Introduction

Pablo Mesejo
pmesejo@go.ugr.es
Universidad de Granada
Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

DECSAI

DaSCI
Instituto Andaluz de Investigación en
Data Science and Computational Intelligence

What is computer vision?

- AI branch that deals with problems related to **visual perception**:
  - Science/Engineering dealing with the **automatic interpretation of images and videos** (i.e. to extract "meaning" from pixels)
- Russell & Norvig's AI book (Ch. 1) explicitly mentions 6 main AI branches:
  - Natural language processing
  - Knowledge representation
  - Automated reasoning
  - Machine learning
  - **Computer Vision (Ch. 27)**
  - Robotics

Necessary to pass the total Turing test, which requires interaction with objects and people in the real world

What is computer vision?

- “In computer vision, [we try] to **describe the world that we see** in one or more images and to **reconstruct its properties**, such as shape, illumination and color distributions.” Szeliski (2022)
- “Computer vision aims to duplicate the effect of human vision by electronically **perceiving and understanding an image**. […] Giving computers the ability to see […]” Sonka et al. (2015).
- “Computer vision aims at using **cameras** for analysing or **understanding scenes in the real world**.” Klette, R. *Concise computer vision*. Springer. 2014.

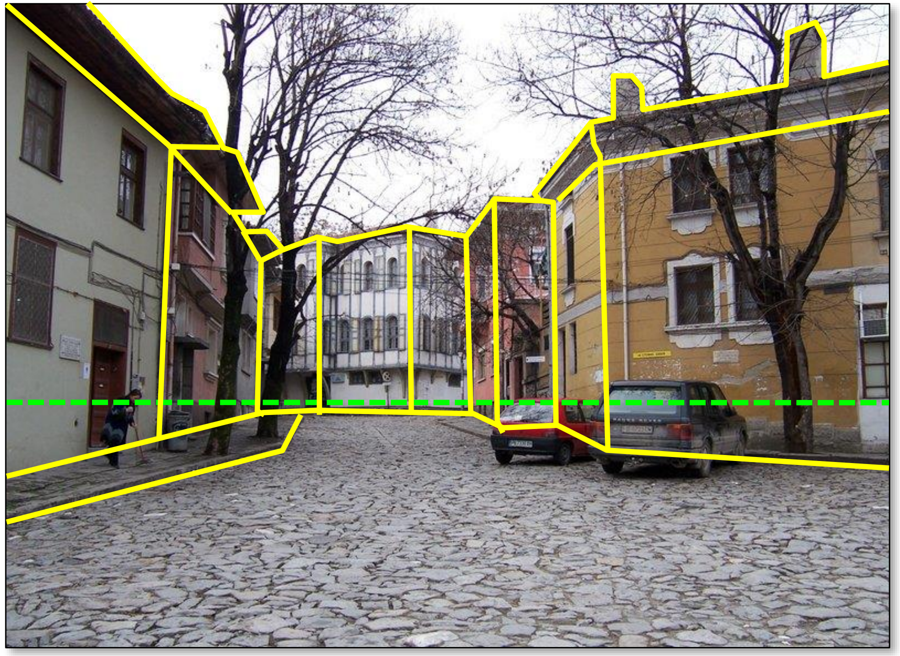
What is computer vision?

- “Computer vision […] uses **statistical methods** to disentangle data using models constructed with the aid of **geometry, physics, and learning theory**. […] Computer vision’s great trick is extracting **descriptions of the world from pictures** or sequences of pictures”. Forsyth, D. A., & Ponce, J. (2012).
- “How to **recover information** from the flood of data that comes **from eyes or cameras**” Russell & Norvig. *Artificial Intelligence: A Modern Approach*. Pearson. 2022.

What is computer vision?

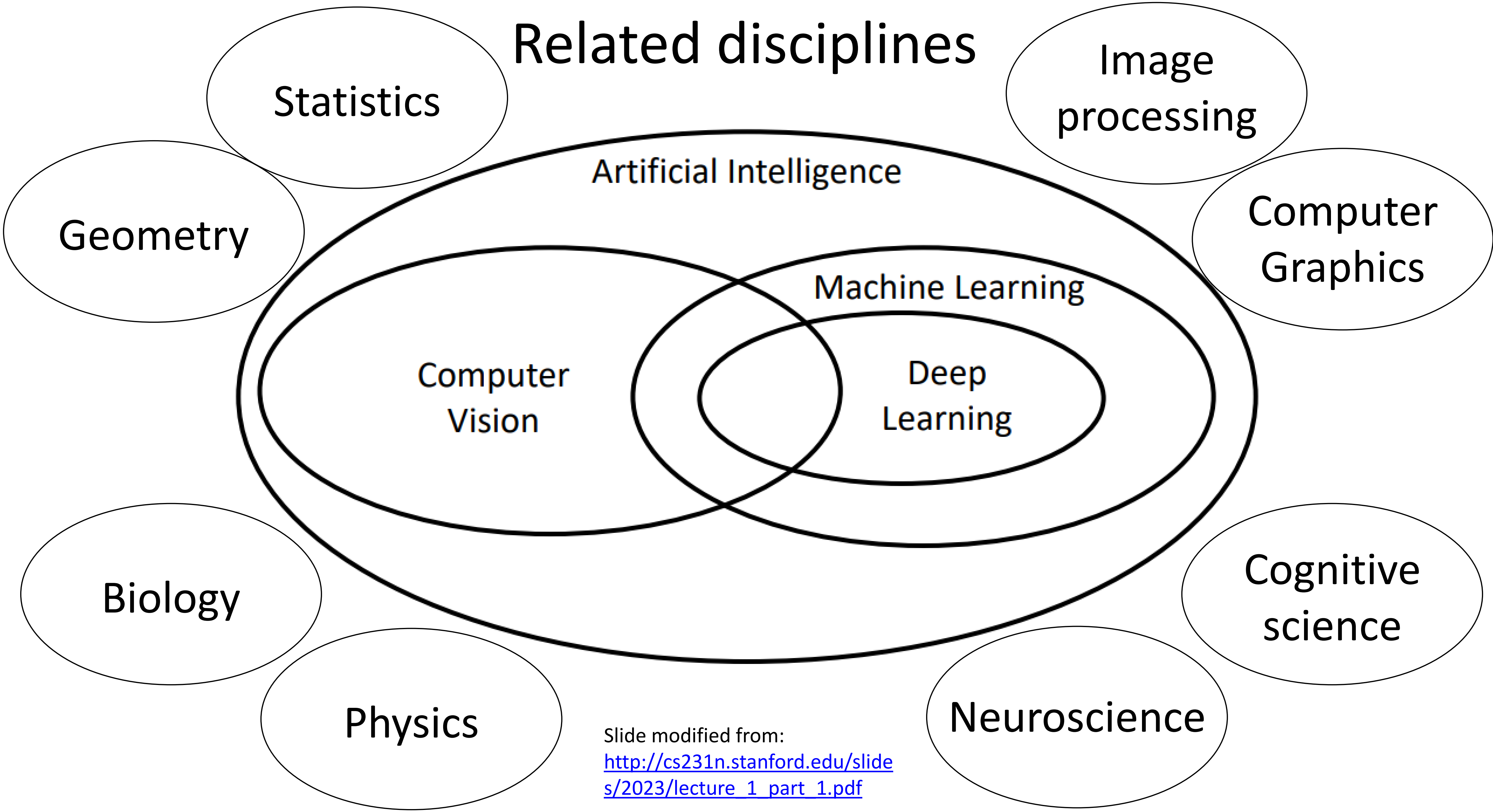
- Automatic understanding of images and videos
  - Computing properties of the 3D world from visual data *(measurement)*
  - Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities *(perception and interpretation)*

Semantic information



Geometric information

Related disciplines



Slide modified from:
http://cs231n.stanford.edu/slides/2023/lecture_1_part_1.pdf

Related disciplines

- Computer Vision vs
  - Computer Graphics (digital synthesis of visual content)

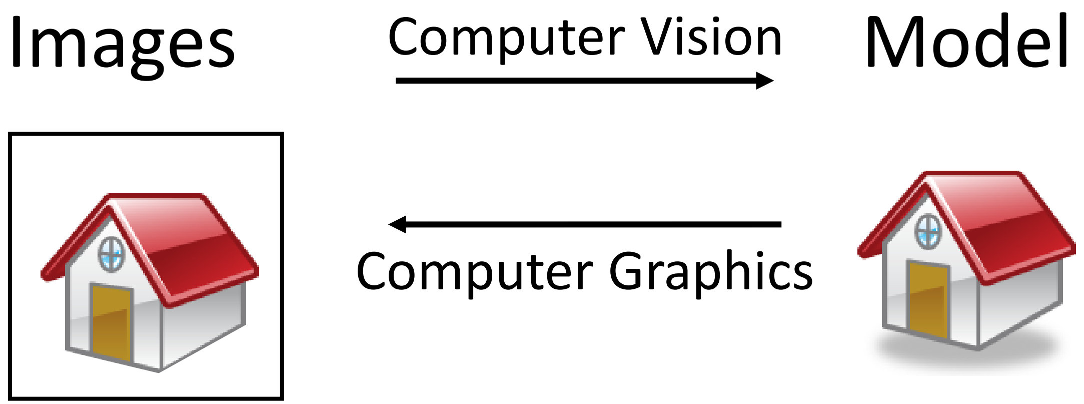
Image extracted from
https://upload.wikimedia.org/wikipedia/commons/5/5f/Utah_teapot_simple_2.png

  - Image Processing ("low level" operations on images)

Image extracted from
https://es.mathworks.com/help/images/ref/imgaussfilt.html

Note: "machine vision" many times refers to the use of computer vision in industrial environments (eg, automatic inspection, process control, etc.)

Vision and graphics



Inverse problems: analysis and synthesis.

Slide credit: “Algorithms and applications in Computer Vision”
(Lihi Zelnik-Manor; https://player.slideplayer.com/16/5020932/)

Low-, mid- and high-level processing

- Different authors employ different definitions and taxonomies
- Forsyth, D. A., & Ponce, J. (2012)

Higher level image representation

  - Early vision
    - Image filtering, texture analysis, Stereopsis, Structure from Motion
  - Mid-level vision
    - It involves coming up with image representations that are simultaneously compact and expressive (Image Segmentation, Tracking)
  - High-level vision
    - Image registration, Image Classification, Object Detection

- Sonka, M., Hlavac, V., & Boyle, R. (2015)

Higher level image representation

  - *Low-level image processing* methods usually use very little knowledge about the content of images.
    - Digital image processing: Image compression, pre-processing methods for noise filtering, edge extraction, and image sharpening
  - *High-level image understanding* tries to imitate human cognition. High-level data represent knowledge about the image content.

Low-, mid- and high-level processing

- Different authors employ different definitions and taxonomies
- Rafael C. González & Richard E. Woods (2008). Digital image processing. Pearson Prentice Hall. 3rd edition.
  - *Low-level processes* involve primitive operations (inputs and outputs are images)
    - image preprocessing to reduce noise, contrast enhancement, and image sharpening.
  - *Mid-level processing* involves (inputs are images but outputs are attributes extracted from those images (e.g., edges, contours, and the identity of individual objects))
    - segmentation, objects description, and objects classification/recognition
  - *Higher-level processing* involves "making sense" of an ensemble of recognized objects and performing the cognitive functions normally associated with vision.

Higher level image representation

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?

Slide credit: Noah Snavely
(https://www.cs.cornell.edu/courses/cs5670/2023sp/lectures/lec00_intro_for_web.pdf)

Can computers match (or beat) human perception?

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

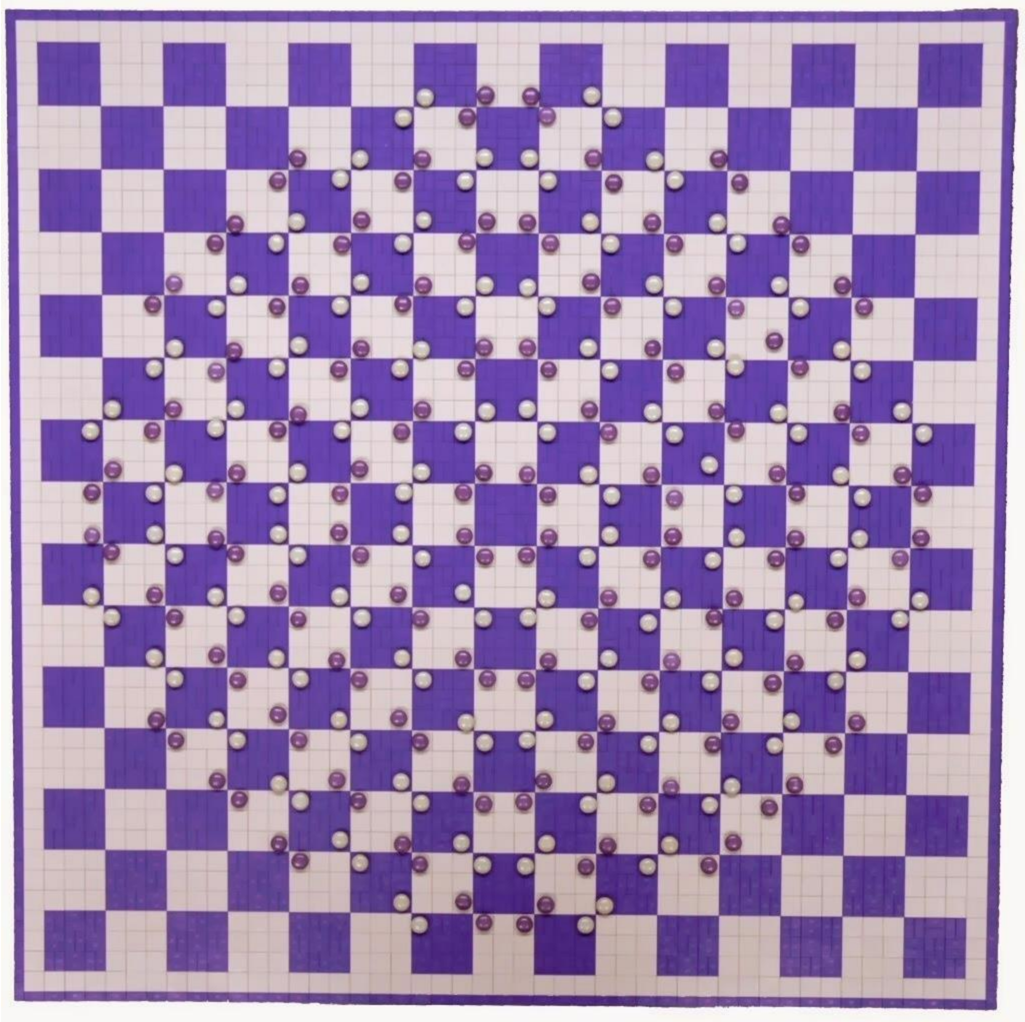
Slide modified from Noah Snavely (https://www.cs.cornell.edu/courses/cs5670/2023sp/lectures/lec00_intro_for_web.pdf)

Can computers match (or beat) human perception?

– **What is considered “hard” keeps changing**

“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**. ” Szeliski (2022)

Human perception has its shortcomings



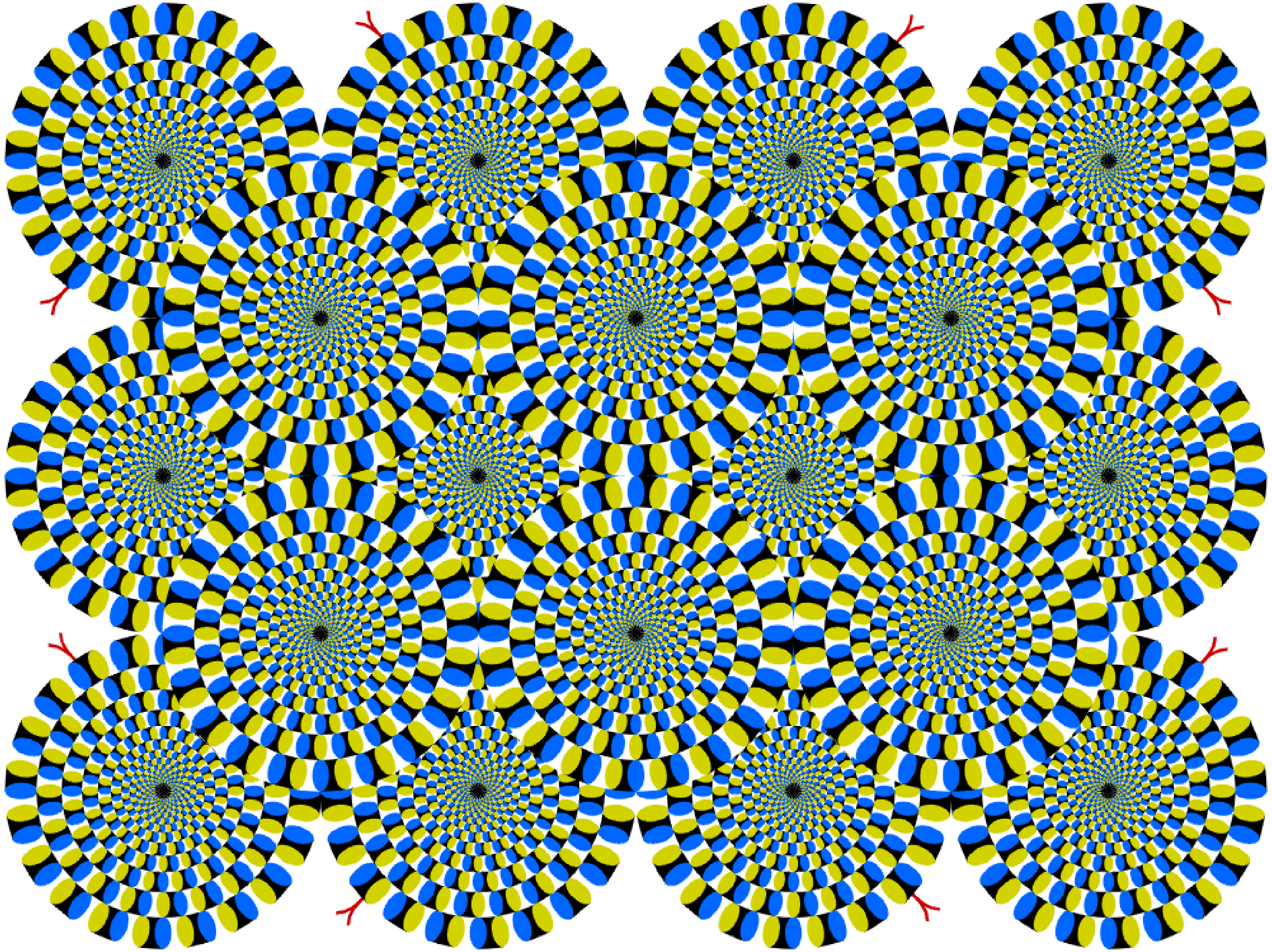
https://twitter.com/pickover/status/1460275132958662657/

Reality shatter. The physics, biology, and mathematics of human vision as it relates to geometrical patterns.

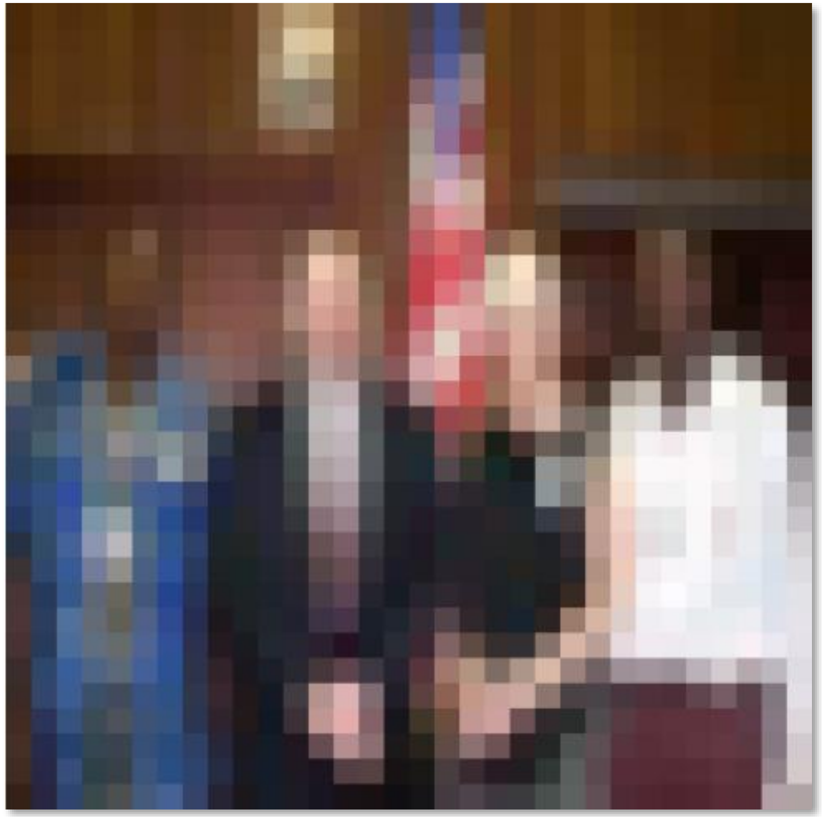
The lines on this checkerboard pattern are straight. (Optical illusion by Akiyoshi Kitaoka), Source: https://bit.ly/3wNxuGQ

Slide modified from Noah Snavely (https://www.cs.cornell.edu/courses/cs5670/2023sp/lectures/lec00_intro_for_web.pdf)

Human perception has its shortcomings



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



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

Slide credit: Noah Snavely
(https://www.cs.cornell.edu/courses/cs5670/2023sp/lectures/lec00_intro_for_web.pdf)

Why is computer vision difficult?

- We just take an example: Image Classification
  – Input: a single image
  – Output: Is there a cat in there? Yes/No?

→ **CAT**

Example taken from http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture2.pdf

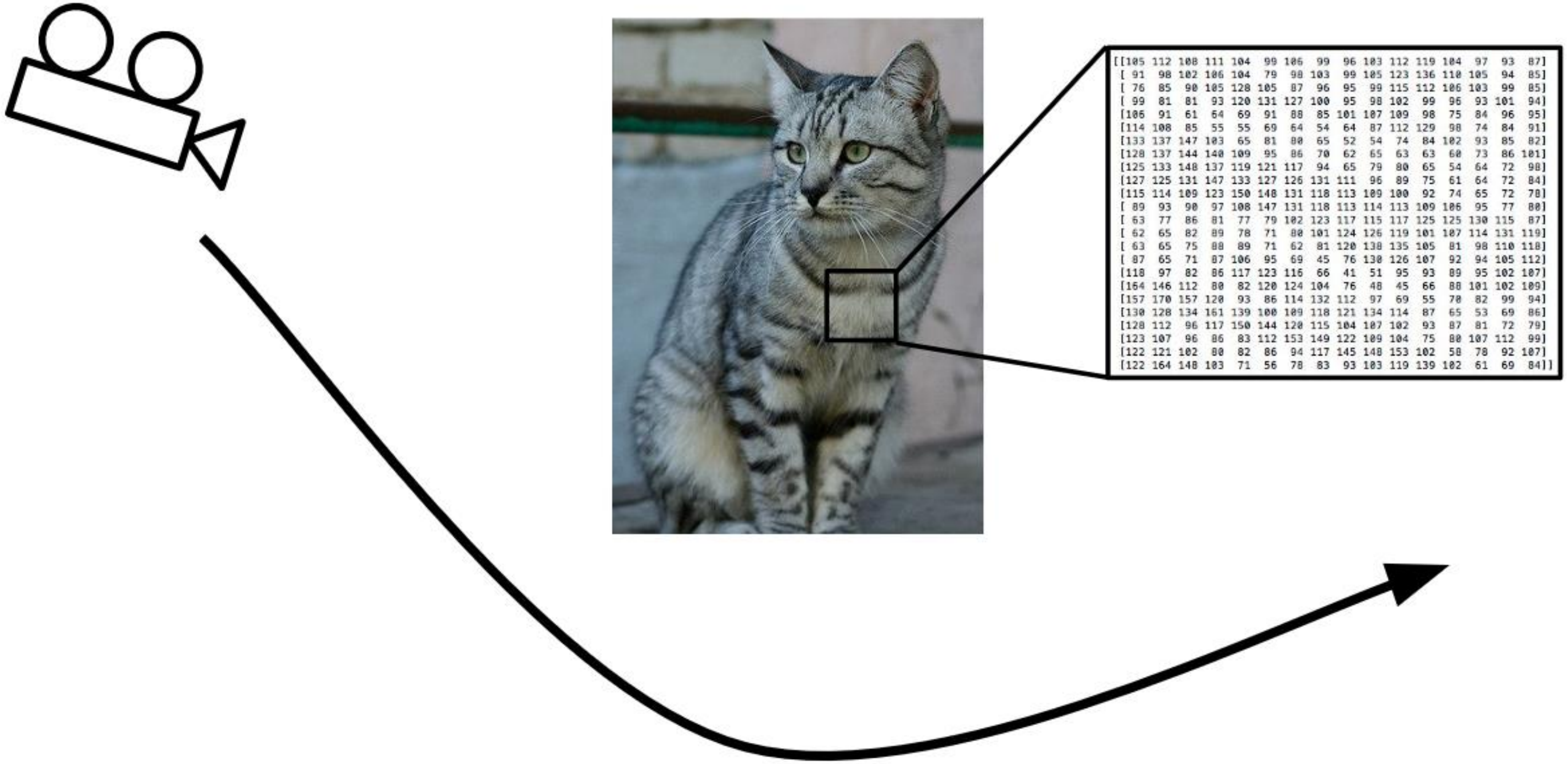
The problem: semantic gap


```
[[105 112 108 111 104  99 106  99  96 103 112 119 104  97  93  87]
 [ 91  98 102 106 104  79  98 103  99 105 123 136 110 105  94  85]
 [ 76  85  90 105 128 105  87  96  95  99 115 112 106 103  99  85]
 [ 99  81  81  93 120 131 127 100  95  98 102  99  96  93 101  94]
 [106  91  61  64  69  91  88  85 101 107 109  98  75  84  96  95]
 [114 108  85  55  55  69  64  54  64  87 112 129  98  74  84  91]
 [133 137 147 103  65  81  80  65  52  54  74  84 102  93  85  82]
 [128 137 144 140 109  95  86  70  62  65  63  63  60  73  86 101]
 [125 133 148 137 119 121 117  94  65  79  80  65  54  64  72  98]
 [127 125 131 147 133 127 126 131 111  96  89  75  61  64  72  84]
 [115 114 109 123 150 148 131 118 113 109 100  92  74  65  72  78]
 [ 89  93  90  97 108 147 131 118 113 114 113 109 106  95  77  80]
 [ 63  77  86  81  77  79 102 123 117 115 117 125 125 130 115  87]
 [ 62  65  82  89  78  71  80 101 124 126 119 101 107 114 131 119]
 [ 63  65  75  88  89  71  62  81 120 138 135 105  81  98 110 118]
 [ 87  65  71  87 106  95  69  45  76 130 126 107  92  94 105 112]
 [118  97  82  86 117 123 116  66  41  51  95  93  89  95 102 107]
 [164 146 112  80  82 120 124 104  76  48  45  66  88 101 102 109]
 [157 170 157 120  93  86 114 132 112  97  69  55  70  82  99  94]
 [130 128 134 161 139 100 109 118 121 134 114  87  65  53  69  86]
 [128 112  96 117 150 144 120 115 104 107 102  93  87  81  72  79]
 [123 107  96  86  83 112 153 149 122 109 104  75  80 107 112  99]
 [122 121 102  80  82  86  94 117 145 148 153 102  58  78  92 107]
 [122 164 148 103  71  56  78  83  93 103 119 139 102  61  69  84]]
```

We perceive an image
(photons processed by our brain).
Holistic view of a cat.

Our computer just "sees" a set of numbers (a matrix).
An image is just a big grid of numbers between [0, 255]:
e.g. 800 x 600 x 3 (3 channels RGB)

Challenges: Viewpoint variation



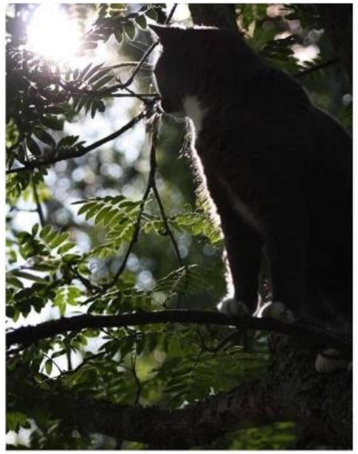
All pixels change when the camera moves!

Challenges: Illumination

This image is CC0 1.0 public domain

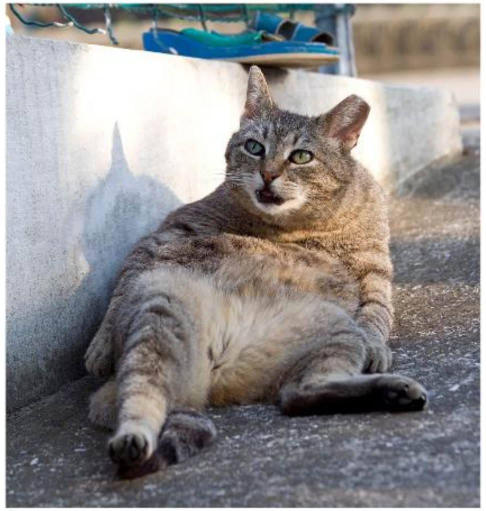
This image is CC0 1.0 public domain

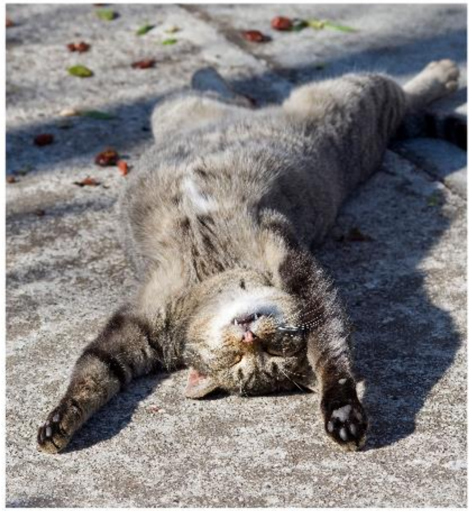
This image is CC0 1.0 public domain

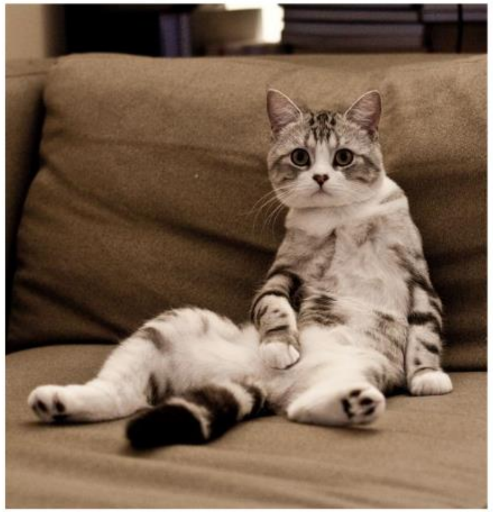


This image is CC0 1.0 public domain

Challenges: Deformation







Challenges: Intraclass variation






slide credit: Fei-Fei, Fergus & Torralba

Challenges: Background Clutter

This image is CC0 1.0 public domain

This image is CC0 1.0 public domain

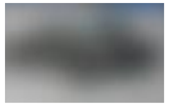
Challenges: Occlusion

Challenges: occlusion, clutter

Challenges: Motion

Source: S. Lazebnik

Challenges: local ambiguity



We need to process the images at different scale levels.

slide credit: Fei-Fei, Fergus & Torralba

Challenges: Inherent ambiguity and scale

- Many different 3D scenes could have given rise to a particular 2D picture

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

But there are lots of visual cues we can use…

Depth cues: Linear perspective

Depth cues: Aerial perspective

Depth ordering cues

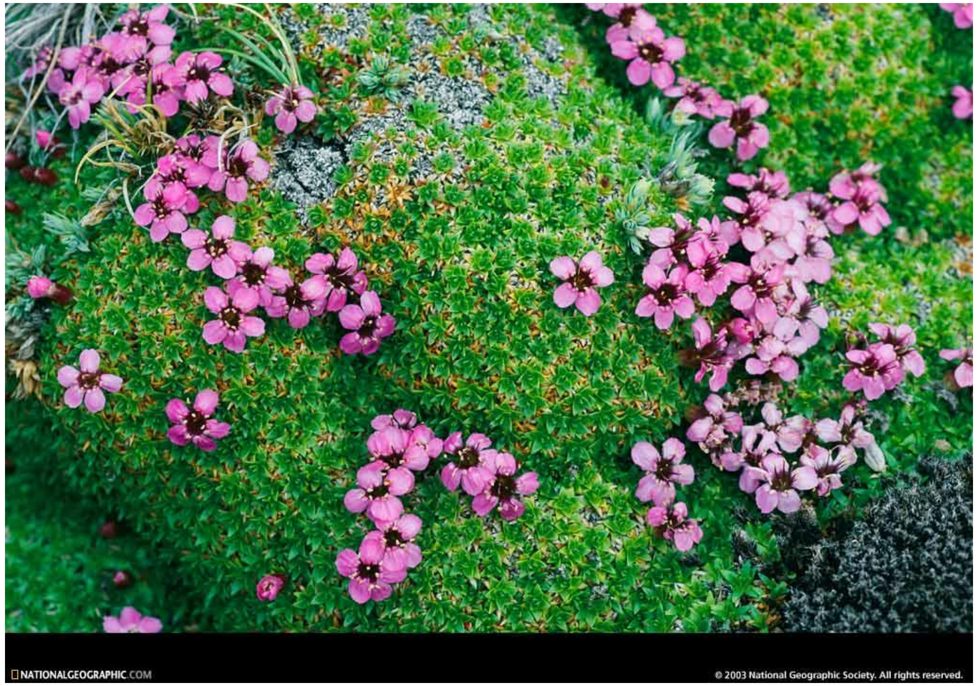
Shape cues: Texture gradient

Shape and lighting cues: Shading

Position and lighting cues: Cast shadows

But there are lots of visual cues we can use…

Grouping cues: Similarity (color, texture, proximity)



Grouping cues: “Common fate”

Why is computer vision difficult?

- Sonka et al. (2015) identifies six main elements:
   i. **Loss of information** in 3D → 2D

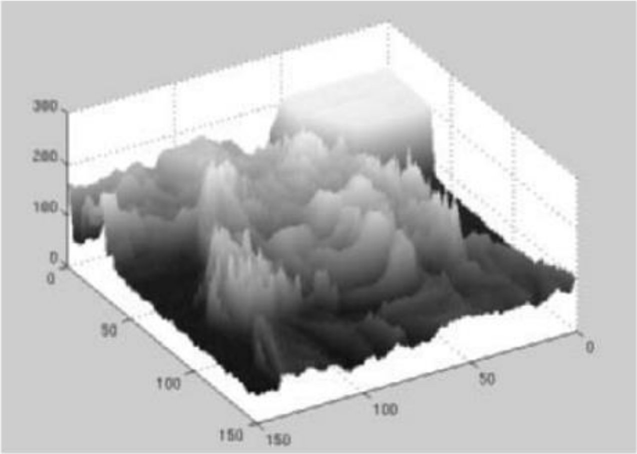
   ii. **Interpretation** is polysemic and requires management, acquisition and representation of knowledge

   iii. **Noise** inherently present in each measurement in the real world

   iv. **Too much data** (need for efficiency)

Why is computer vision difficult?

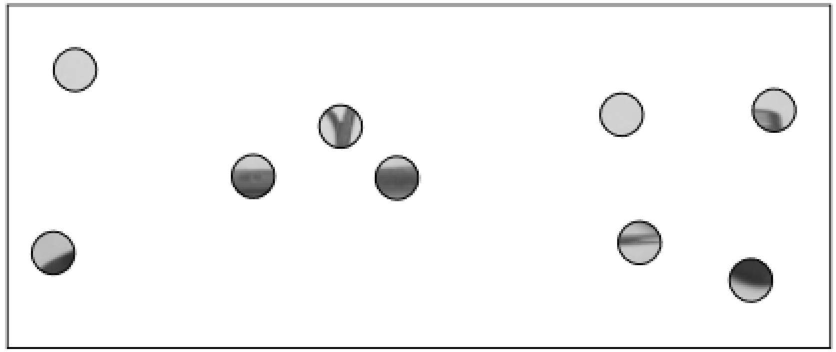
- Sonka et al. (2015) identifies six main elements:
  - v. We commonly operate with intensity images, but **brightness** is dependent on a number of factors (related to complicated image formation physics).
    - **Inverse tasks** (e.g., to reconstruct the aforementioned factors (like object surface reflectance properties, illumination properties, and object surface orientation) from intensity variations) **are ill-posed** in many cases (i.e. no unique solution).



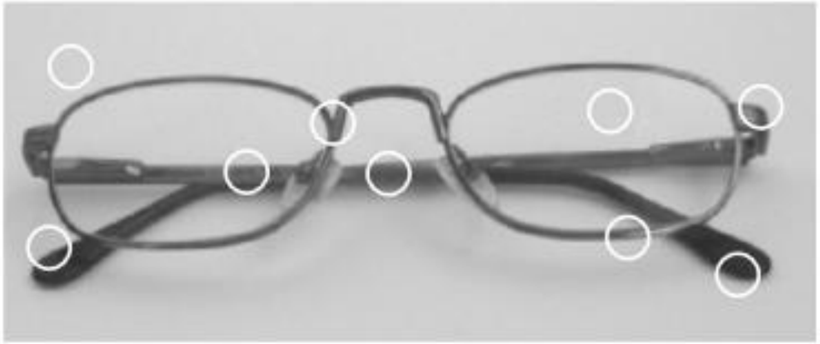
Why is computer vision difficult?

- Sonka et al. (2015) identifies six main elements:

vi. Local vs global view



It is very difficult to guess what object is depicted if you don't have a global view



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

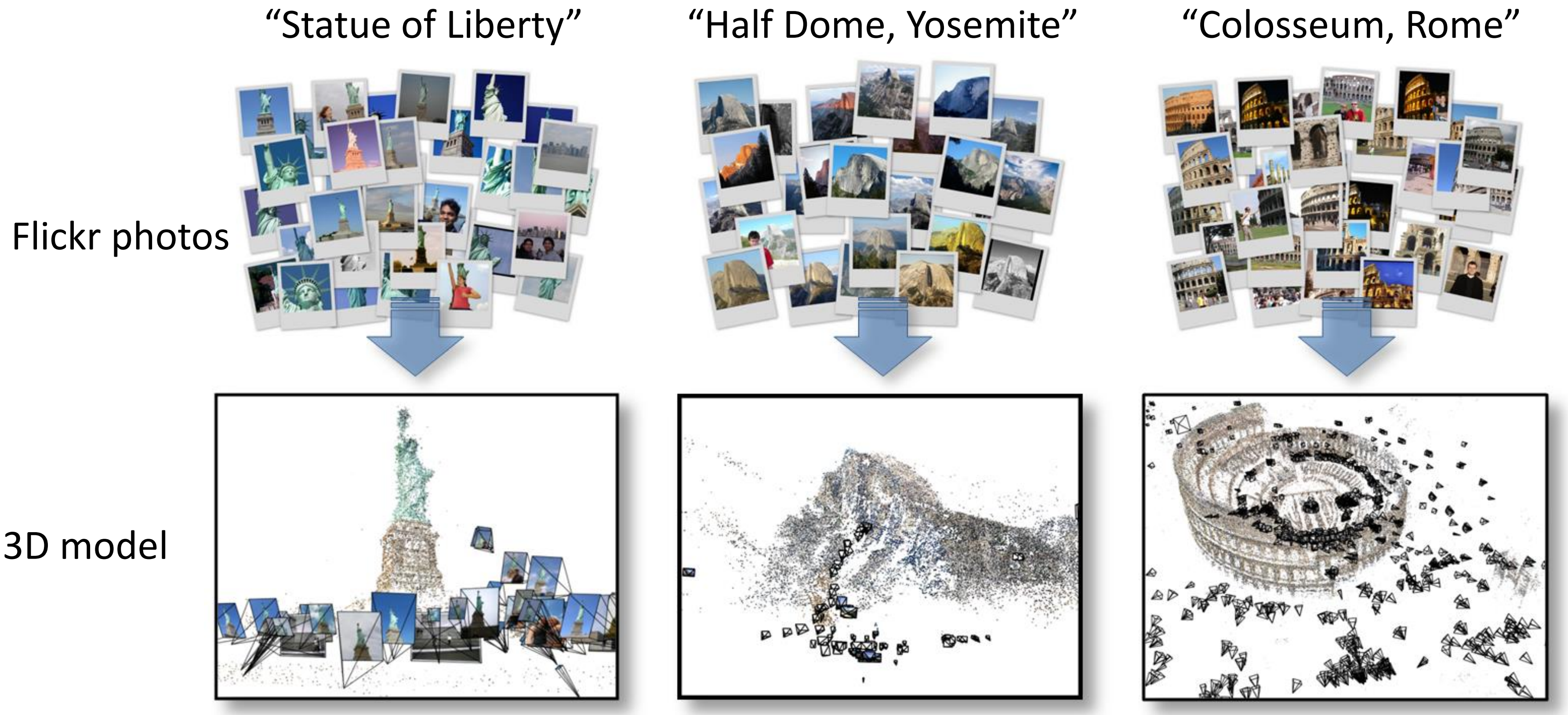
Oct 22, 2012

https://karpathy.github.io/2012/10/22/state-of-computer-vision/

- Bunch of people in a hallway.
- There are 3 mirrors in the scene so some of those people are "fake" replicas from different viewpoints.
- You recognize Obama […] he is in his suit […] he is surrounded by other people with suits.
- There's a person standing on a scale.
- Obama has his foot positioned just slightly on top of the scale.
- Obama is leaning in on the scale, which applies a force on it. Scale measures force that is applied on it => it will over-estimate the weight of the person standing on it.
- The person measuring his weight is not aware of Obama doing this.
- People are self-conscious about their weight.
- There are people in the back who find the person's imminent confusion funny.
- The perpetrator is the president, making it even a little more funnier.

Computer Vision Applications and Tasks

Automatic 3D reconstruction from Internet photo collections

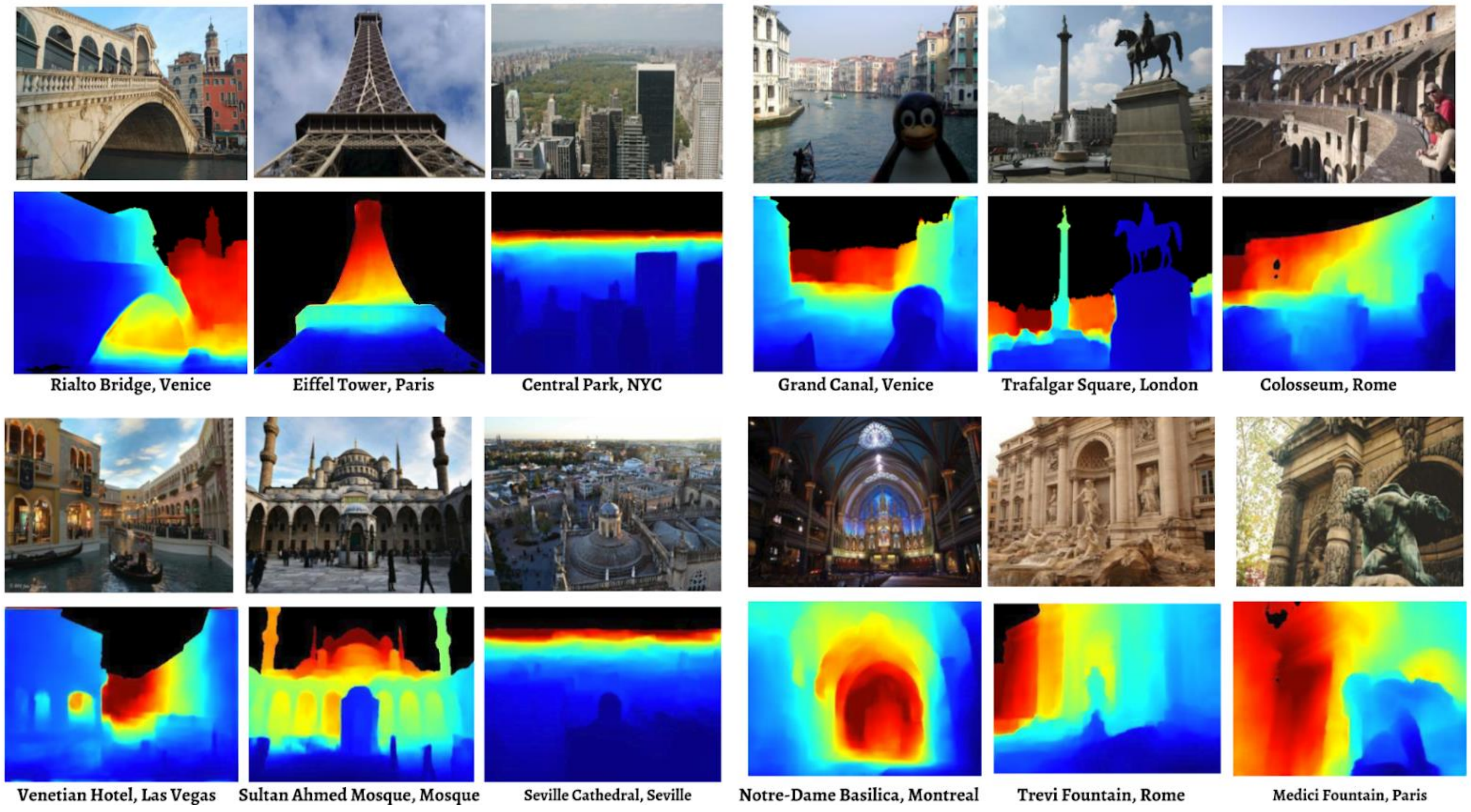


Snavely, Noah, et al. "Scene reconstruction and visualization from community photo collections." *Proceedings of the IEEE* 98.8 (2010): 1370-1390.

City-scale 3D reconstruction
(Large-scale Structure-from-Motion)

Reconstruction of Dubrovnik, Croatia, from ~40,000 images (Source: Noah Snavely)
https://www.youtube.com/watch?v=mUhYJnd1TqE

Depth estimation from a single image
(Monocular Depth Estimation)



Li, Zhengqi, and Noah Snavely. "Megadepth: Learning single-view depth prediction from internet photos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Improve photos (“Computational Photography”)

Super-resolution
(Yang, Wenming, et al. "Deep learning for single image super-resolution: A brief review." *IEEE Transactions on Multimedia* 21.12 (2019): 3106-3121)

Depth of field on cell phone camera (Picture without (left) and with (right) portrait mode; good way to draw the viewer's attention to a subject https://blog.research.google/2017/10/portrait-mode-on-pixel-2-and-pixel-2-xl.html)

Removing objects
(https://blog.google/products/photos/magic-eraser/)

Low-light photography
(Hasinoff, Samuel W., et al. "Burst photography for high dynamic range and low-light imaging on mobile cameras." *ACM Transactions on Graphics (ToG)* 35.6 (2016): 1-12)

Slide adapted from Noah Snavely

Self-driving cars

Source: https://analyticsindiamag.com/what-is-the-difference-between-computer-vision-and-image-processing/

Tesla AI Day: https://www.youtube.com/watch?v=j0z4FweCy4M&t=2814s

Image generation

“Progressive Growing of GANs for Improved Quality, Stability, and Variation” (Karras et al., 2017)

Which face is real?

Click on the person who is real.

https://www.whichfaceisreal.com/

Image generation (from text prompts)

color photograph of marie curie with a marvel ironman suit

color photograph of batman and queen elizabeth the second of united kingdom

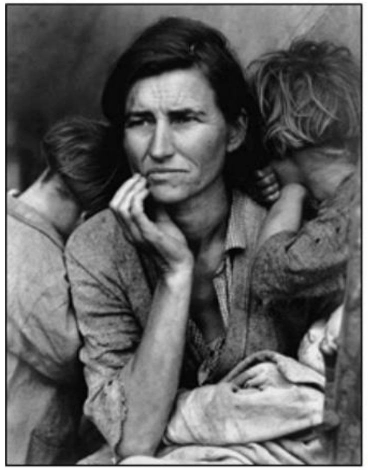
color photograph of a dog in a coral reef

https://stability.ai/blog/stable-diffusion-public-release

Style Transfer

“A Neural Algorithm of Artistic Style”
(Gatys et al., 2015)

Image colorization

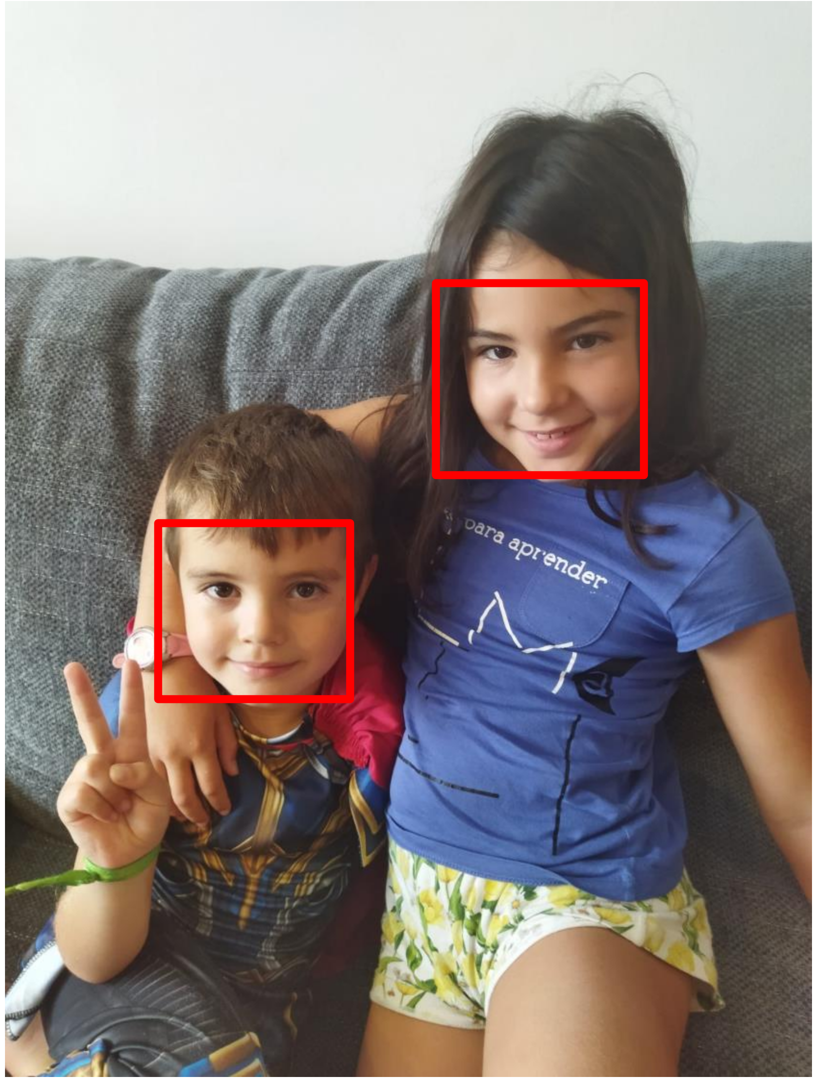


"Colorful Image Colorization" (Zhang et al., 2016)

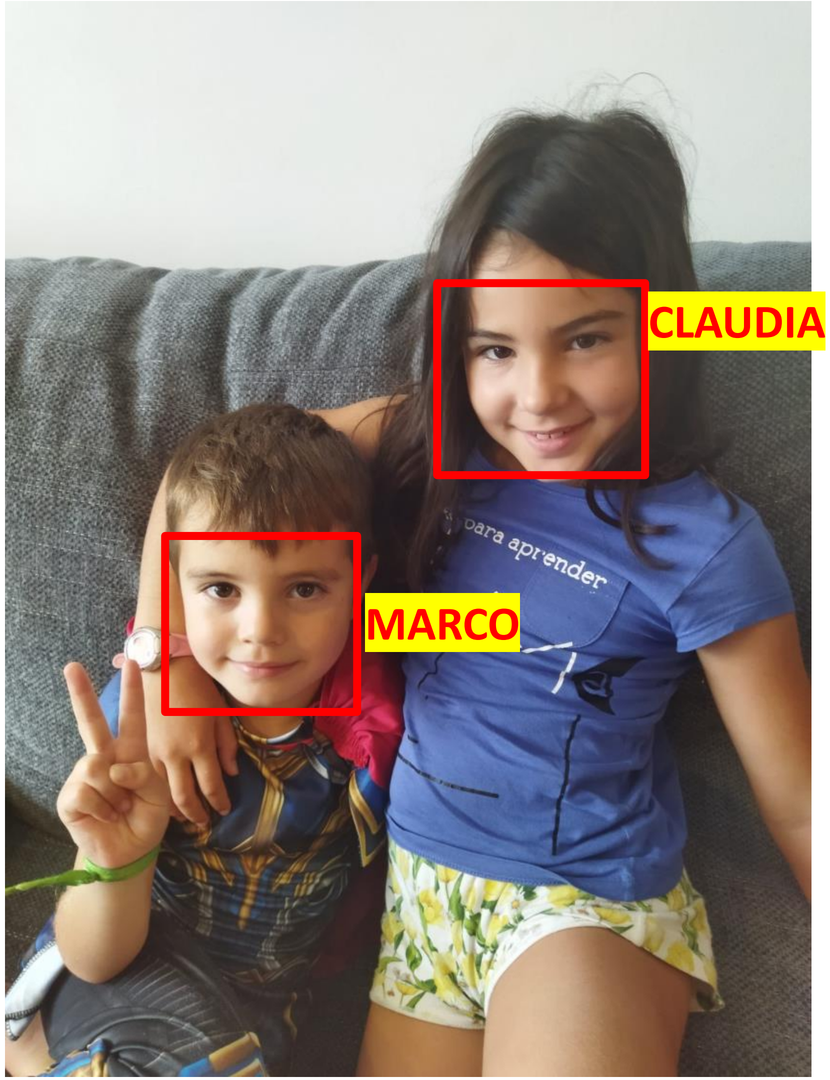
Image reconstruction/inpainting

"Image Inpainting for Irregular Holes Using Partial Convolutions" (Liu et al., 2018)

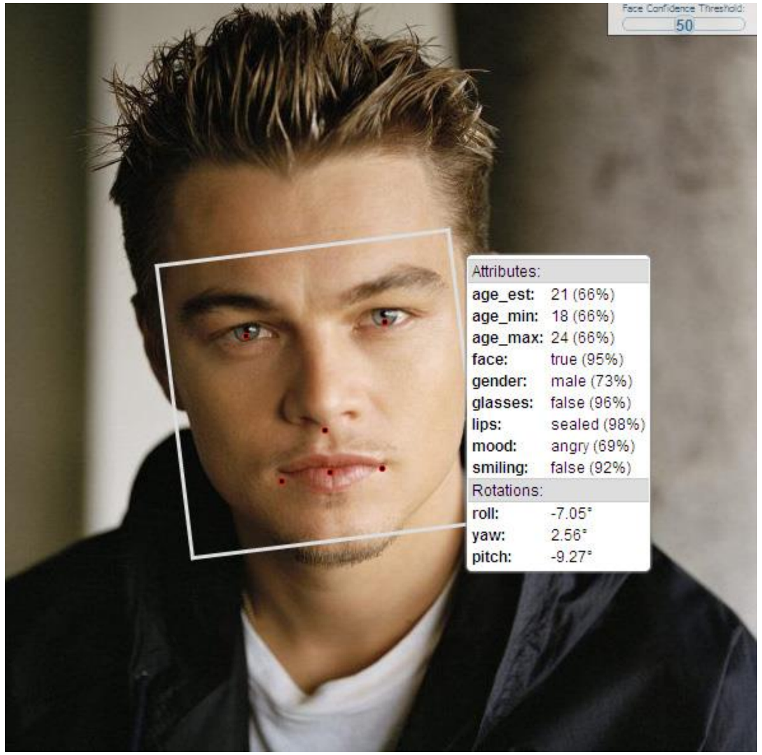
Face detection



Face recognition

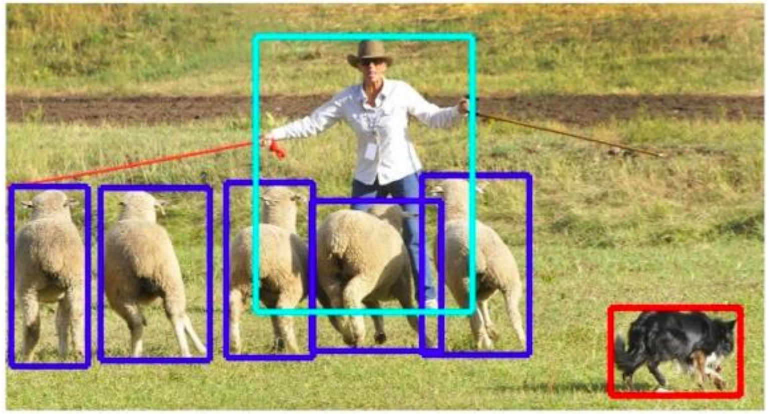


Face analysis

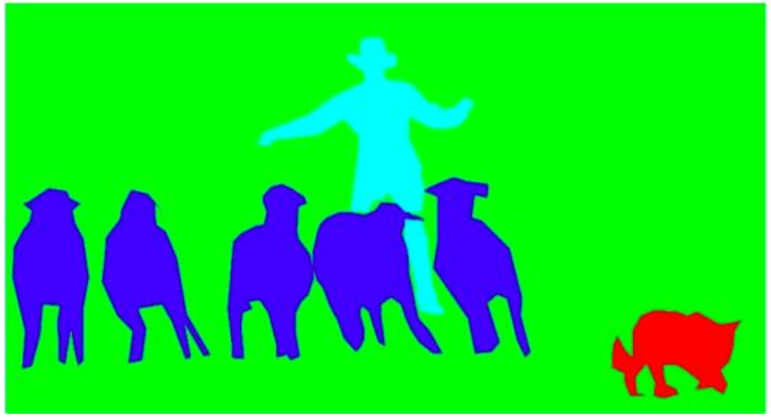


Detection, semantic segmentation, instance segmentation

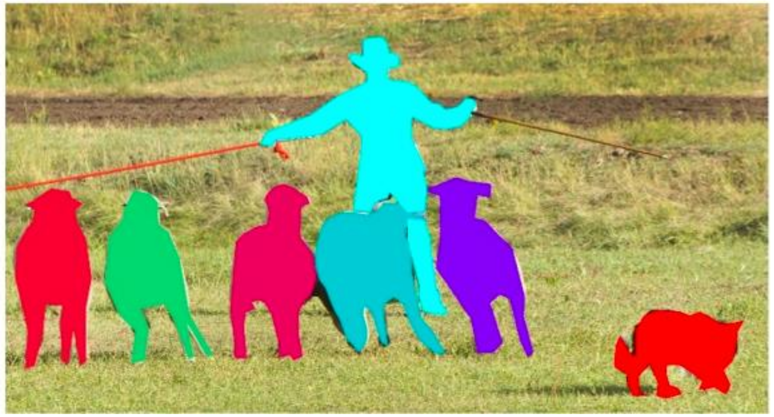
image classification



object detection



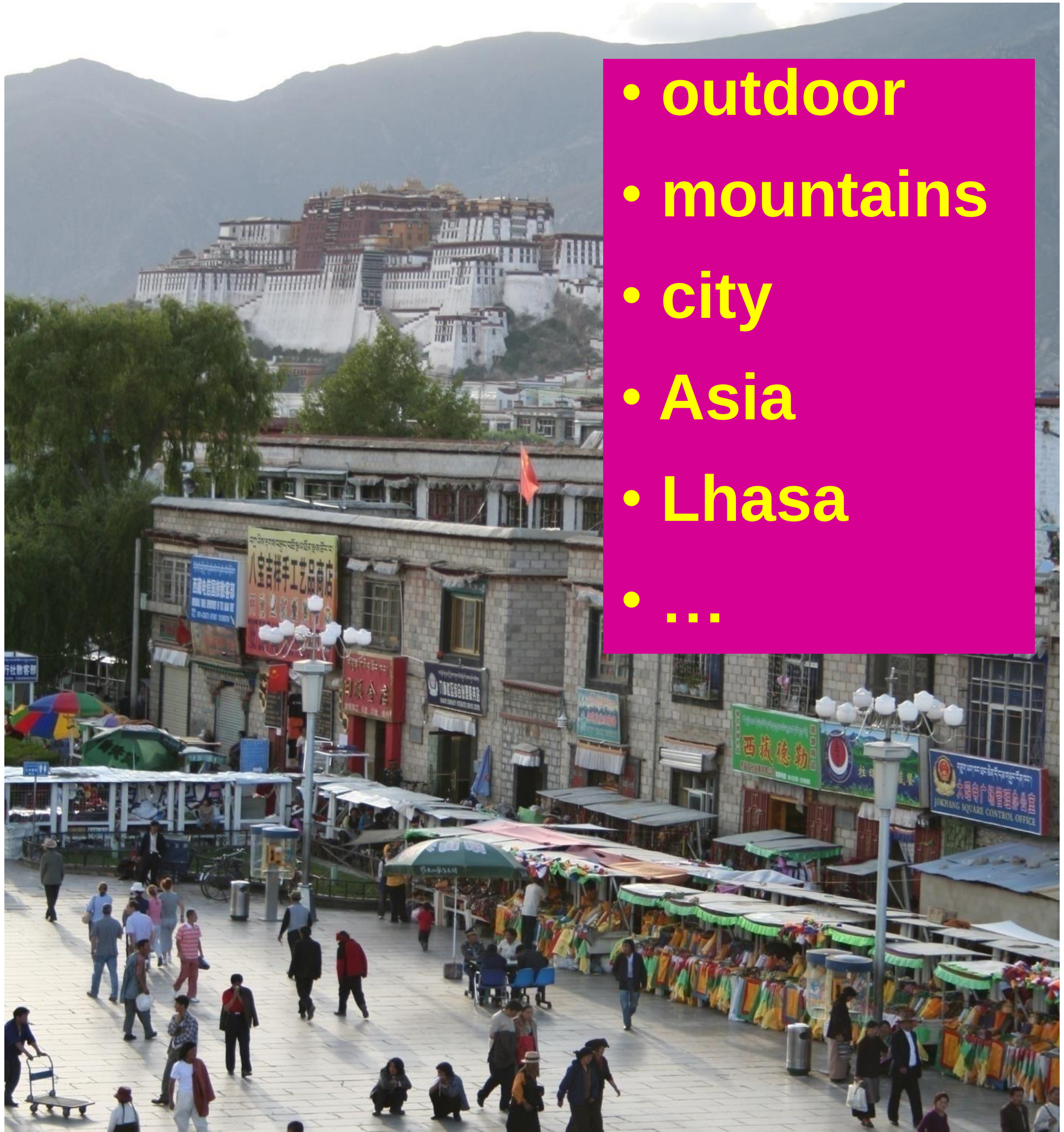
semantic segmentation



instance segmentation

Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer International Publishing, 2014.

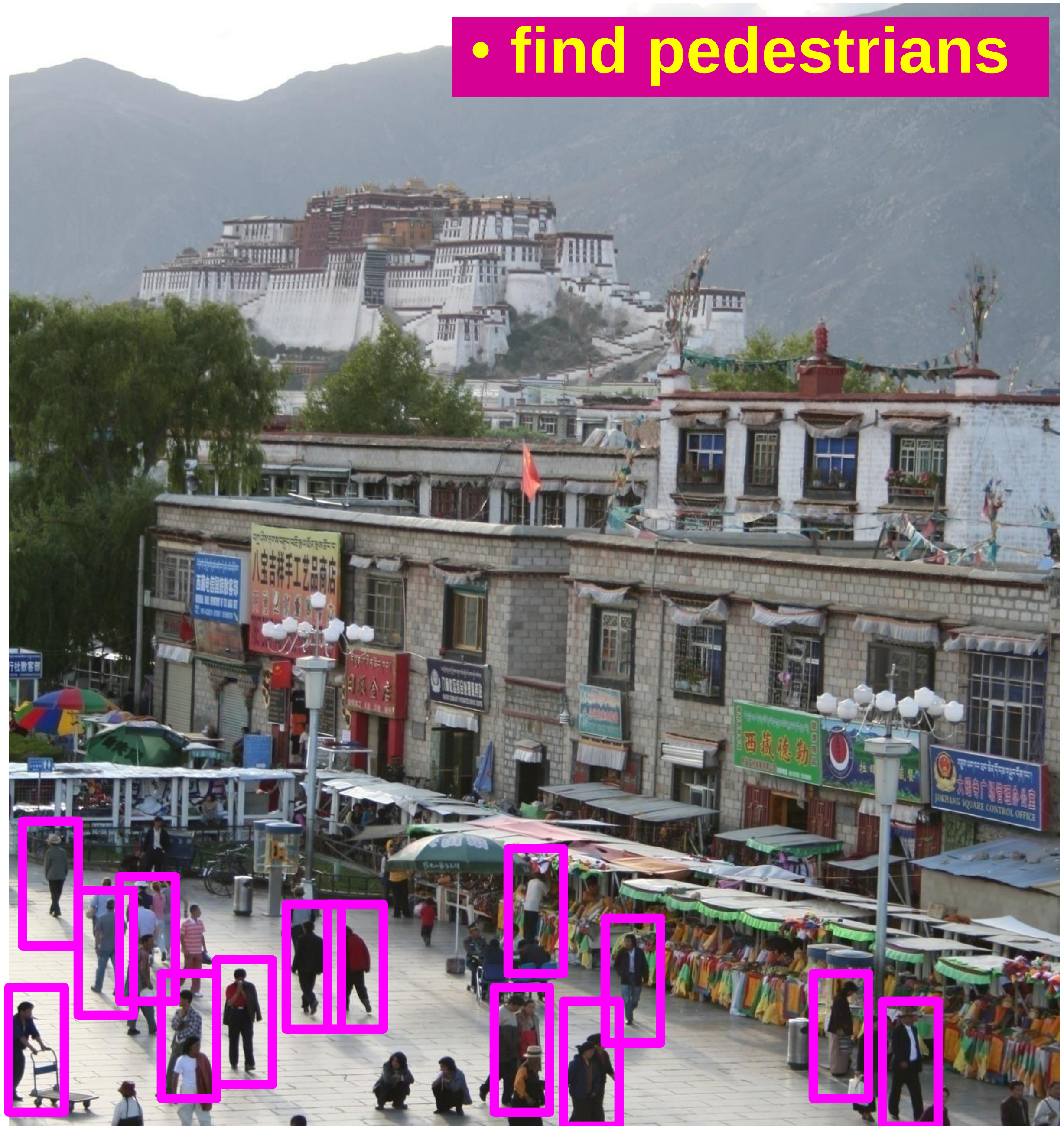
Image classification and tagging



- **outdoor**
- **mountains**
- **city**
- **Asia**
- **Lhasa**
- **…**

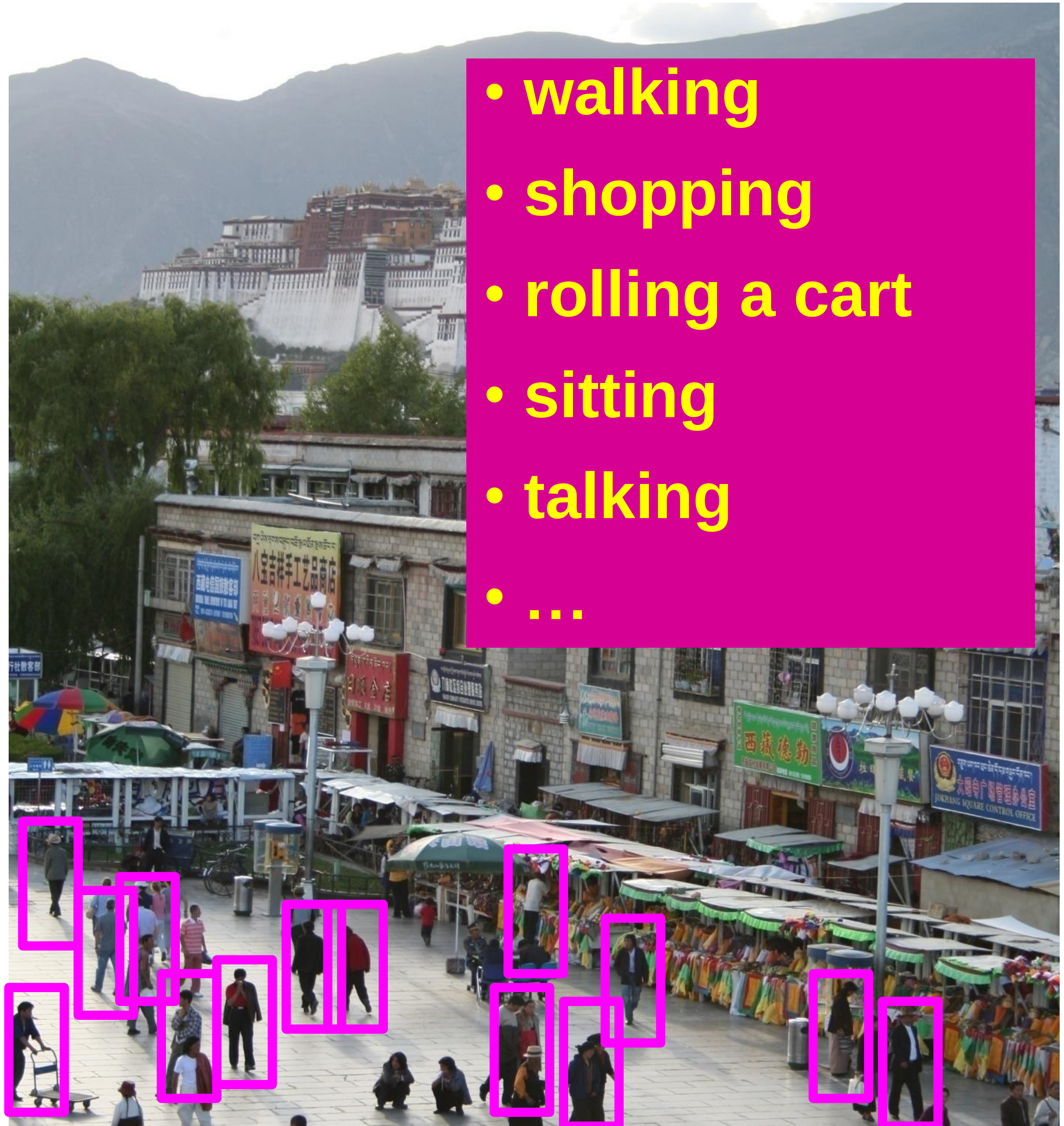
Adapted from
Fei-Fei Li

Object detection



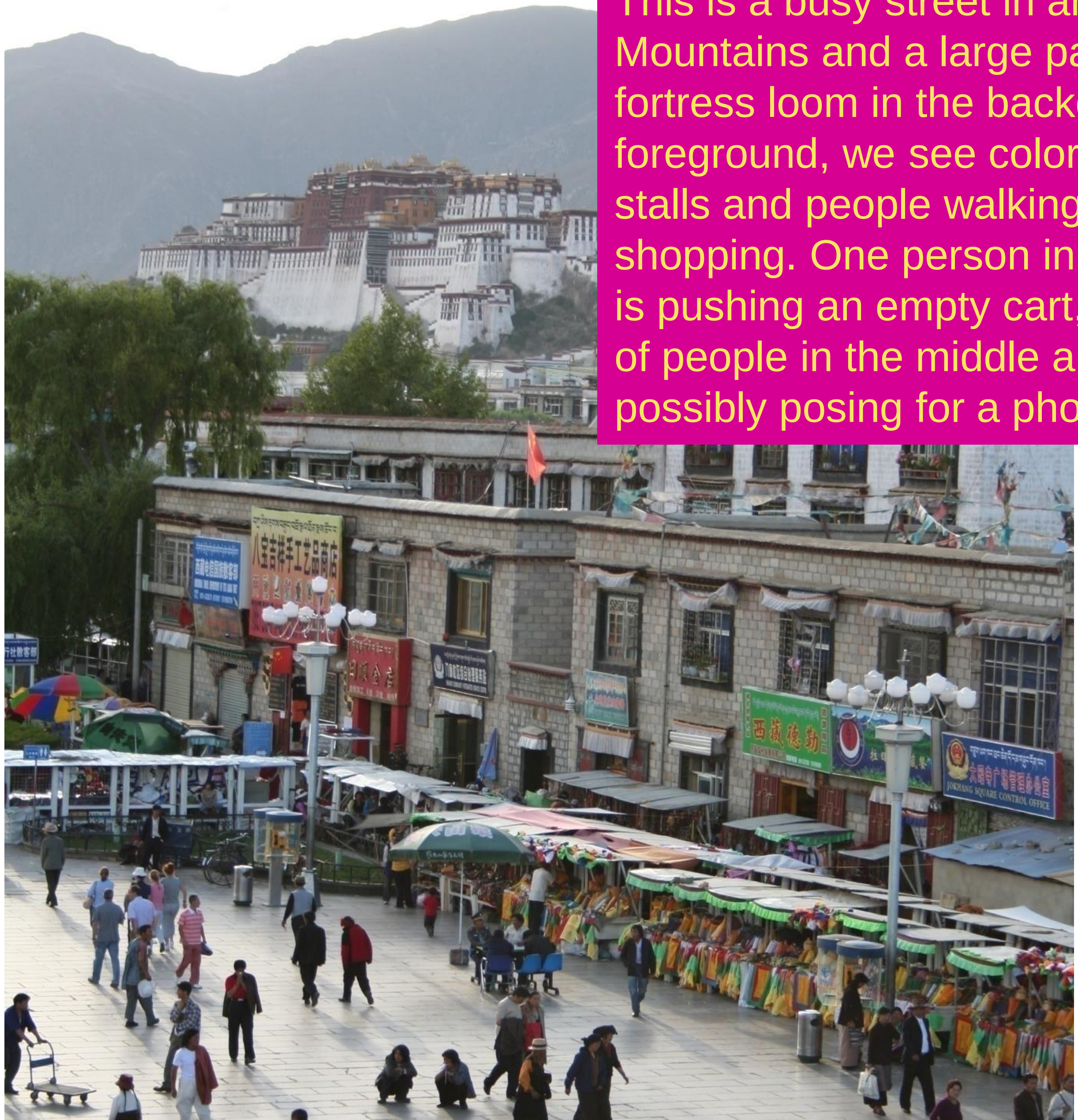
Adapted from
Fei-Fei Li

Activity recognition



Adapted from
Fei-Fei Li

Image description/captioning



This is a busy street in an Asian city. Mountains and a large palace or fortress loom in the background. In the foreground, we see colorful souvenir stalls and people walking around and shopping. One person in the lower left is pushing an empty cart, and a couple of people in the middle are sitting, possibly posing for a photograph.

Adapted from Fei-Fei Li

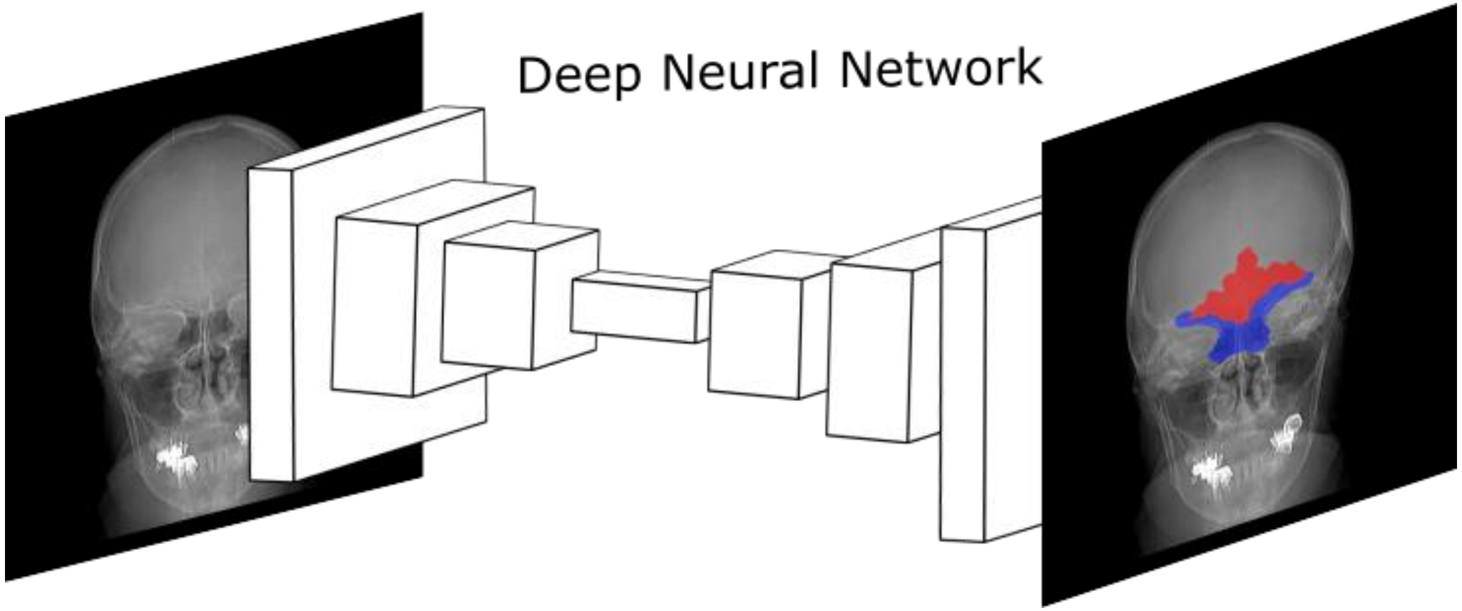
Human pose estimation in videos

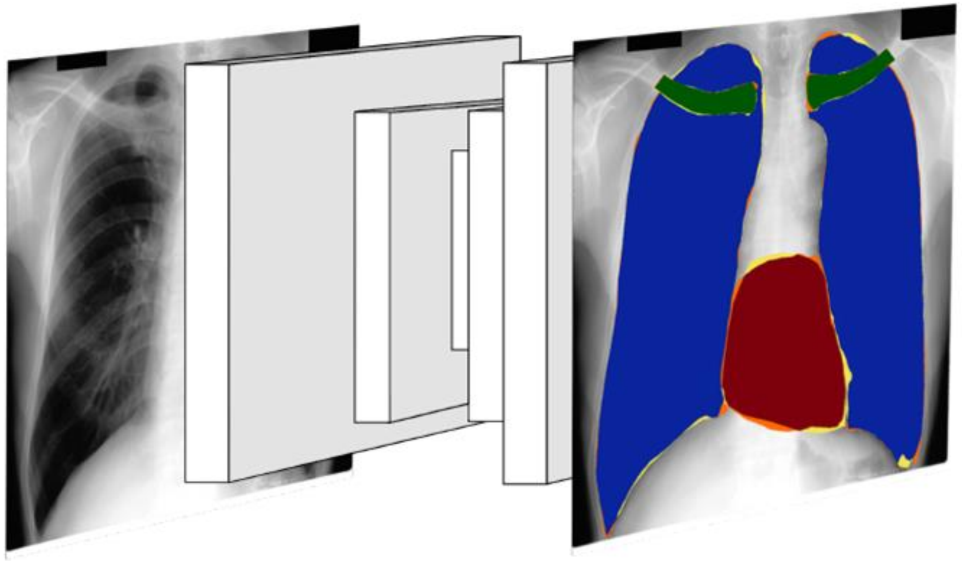
Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh

Carnegie Mellon University

Biomedical Image Segmentation

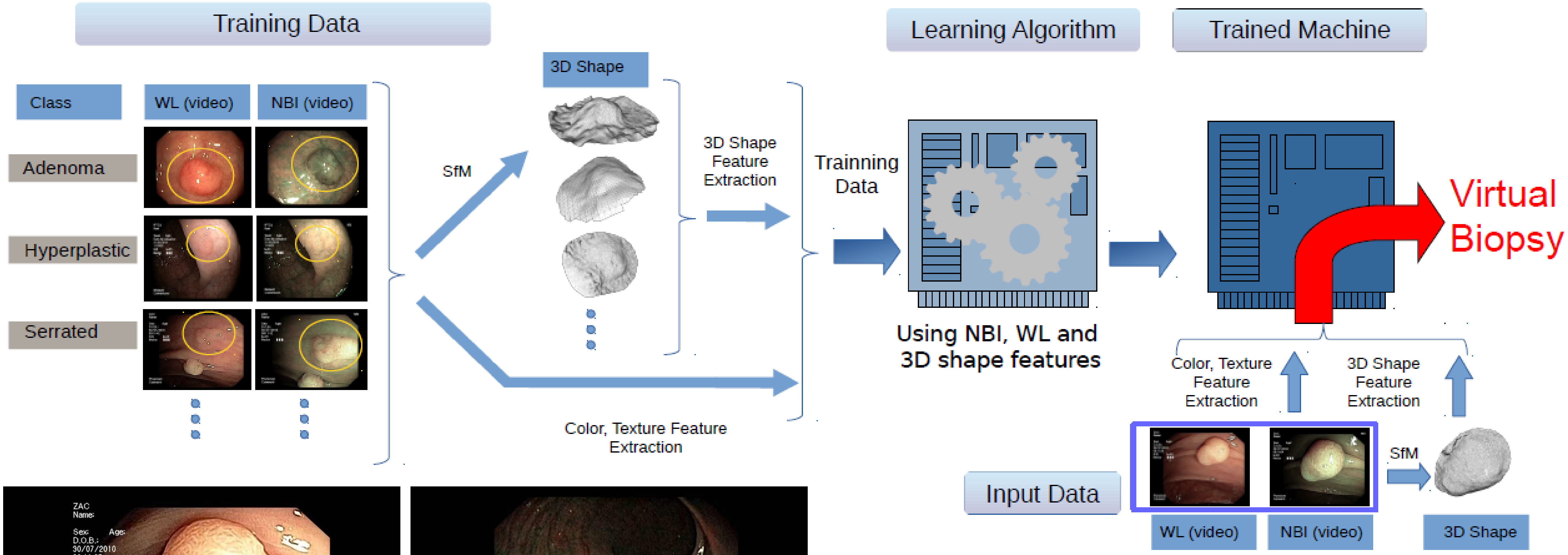


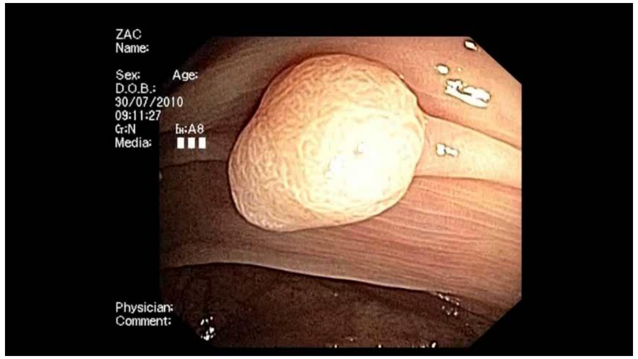


Gómez et al., "Deep architectures for high-resolution multi-organ chest X-ray image segmentation", Neural Computing & Applications 32, 15949-15963, Springer, October – 2019

Gómez et al. "Deep architectures for the segmentation of frontal sinuses in X-Ray images: towards an automatic forensic identification system in comparative radiography", Neurocomputing 456, 575-585, Elsevier, October - 2021

Biomedical Image Classification

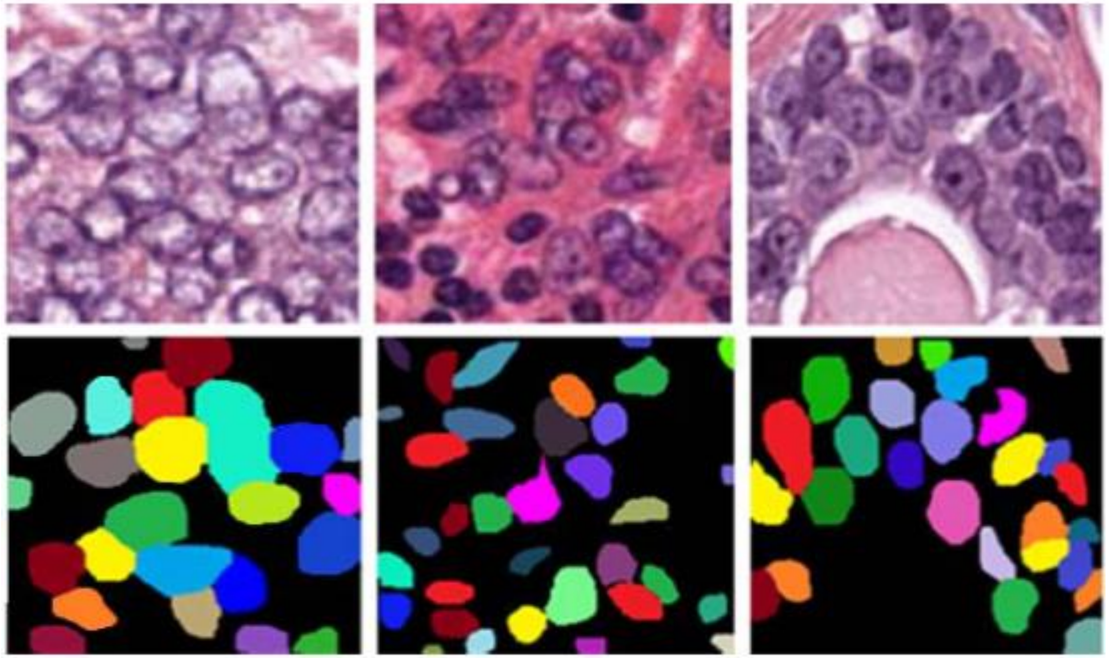




Mesejo et al., “Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy”, IEEE Transactions on Medical Imaging 35(9), 2051-2063, 2016

Biomedical Image Analysis

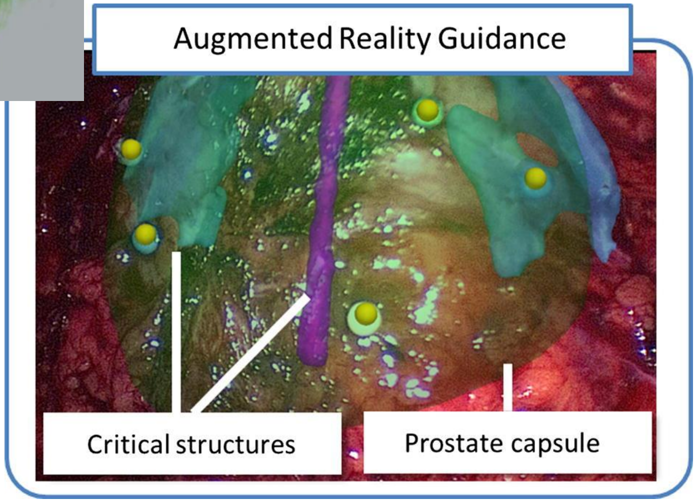
Computational Pathology



Kumar, Neeraj, et al. "A dataset and a technique for generalized nuclear segmentation for computational pathology." *IEEE transactions on medical imaging* 36.7 (2017): 1550-1560.

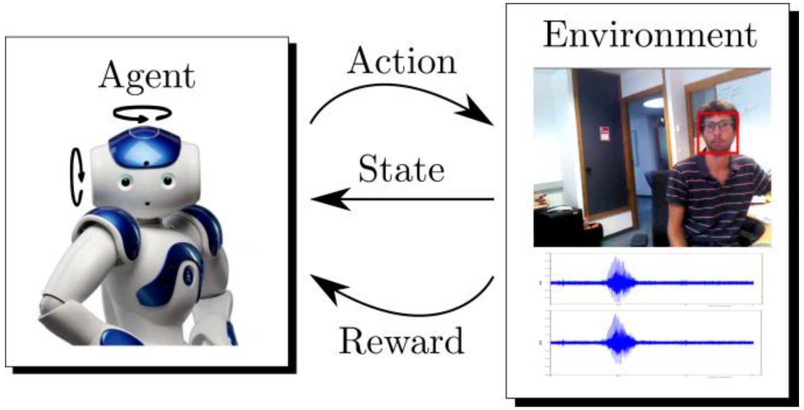
Grimson, W. Eric L., et al. "Image-guided surgery." *Scientific American* 280.6 (1999): 62-69.

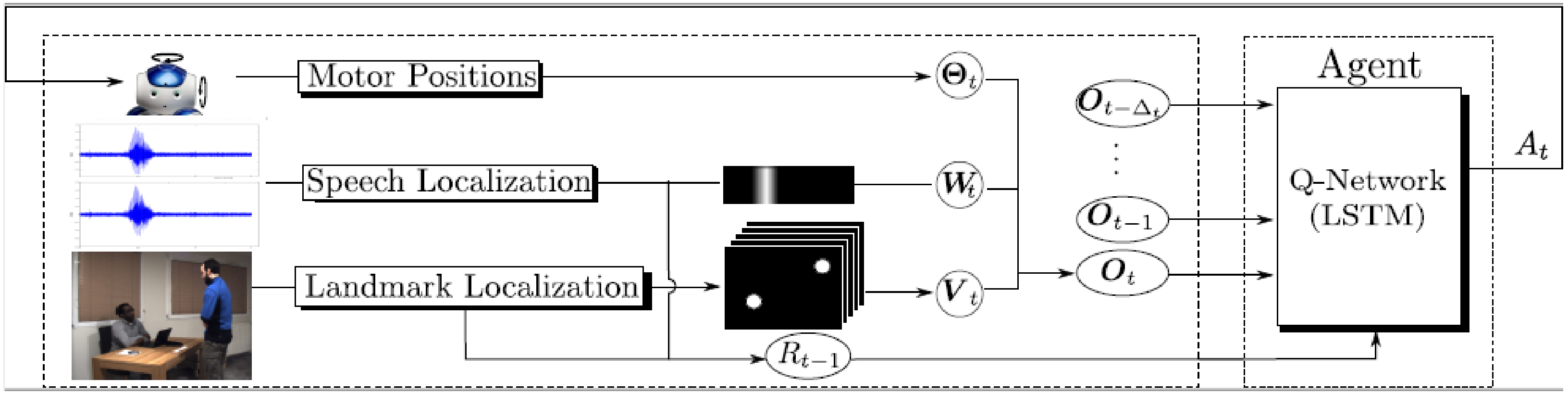
Image-guided Surgery



Maier-Hein, Lena, et al. "Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery." *Medical image analysis* 17.8 (2013): 974-996.

Audio-visual fusion for human-robot interaction

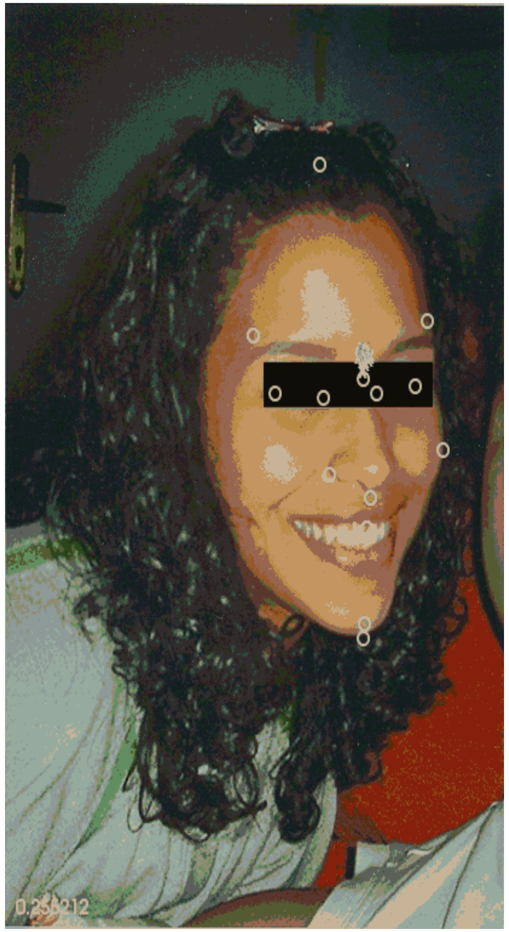




Lathuilière et al., “Neural Network-based Reinforcement Learning for Audio-Visual Gaze Control in Human-Robot Interaction”, Pattern Recognition Letters, 2019

Lathuilière et al., “Deep reinforcement learning for audio-visual gaze control”, IROS, 2018

Forensic Human Identification through Craniofacial Superimposition

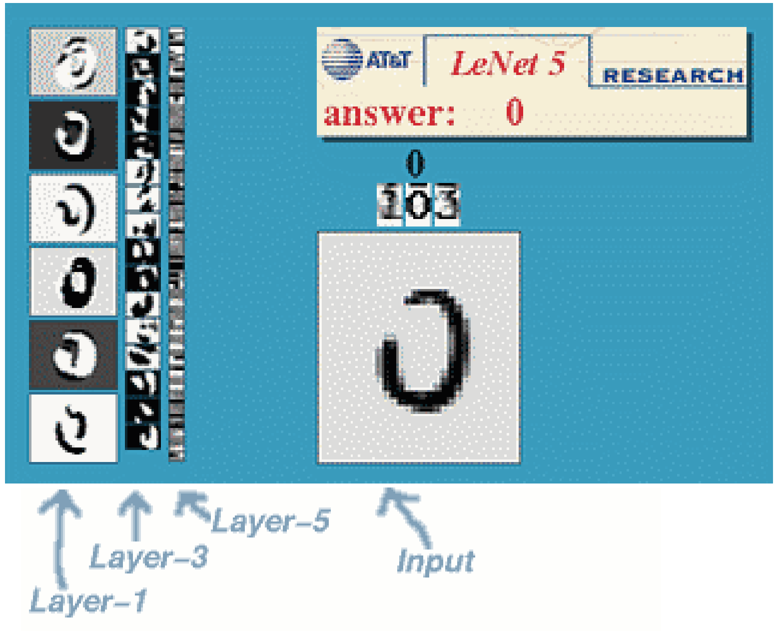


https://panacea-coop.com

Panacea
COOPERATIVE RESEARCH

https://skeleton-id.com/

Optical character recognition (OCR)



Digit recognition, AT&T labs (1980s)
http://yann.lecun.com/exdb/lenet/

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

Earth viewers (3D modeling)

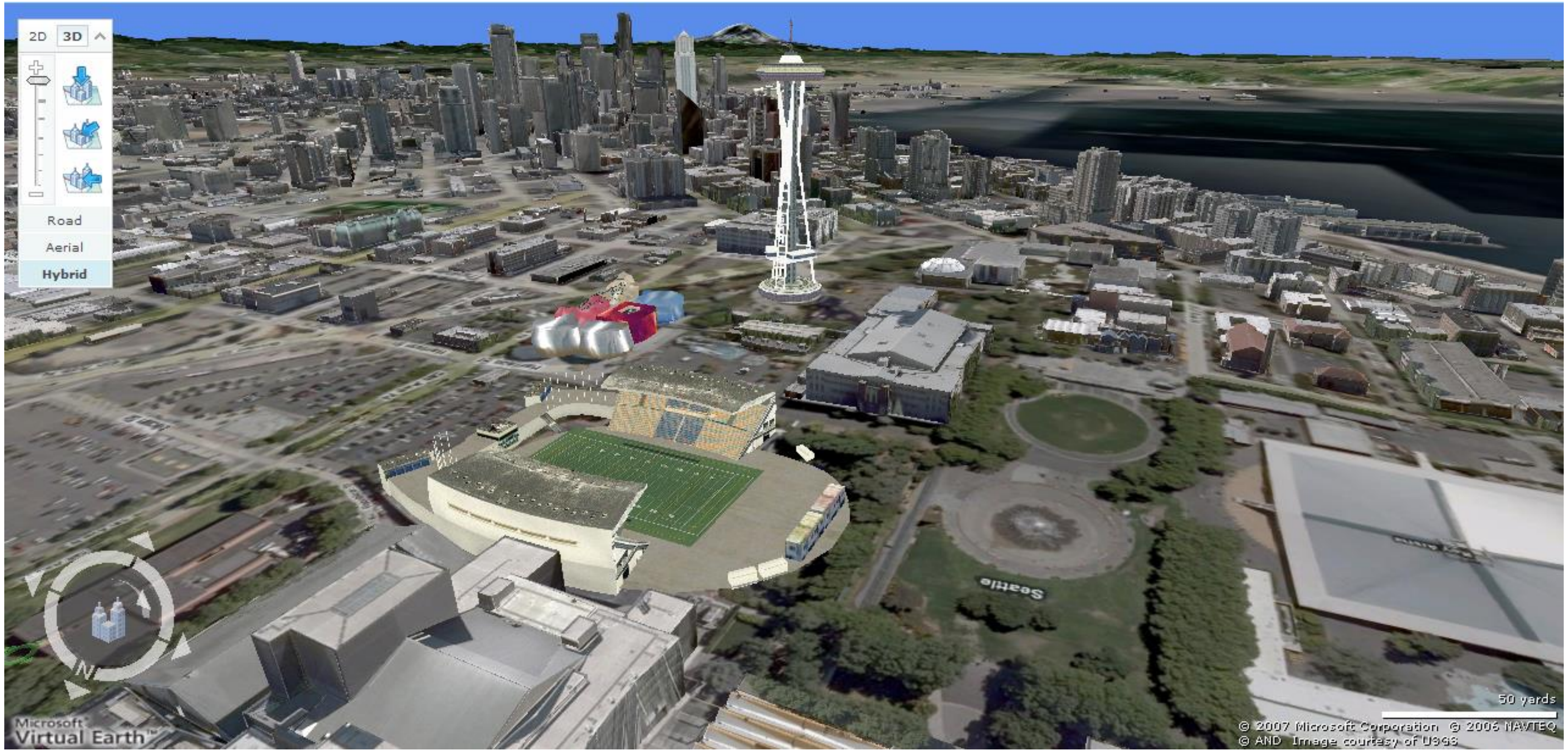


Image from Microsoft's Virtual Earth
(see also: Google Earth)

Special effects: shape and motion capture

Source: S. Seitz

Sports

Sportvision first down line
https://entertainment.howstuffworks.com/first-down-line.htm

Skillcorner Broadcast Tracking Technology
https://medium.com/skillcorner/a-new-world-of-performance-insight-from-video-tracking-technology-f0d7c0deb767

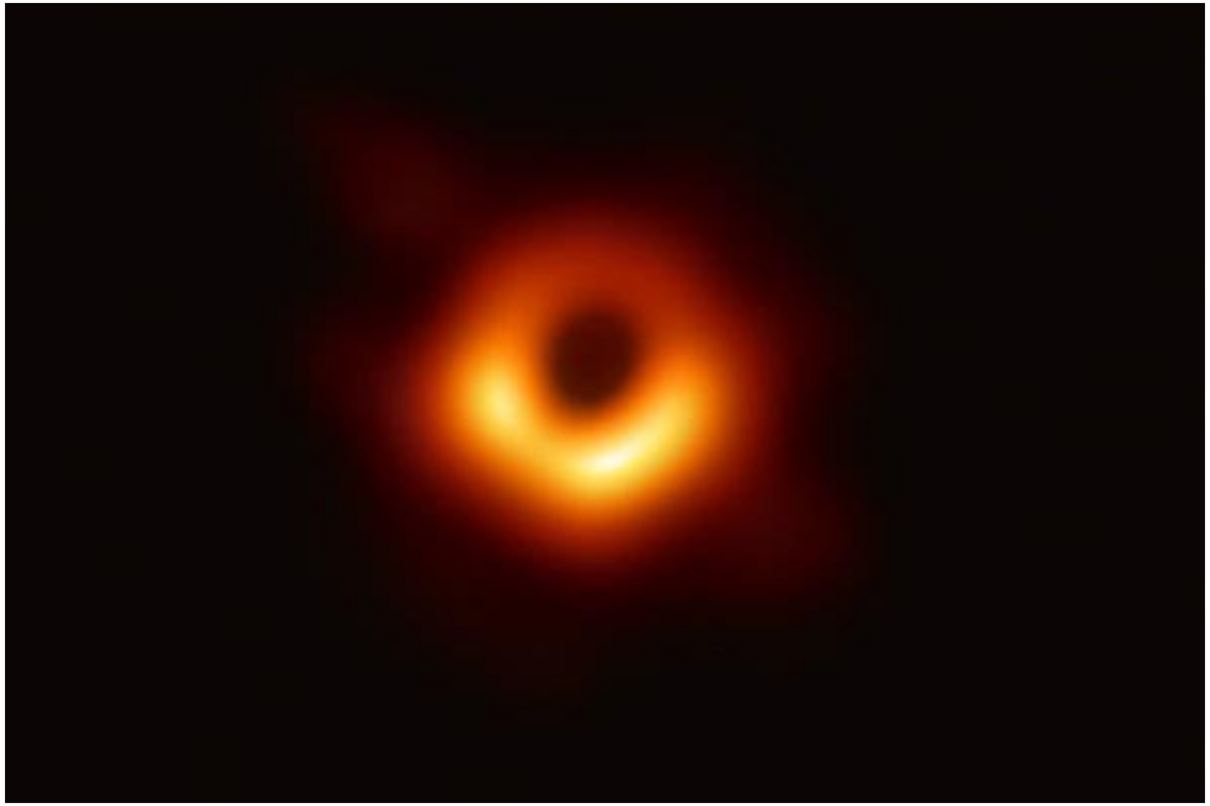
Astronomy

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



Black Holes | The Edge of All We Know (2020)
https://www.youtube.com/watch?v=2MLzrQ7dpIY

Bouman, Katherine L., et al. "Computational imaging for VLBI image reconstruction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

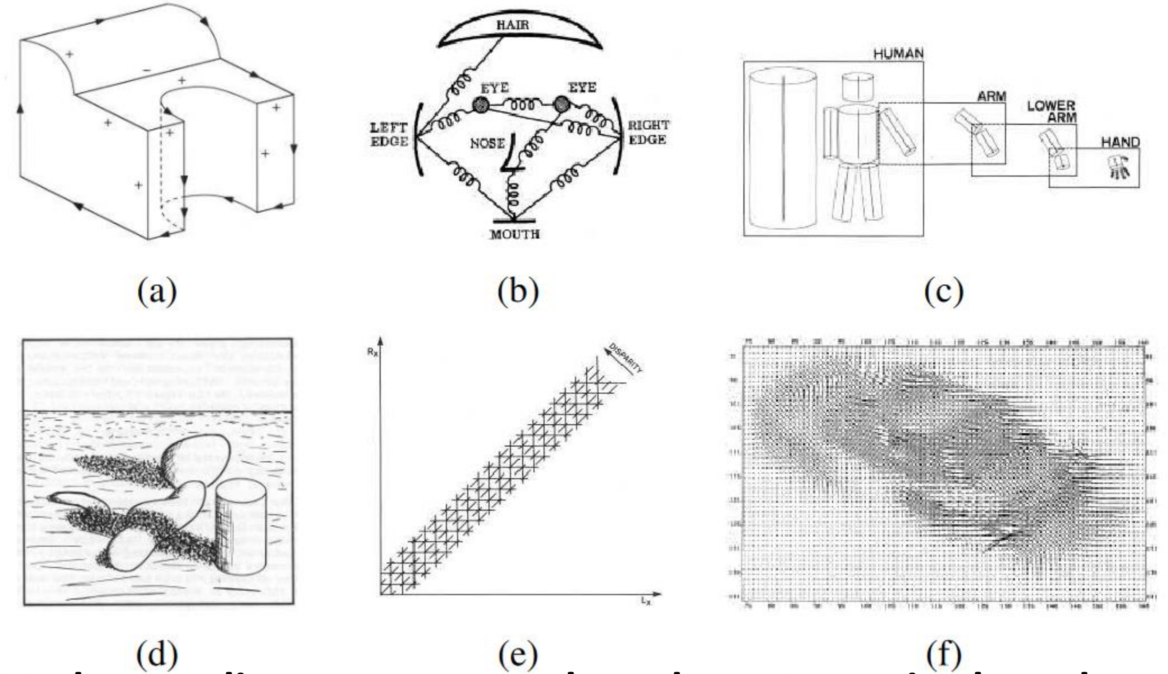
A bit of history

1960s-1970s

- It was initially considered an easy step
  - In 1966, Marvin Minsky at MIT asked his undergraduate student Gerald Jay Sussman to “spend the summer linking a camera to a computer and getting the computer to describe what it saw” (Boden 2006, p. 781)
- recover the **three-dimensional structure** of the world **from images** and to use this as a stepping stone **towards full scene understanding**

Edge extraction (line labeling)

Three-dimensional modeling of objects



Understanding intensities and shading variations

Feature-based stereo correspondence

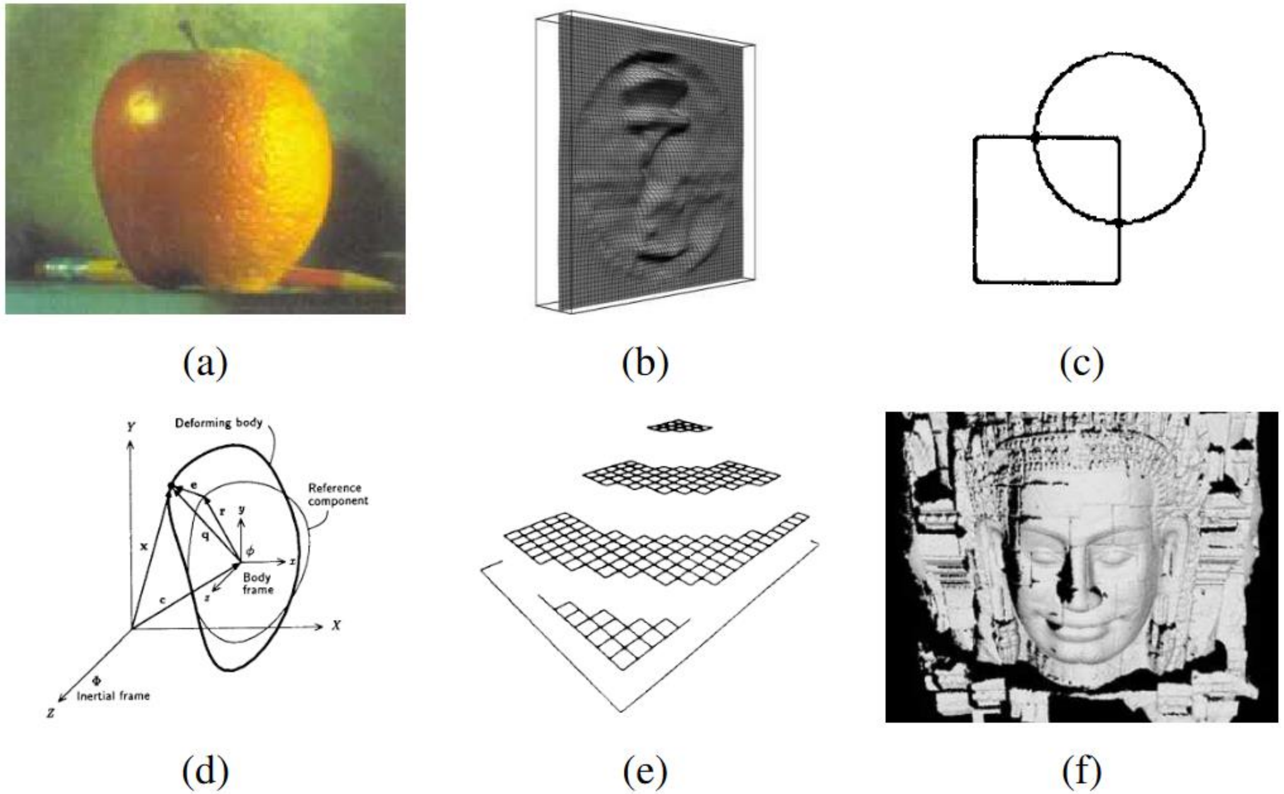
Intensity-based optical flow

Szeliski, R. (2022). Computer vision: algorithms and applications. Pages 9-18

1980s

- More sophisticated mathematical techniques for performing **quantitative image and scene analysis**

Image blending | **Shape from shading** | **Edge detection**



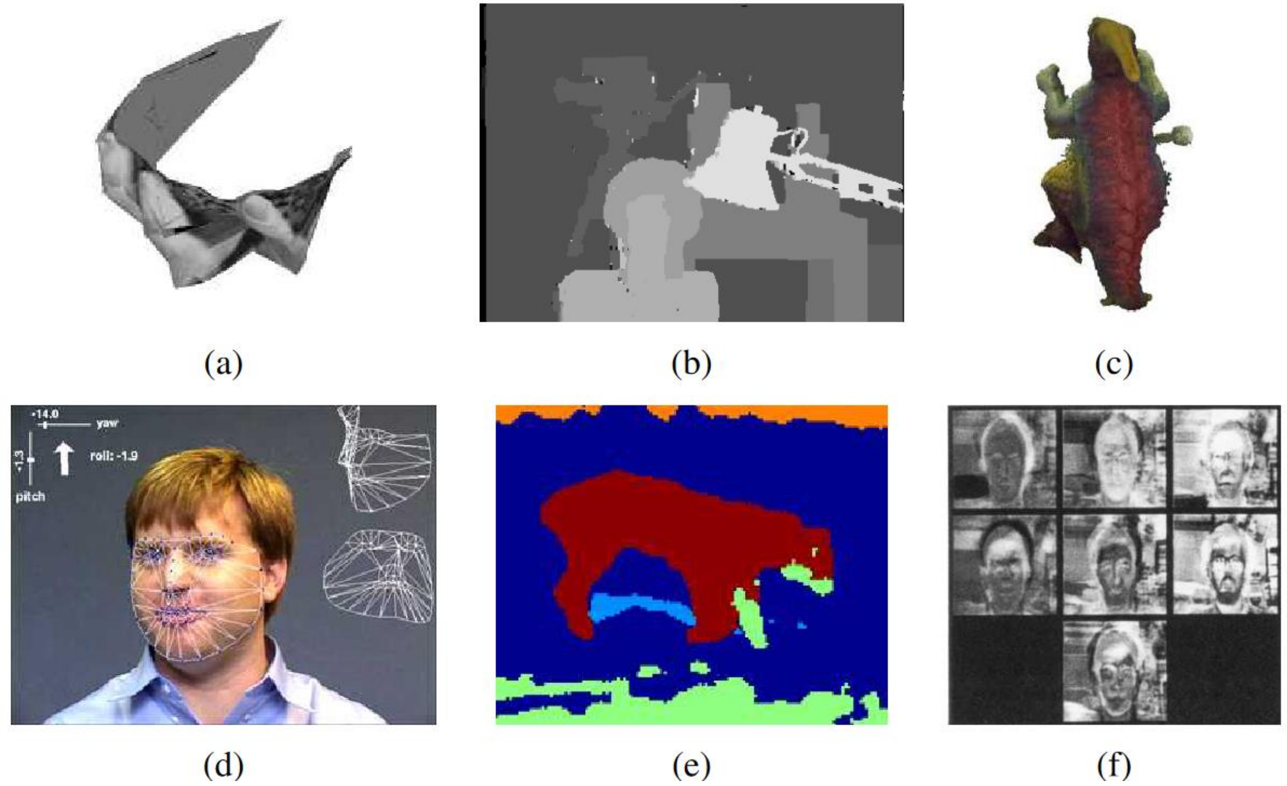
Deformable models | **Regularization-based surface reconstruction** | **Range data acquisition and merging**

Szeliski, R. (2022). Computer vision: algorithms and applications. Pages 9-18

1990s

- Much of the previously mentioned topics continued to be explored.
- A lot of the initial activity was directed at **projective reconstructions.**
- **Multi-view algorithms** that produce **complete 3D surfaces** were also an active topic of research.
- Increased **interaction with computer graphics**.

Structure-from-motion using orthographic camera approximations | **Dense stereo matching** | **Multi-view reconstruction**



Face tracking | **Image segmentation** | **Face recognition**

Szeliski, R. (2022). Computer vision: algorithms and applications. Pages 9-18

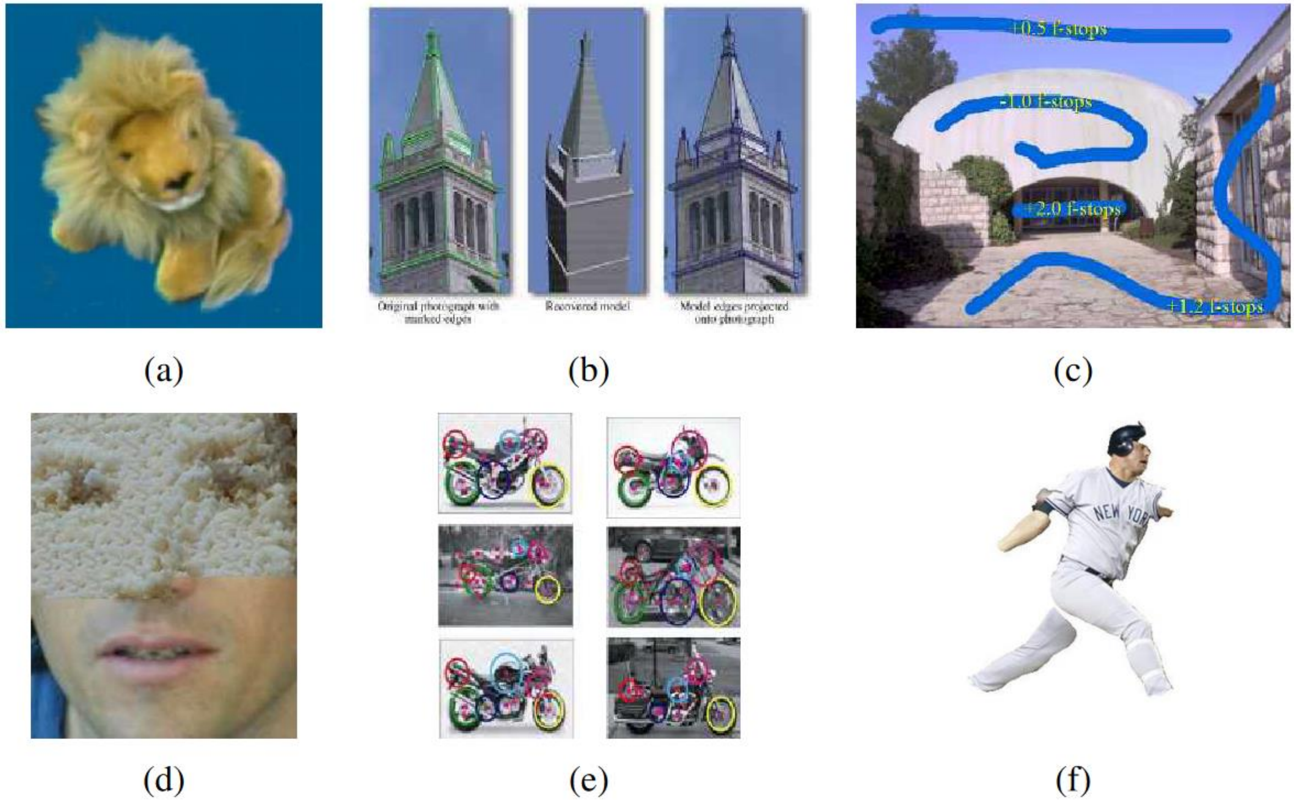
2000s

- This decade continued to deepen the **interplay between the vision and graphics fields**, but more importantly embraced **data-driven and learning approaches** as core components of vision.

Image-based rendering (new views of the scene)

Image-based modeling (generate a 3D model)

Interactive tone mapping (computational photography)



Texture synthesis

Feature-based techniques for object recognition

Region segmentation

Szeliski, R. (2022). Computer vision: algorithms and applications. Pages 9-18

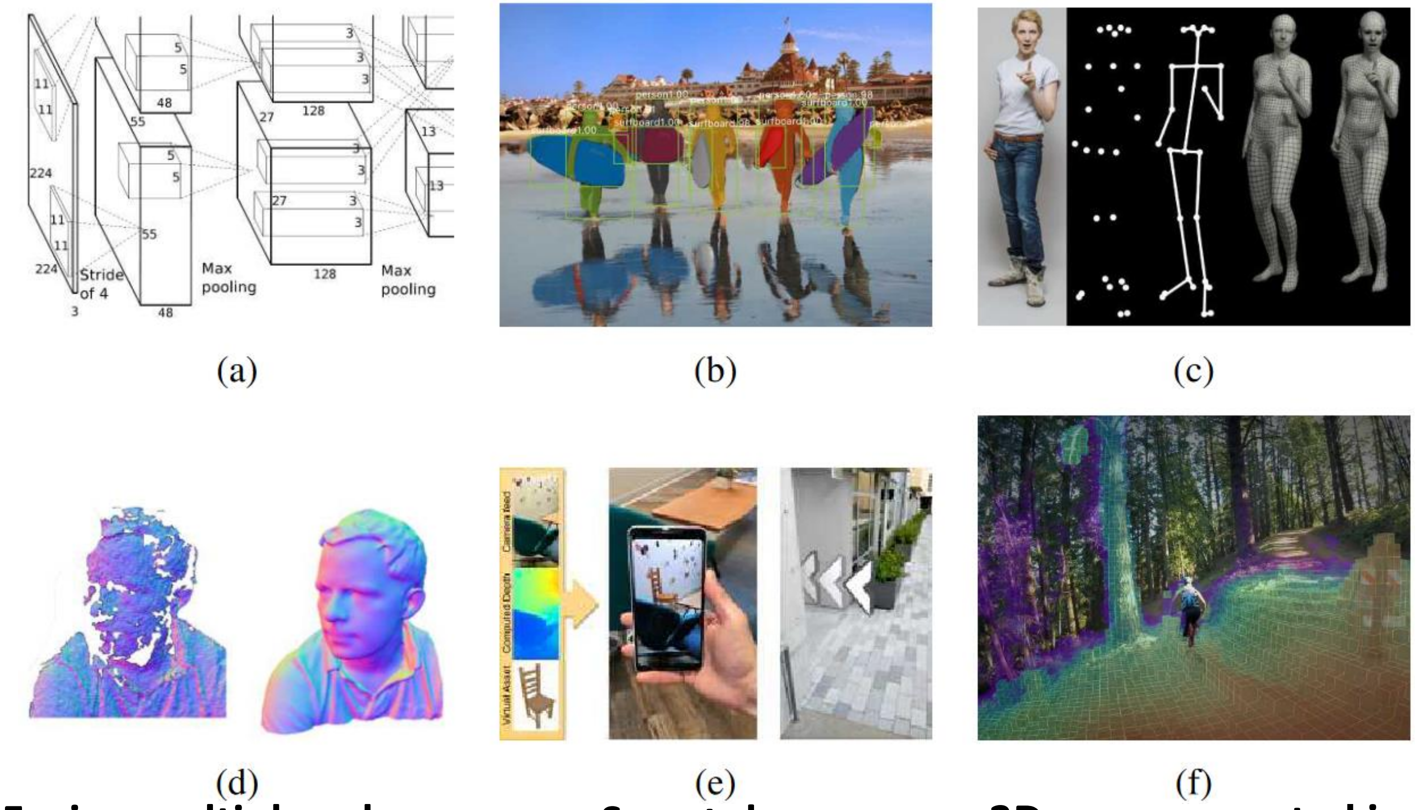
2010s

- Trend towards using **large labeled (and also self-supervised) datasets** to develop **machine learning** algorithms.
- **Specialized sensors and hardware** for computer vision tasks also continued to advance
- Incredible advances in the **performance and reliability** of computer vision algorithms.

Deep convolutional neural networks for visual recognition

Object instance segmentation

Person's 3D model with gestures and expression from a single image

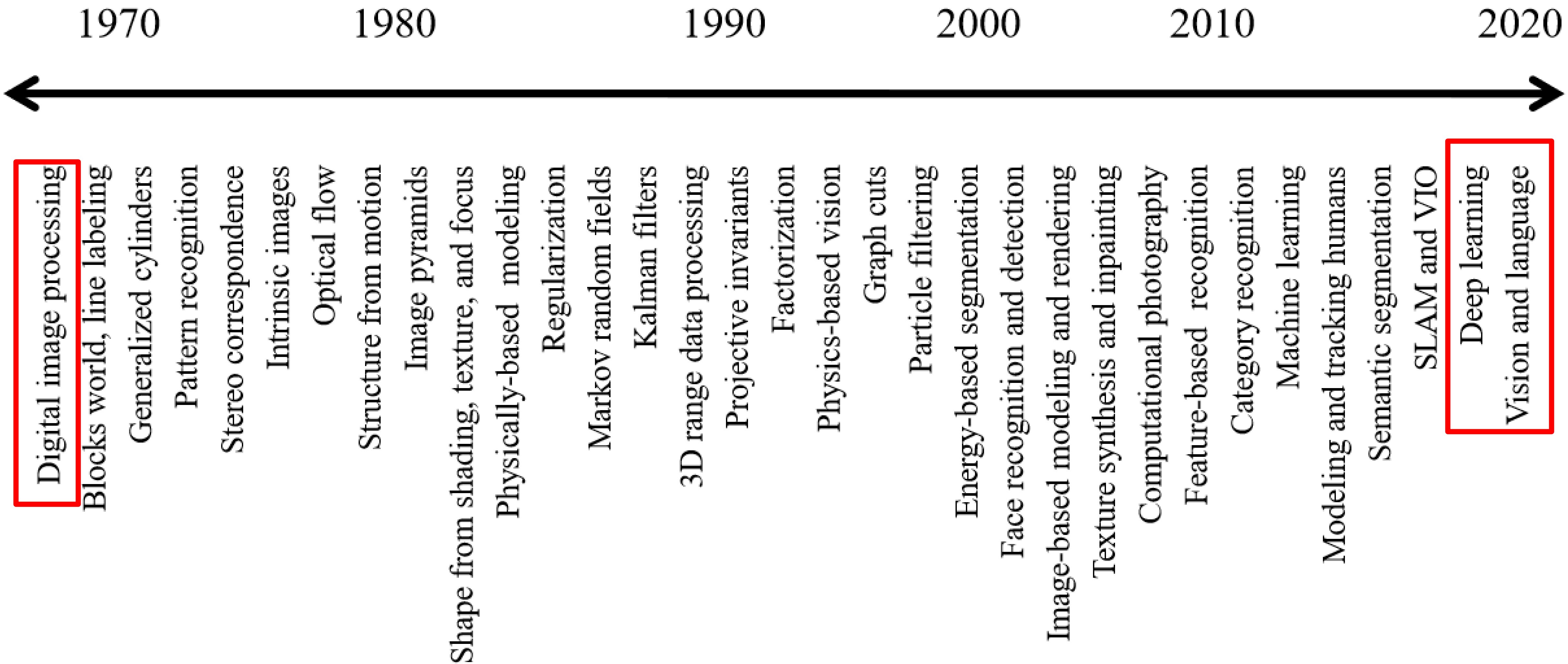


Fusing multiple color depth images in real-time

Smartphone augmented reality

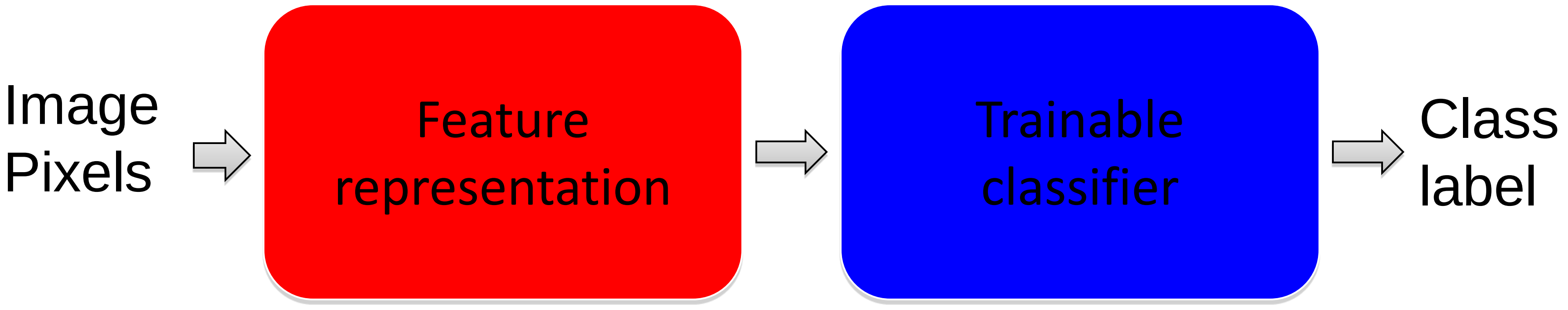
3D map computed in real-time on a fully autonomous drone

Szeliski, R. (2022). Computer vision: algorithms and applications. Pages 9-18



A rough timeline of some of the most active topics of research in computer vision.
(Szeliski, 2022)

“Classic” recognition pipeline

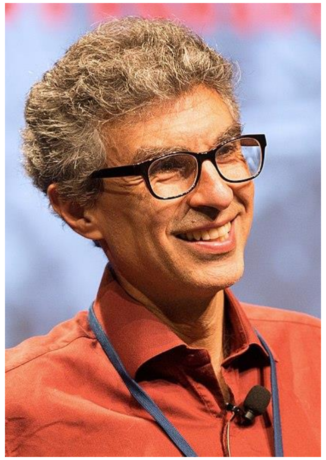


Hand-crafted feature representation

+

Off-the-shelf trainable classifier

Deep Learning



Yoshua Bengio
University of Montreal

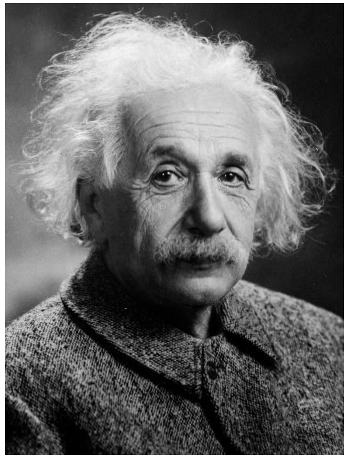
Geoffrey Hinton
University of Toronto & Google

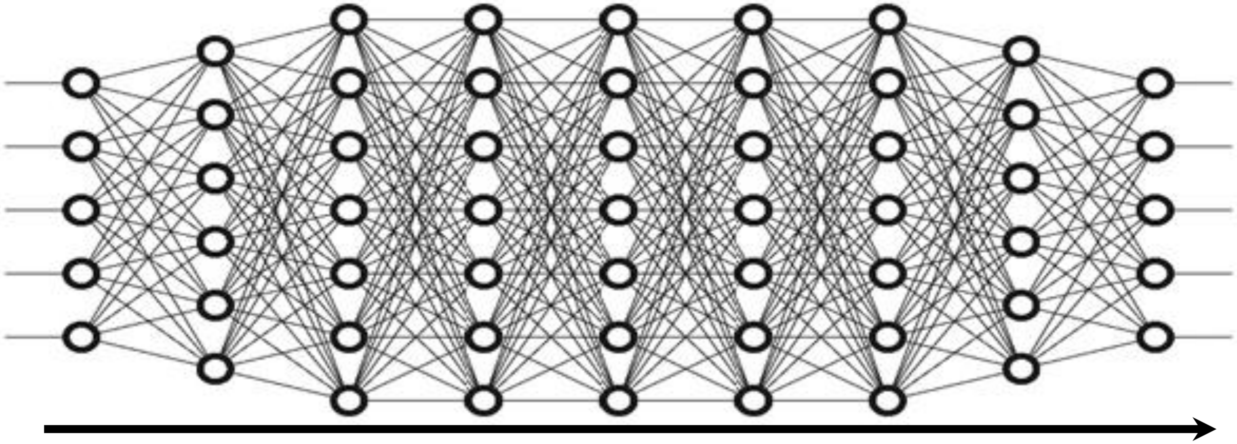
Yann LeCun
New York University & Facebook

Turing Award 2018

https://awards.acm.org/about/2018-turing

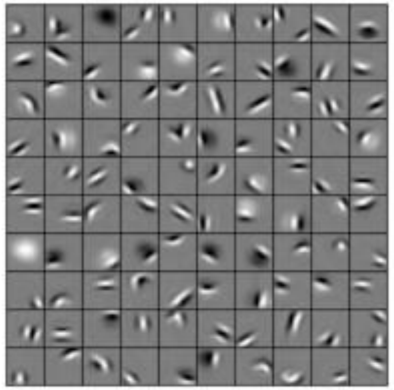
Hierarchical Learning



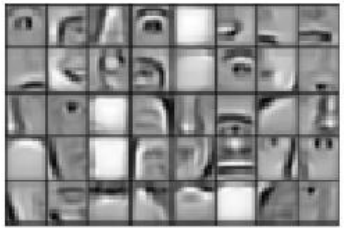


“Albert Einstein”

Low-level Features



Mid-level Features

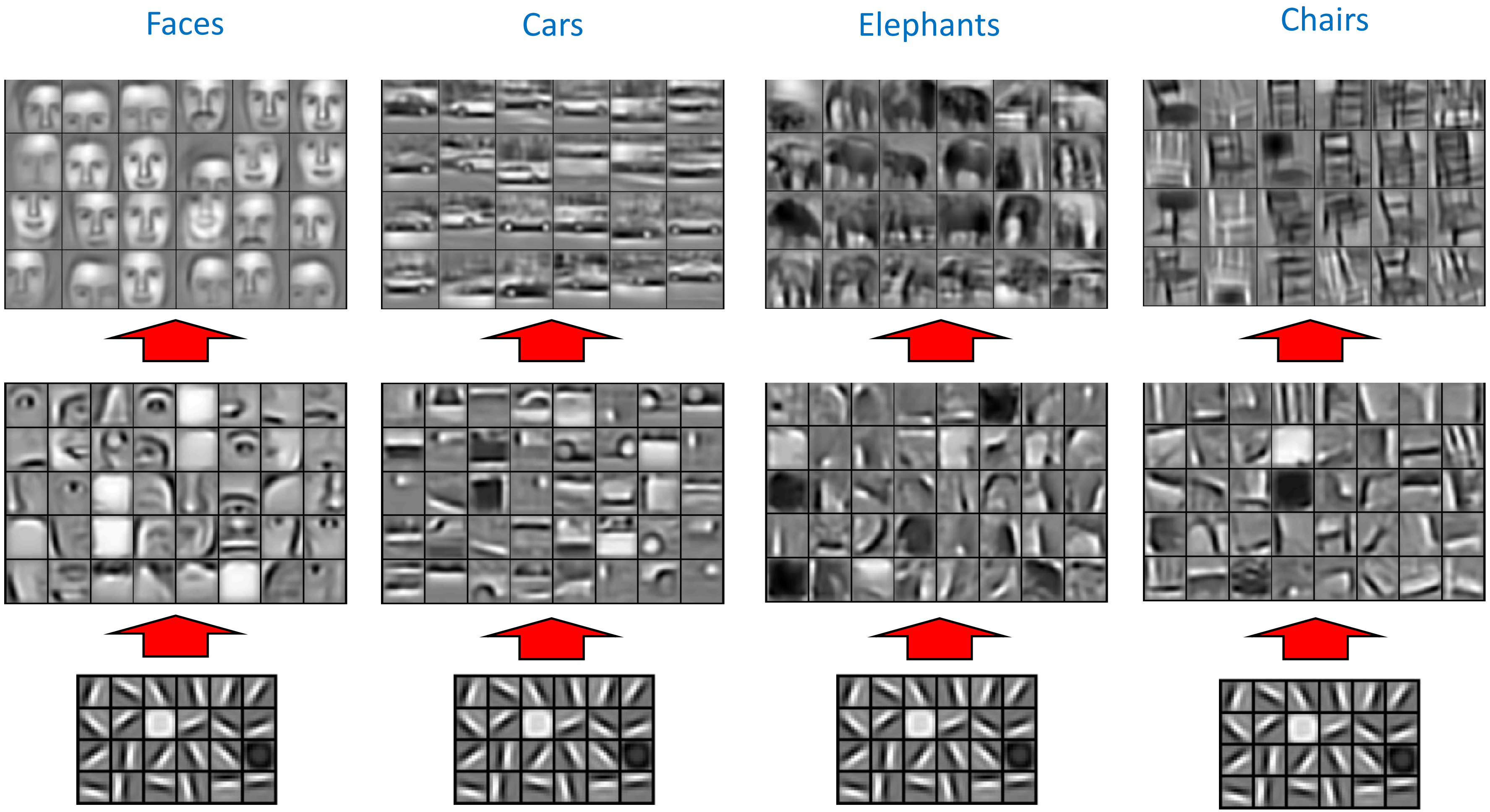


High-level Features

Deep learning allows computational models that are composed of **multiple processing layers** to **learn representations of data** with **multiple levels of abstraction**.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *Nature* 521.7553 (2015): 436-444.

Hierarchical Learning



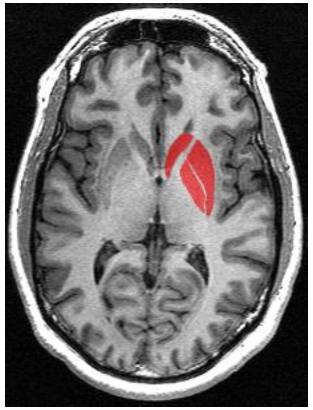
Summary

- 1970 – 2010
  - Create good linear mathematical model
  - Most effort focused on hand-crafted feature extraction

- Current:
  - Learn features from data
  - Many concepts share the same features
  - Internet is plenty of data to training models for many concepts
  - We have resources to train very large models (GPU computing)

Main limitations

- Machines can outperform humans at **concrete, well-defined tasks**, but they struggle at more general and abstract tasks.



Main limitations (2)

- Many times, your system is hard to analyze (**black box**) and it's only going to be **as good as the data you have**.
  - Serious ethical, practical, legal problems...

This is not a tank (Yudkowsky, 2008)

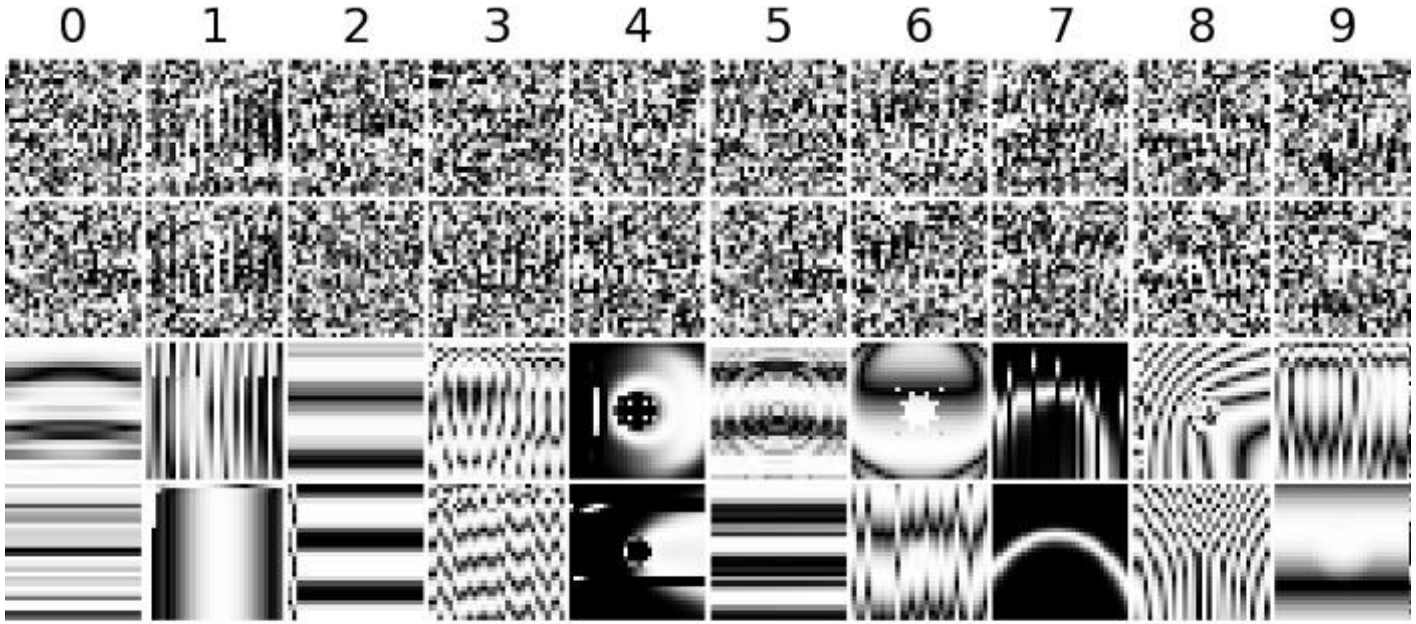
Main limitations (3)

- **Deep learning generally requires too much data**
  - Humans don't need to see 10,000 images of different cats to reliably identify one.

Image taken from https://www.fourpaws.com/

Main limitations (and 4)

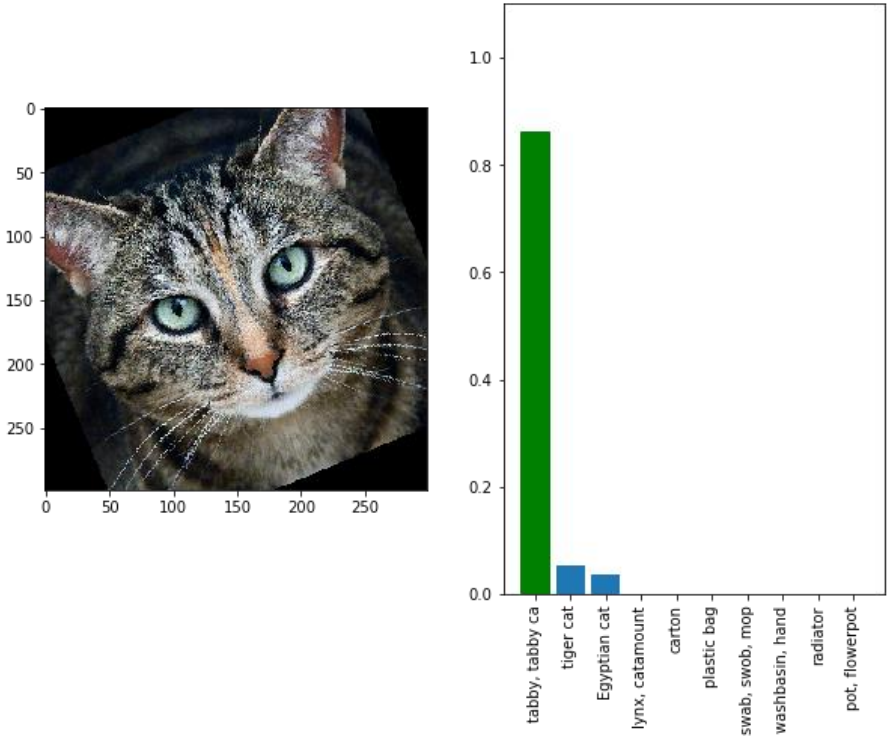
- **Machines lack context**
  - They only see signals, and apply the same solution every time
  - It is difficult for them to “see the whole”

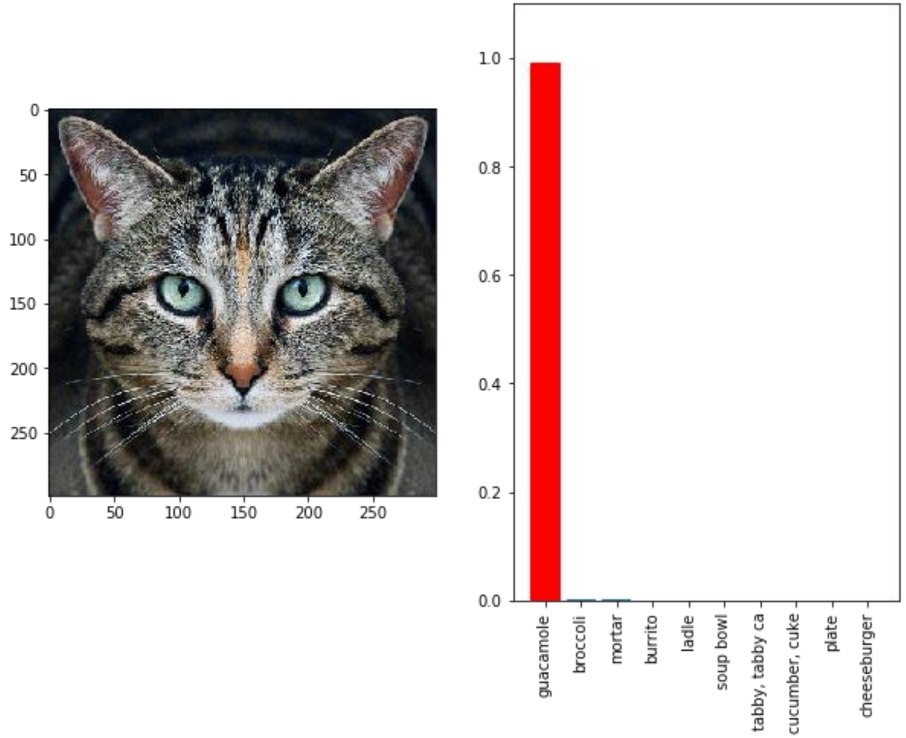


Deep neural networks are easily fooled: High confidence predictions for unrecognizable images (CVPR 2015)

Main limitations (and 4)

- **Machines lack context**
  - They only see signals, and apply the same solution every time
  - It is difficult for them to "see the whole"





Adversarial Example extracted from https://www.labsix.org/physical-objects-that-fool-neural-nets/ (2017)

Main limitations (and 4)

- **Machines lack context**
  - They only see signals, and apply the same solution every time
  - It is difficult for them to "see the whole"

https://spectrum.ieee.org/slight-street-sign-modifications-can-fool-machine-learning-algorithms

Introduction

Pablo Mesejo

pmesejo@go.ugr.es

Universidad de Granada

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

DECSAI

DaSCI

Instituto Andaluz de Investigación en **Data Science** and **Computational Intelligence**