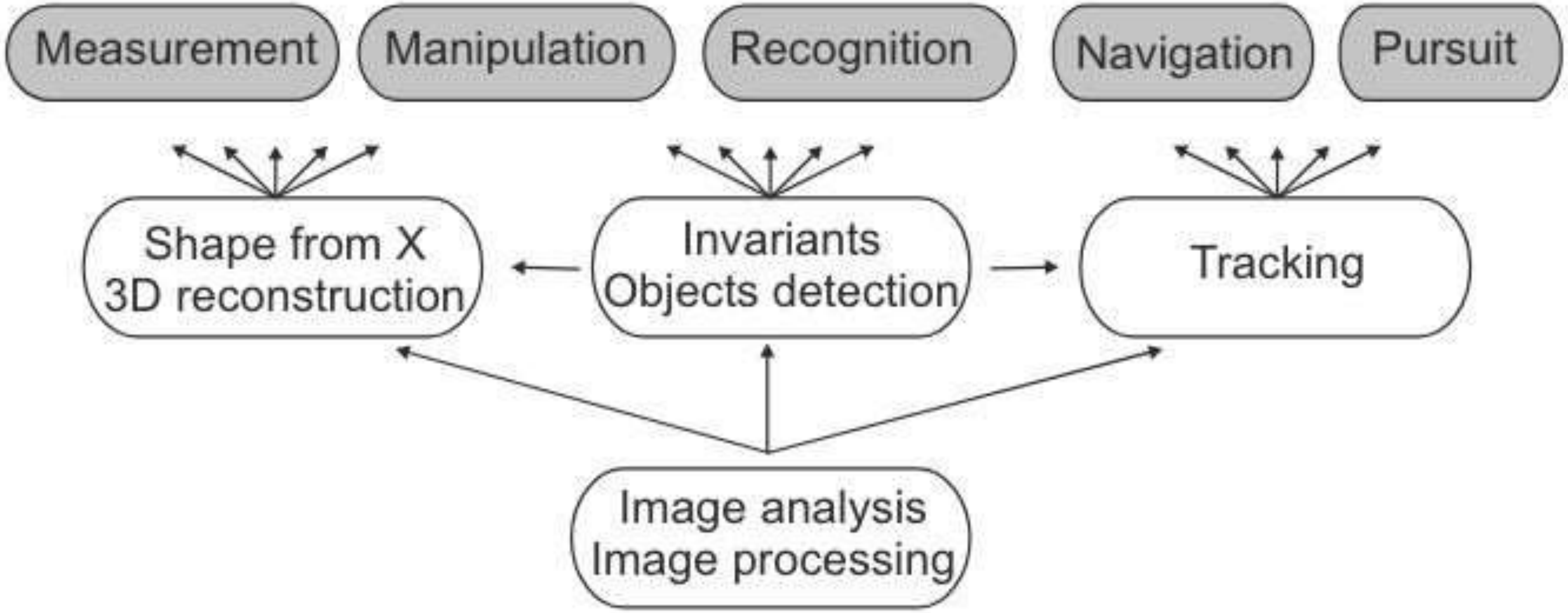


Figure 1.8: An unusual image representation. © *R.D. Boyle 2015.*

Figure 1.9: Another representation of Figure 1.8. © *R.D. Boyle 2015.*

3D vision tasks and algorithmic components



- Readings:
  1. What is Computer Vision?
  2. Why Computer Vision is hard? (T1 Ch 1.2)
  3. Applications of Computer Vision (R1 Ch 1.1)
  4. Image representation and image analysis tasks (T1 Ch 1.3)

- Topics for Next Class
  5. Image digitization - Sampling and resolution (T1 Ch 2.2)
  6. Digital Images (T1 Ch 2.3)
  7. Digital Image types -Binary, Gray-scale and Color (Class Notes)
  8. Color Images (T1 Ch 2.4)
  9. Color spaces: RGB and HSV (T1 Ch 2.4)

DHRUBA ADHIKARY

Data Scientist - Generative AI & Computer Vision | Applying Deep Learning to Advance...

Thank you