Computer Vision

Teaching cameras to “see”

Arjun Jain
CS 763, Spring 2018
IIT Bombay, CSE Department

What is **Computer Vision**?

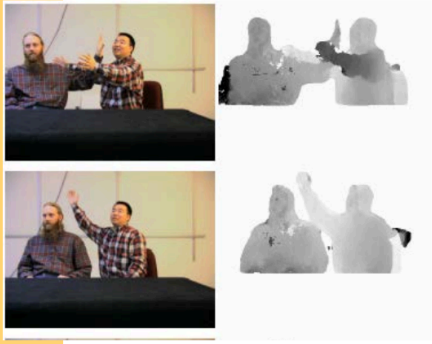
- Automatic understanding of images and video
  - **Measurement** : Computing properties of the 3D world from visual data
  - **Perception and interpretation**: Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities.

Computer Vision: Measurement

- Measurement

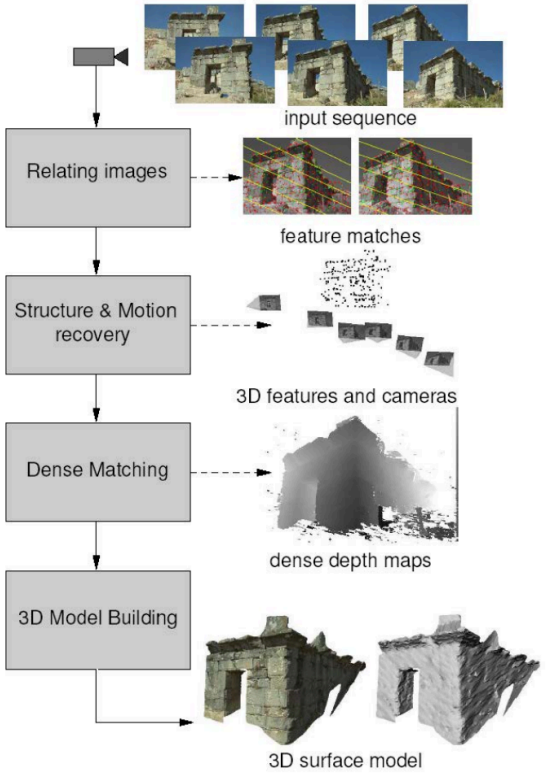
Real-time stereo

NASA Mars Rover

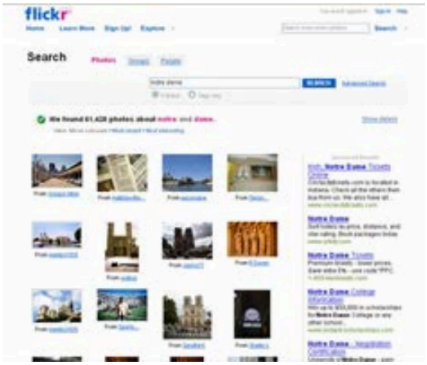


Pollefeys et al.

Structure-from-motion



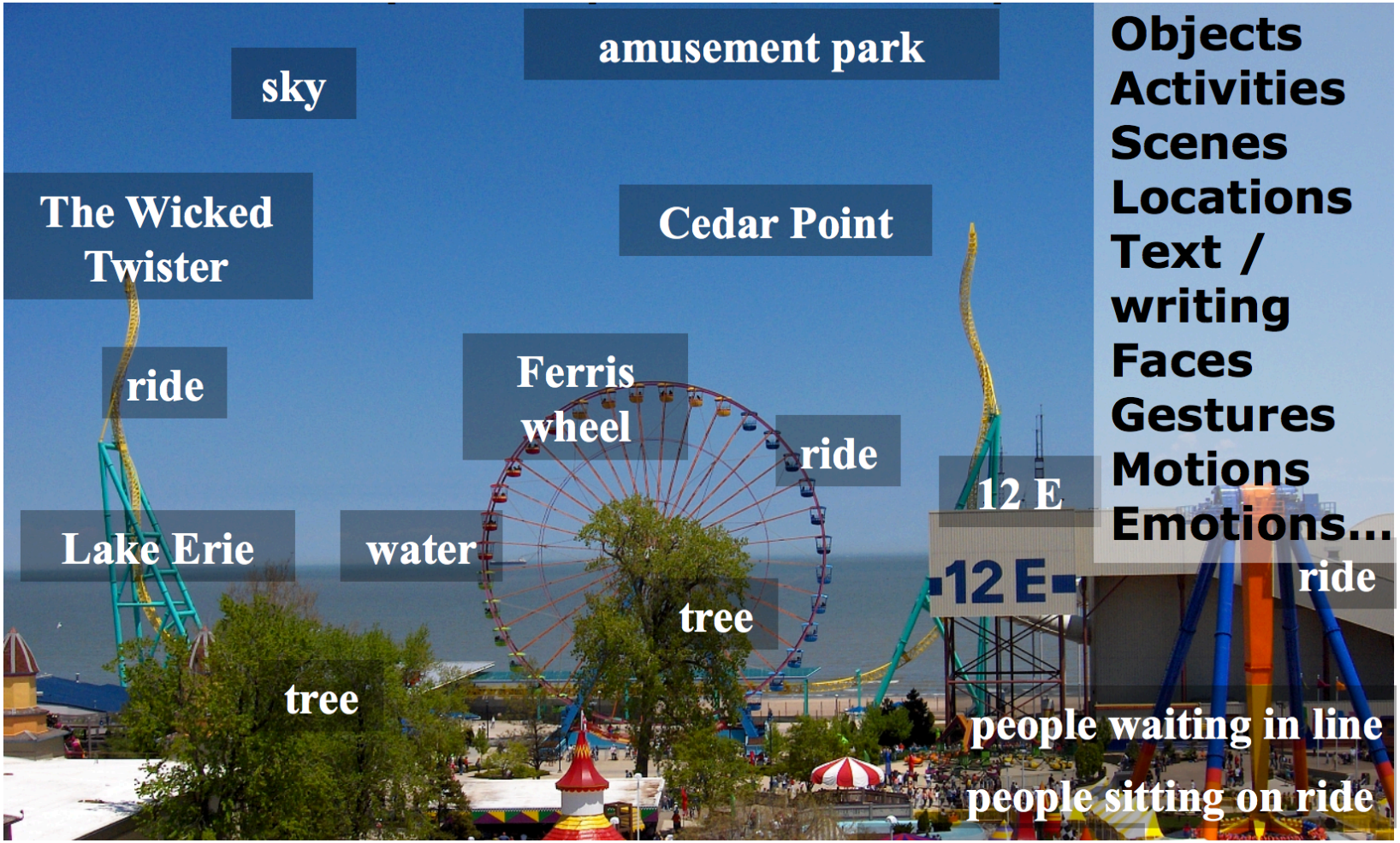
Multi-view stereo for community photo collections



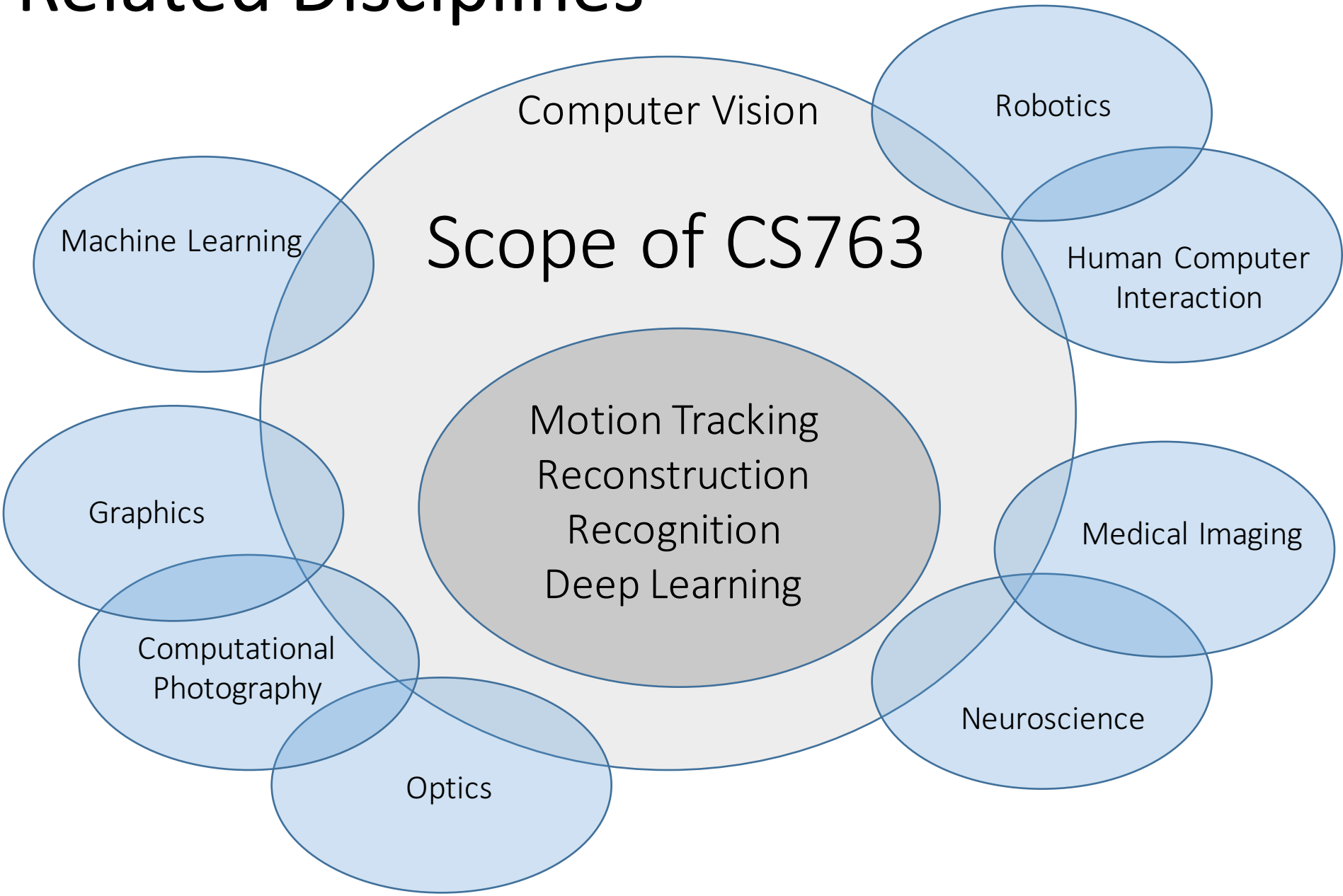
Goesele et al.

Computer Vision: Perception

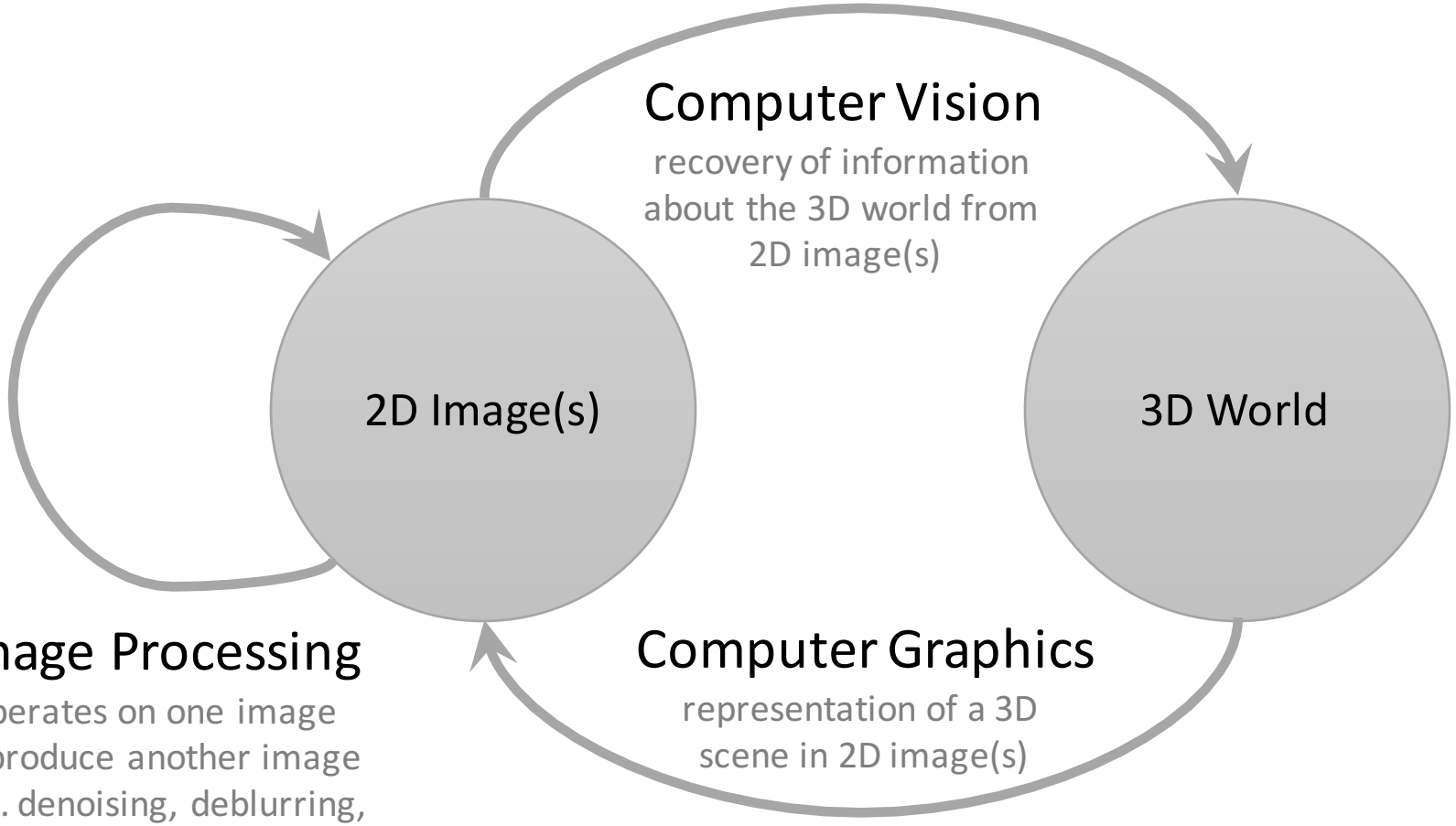
- Perception and interpretation



Related Disciplines



Computer Vision, Computer Graphics and Image Processing



Why do Computer Vision?

- As image sources multiply, so do applications
  - Relieve humans of boring, easy tasks
  - Enhance human abilities: human computer interaction, visualization
  - Perception for robotics / autonomous agents
  - Organize and give access to visual content

Why Computer Vision?

Images are everywhere!

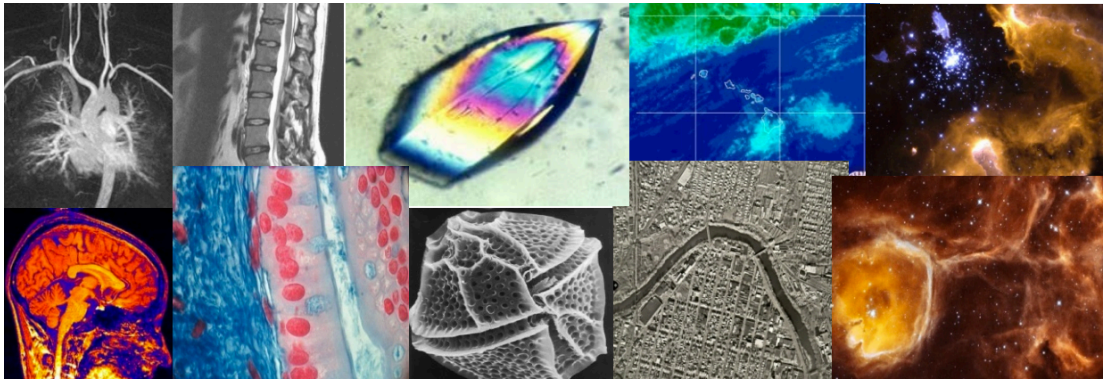
Personal Photo Albums

Movies, News, Sports

Google Image Search Picasa flickr webshots picsearch YouTube Broadcast Yourself

Surveillance and Security

Medical and Scientific Images



Why is Computer Vision so Hard?

Consider Image Classification: a core task in Computer Vision

(assume given set of discrete labels)

{dog, cat, truck, plane, ...}

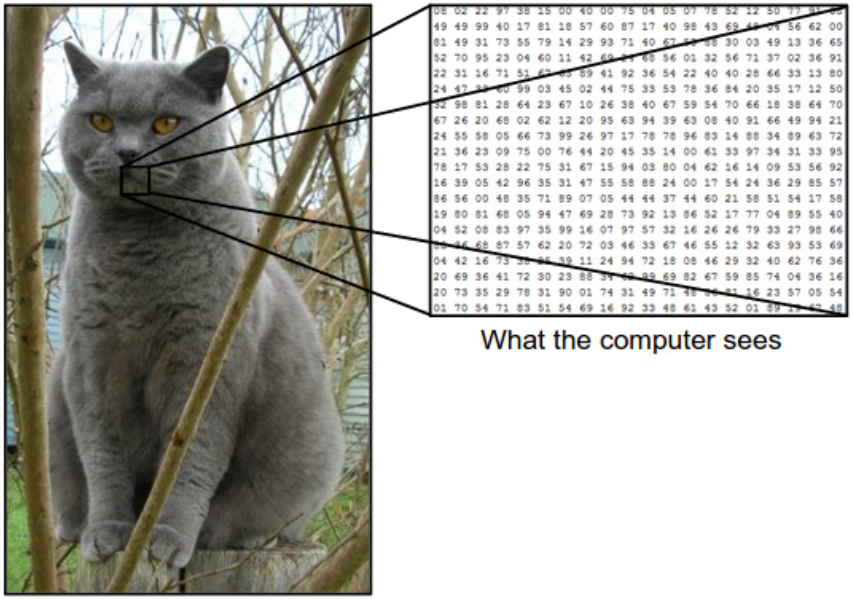
cat

Why is Computer Vision so Hard?

Images are represented as 3D arrays of numbers, with integers between [0, 255].

E.g.
300 x 100 x 3

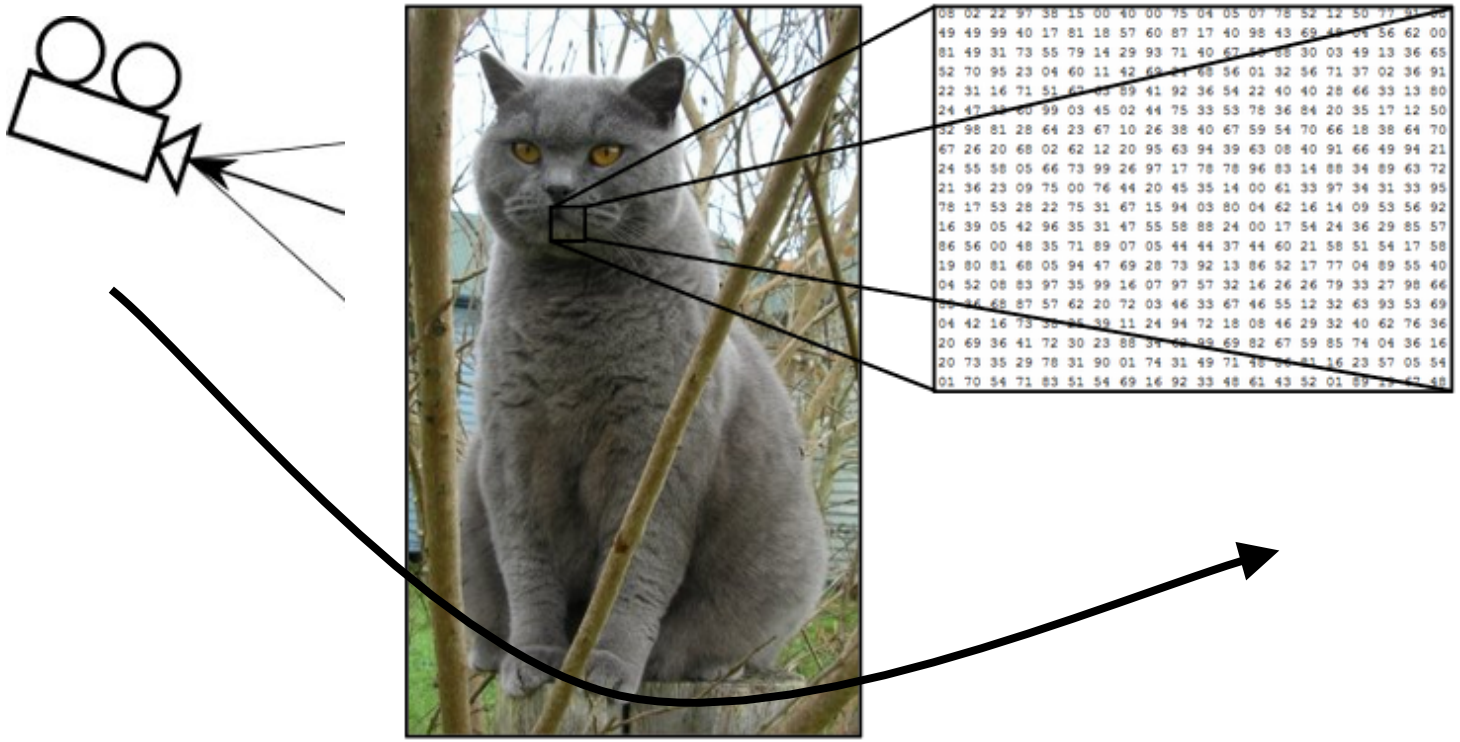
(3 for 3 color channels RGB)



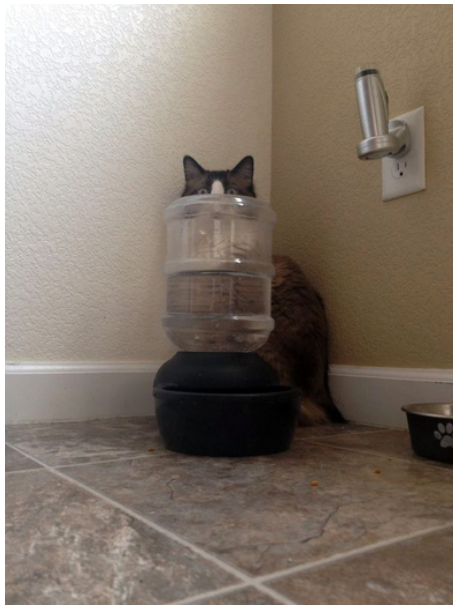
What the computer sees

Challenges: Invariant to Illumination

Challenges: Invariant to Viewpoint



Challenges: Deal with Occlusion

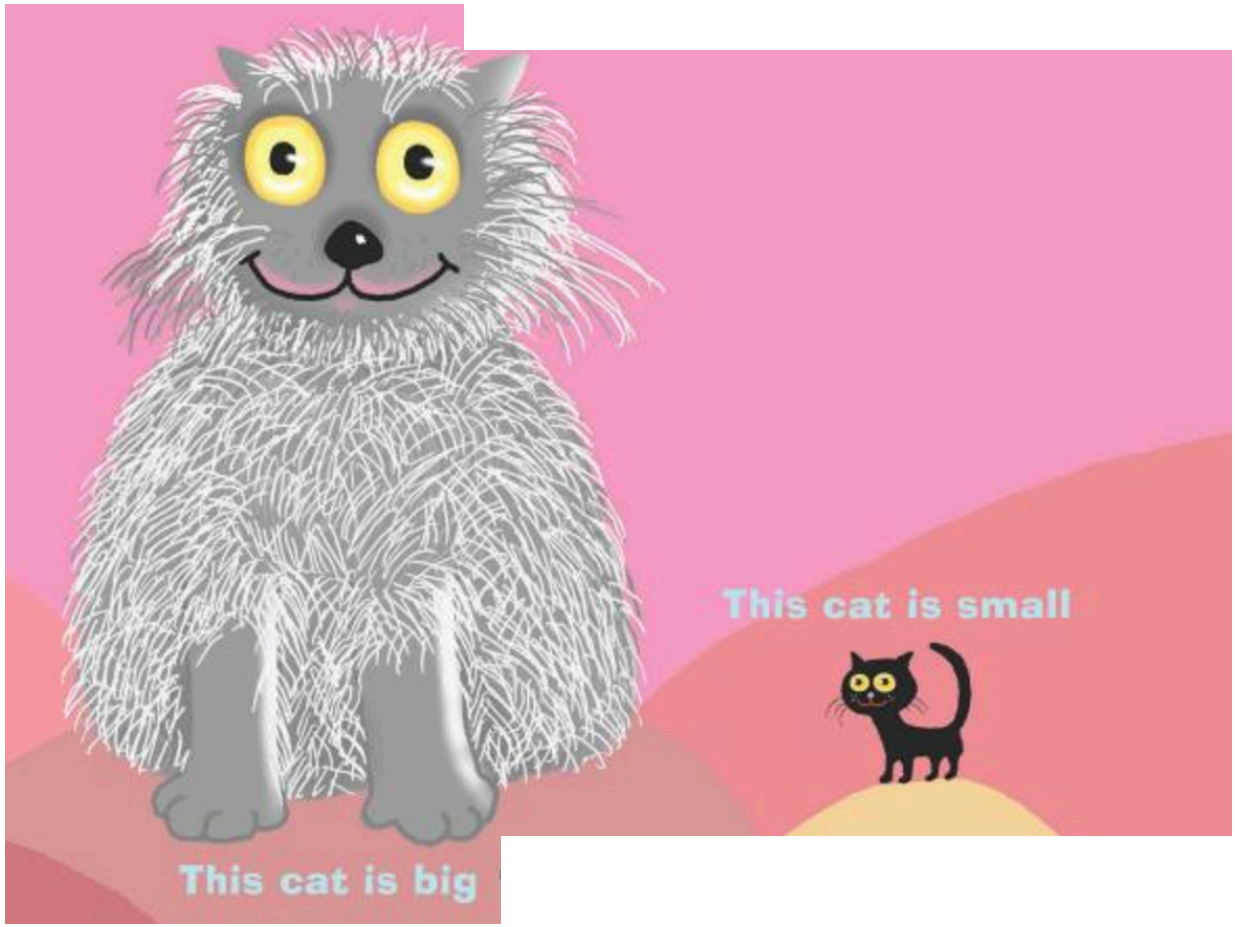


Challenges: Invariant to Deformation

Challenges: Deal with Background Clutter

Challenges: Deal with Intra-class Variation

Challenges: Deal with Scale Changes



Challenges: Deal with Motion

Challenges or Opportunities?

- Images are confusing, but they also reveal the structure of the world through numerous cues
- Computer Vision: interpret the cues (the human visual system does this all the time!)
- E.g. we interpret depth in images using both **physiological** and **psychological** cues
  - **Physiological** cues require both eyes to be open (binocular)
  - Other cues are available also when looking at images with only one open eye (monocular). All **psychological** cues are monocular

Depth Cues: Linear Perspective

When looking down a straight level road we see the parallel sides of the road meet in the horizon

Depth Cues: Aerial Perspective

The mountains in the horizon look always slightly bluish or hazy. The reason for this are small water and dust particles in the air between the eye and the mountains. The farther the mountains, the hazier they look.

Depth Ordering Cues: Occlusion

When objects block each other out of our sight, we know that the object that blocks the other one is closer to us. The object whose outline pattern looks more continuous is felt to lie closer.

Depth Cues: Texture Gradient

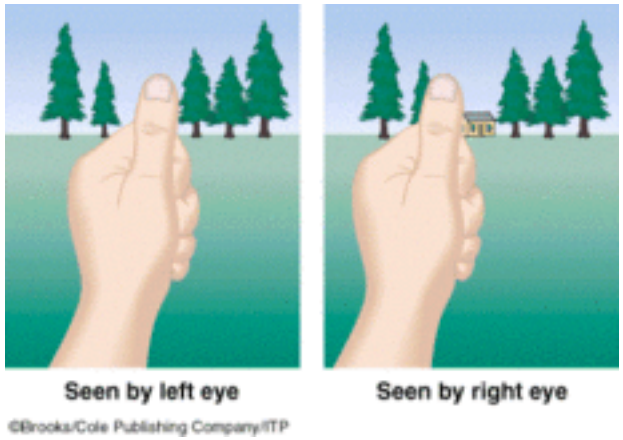
The closer we are to an object the more detail we can see of its surface texture. So objects with smooth textures are usually interpreted being farther away.

Depth Cues: Texture Gradient

When we know the location of a light source and see objects casting shadows on other objects, we learn that the object shadowing the other is closer to the light source.

Depth Cues

- **Binocular Parallax:** As our eyes see the world from slightly different locations, the images sensed by the eyes are slightly different. This difference in the sensed images is called binocular parallax. Human visual system is very sensitive to these differences, and binocular parallax is the most important depth cue for medium viewing distances. The sense of depth can be achieved using binocular parallax even if all other depth cues are removed.



The closer the object, the larger the disparity. Far away objects will seem almost the same by both eyes.

Depth Cues

- **Monocular Movement Parallax:** If we close one of our eyes, we can perceive depth by moving our head. This happens because human visual system can extract depth information in two similar images sensed after each other, in the same way it can combine two images from different eyes.

- **Retinal Image Size:** When the real size of the object is known, our brain compares the sensed size of the object to this real size, and thus acquires information about the distance of the object.

Grouping Cues
(color, texture, proximity)

Grouping Cues: “Common Fate”

Bottom Line

- Perception is an inherently ambiguous and ill posed problem:
  - Many different 3D scenes could have given rise to the same 2D picture

- Possible solutions: Bring in more constraints (more images)
- Use prior knowledge about the structure of the world
- Need a combination of different methods!

Every Picture Tells a Story

Goal of computer vision is to write computer programs that can interpret images

Computer Vision

- Has been around since the 1960s
- What has changed?

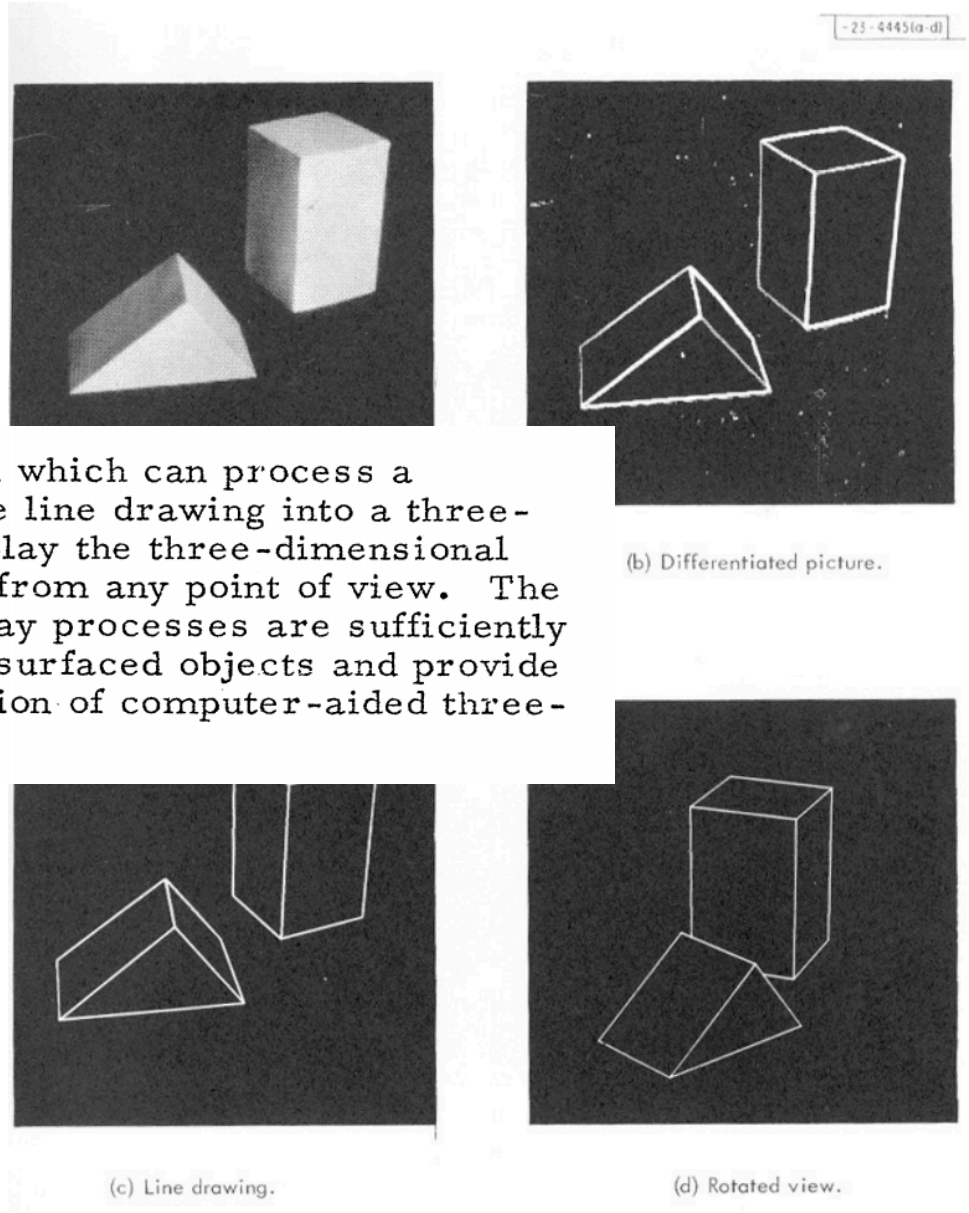
1. Increasing availability of cheap, powerful cameras (e.g. digital cameras, webcams) and other sensors

2. Increasing availability of massive amounts of labeled and unlabeled image and multimedia content on the web (e.g. face databases, etc.)

3. Increasing availability of cheap, powerful computers (processing speed and memory capacity - 10 Tflops by 1 Titan X!). Anyone heard of Titan V? Tesla V100?

4. Techniques from machine learning and statistics which lead to more complex, data-driven models and algorithms (e.g. deep learning!)

L. G. Roberts, Machine Perception of Three Dimensional Solids, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.



(b) Differentiated picture.

(c) Line drawing.

(d) Rotated view.

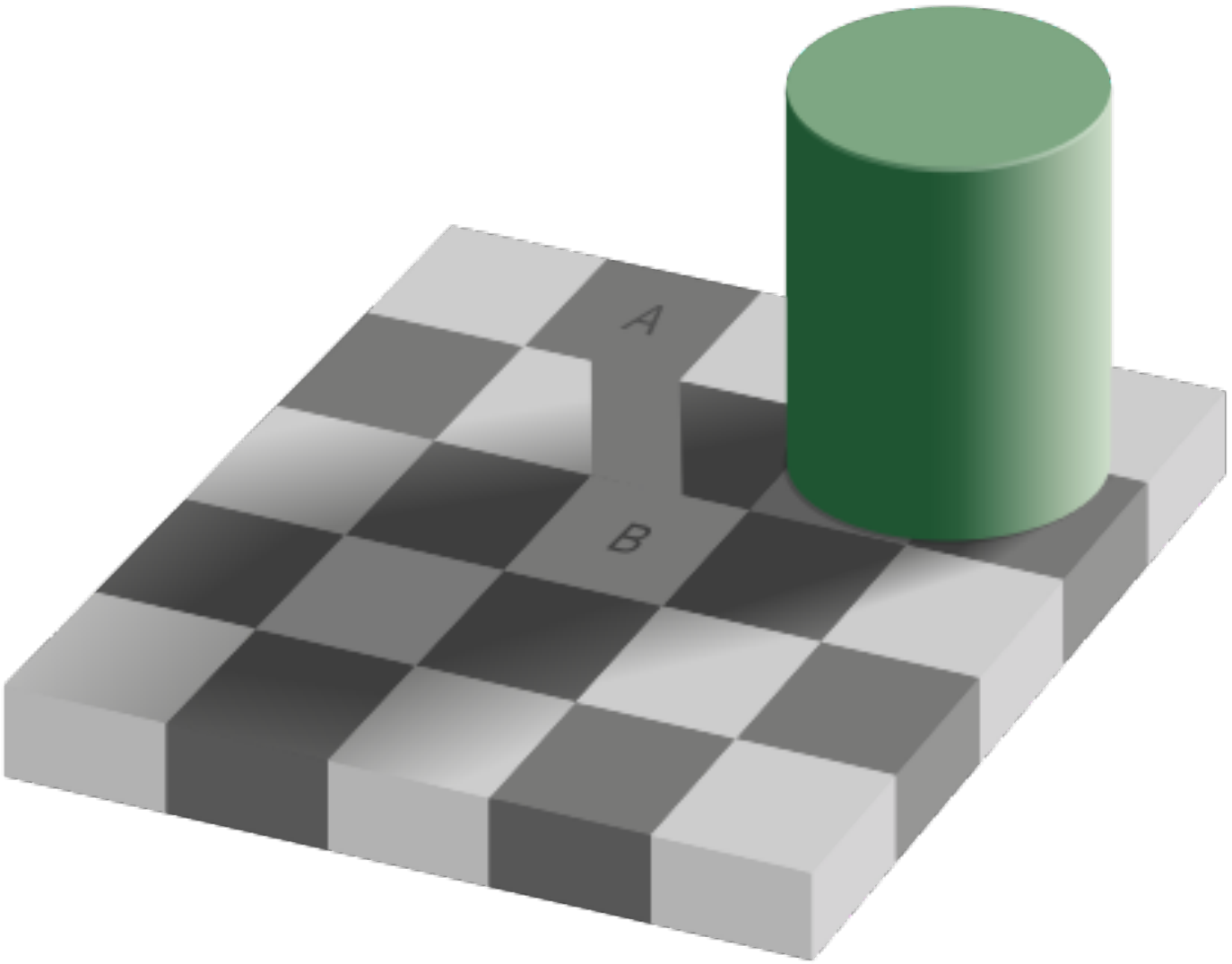
A computer program has been written which can process a photograph into a line drawing, transform the line drawing into a three-dimensional representation, and finally, display the three-dimensional structure with all the hidden lines removed, from any point of view. The 2-D to 3-D construction and 3-D to 2-D display processes are sufficiently general to handle most collections of planar-surfaced objects and provide a valuable starting point for future investigation of computer-aided three-dimensional systems.

Human Perception Has Issues

Sinha and Poggio, Nature, 1996

- Absolutely identical in terms of nose, eyes, mouth and their spatial arrangements!

Human Perception Has Issues



Can Computers Beat Human Vision?

- Yes and no
  - humans are usually much better at “generic” problems
  - computers can be better at “specific” problems

Can Computers Beat Human Vision?

http://www.bbc.com/news/technology-39298199

Vision for Touchless UI

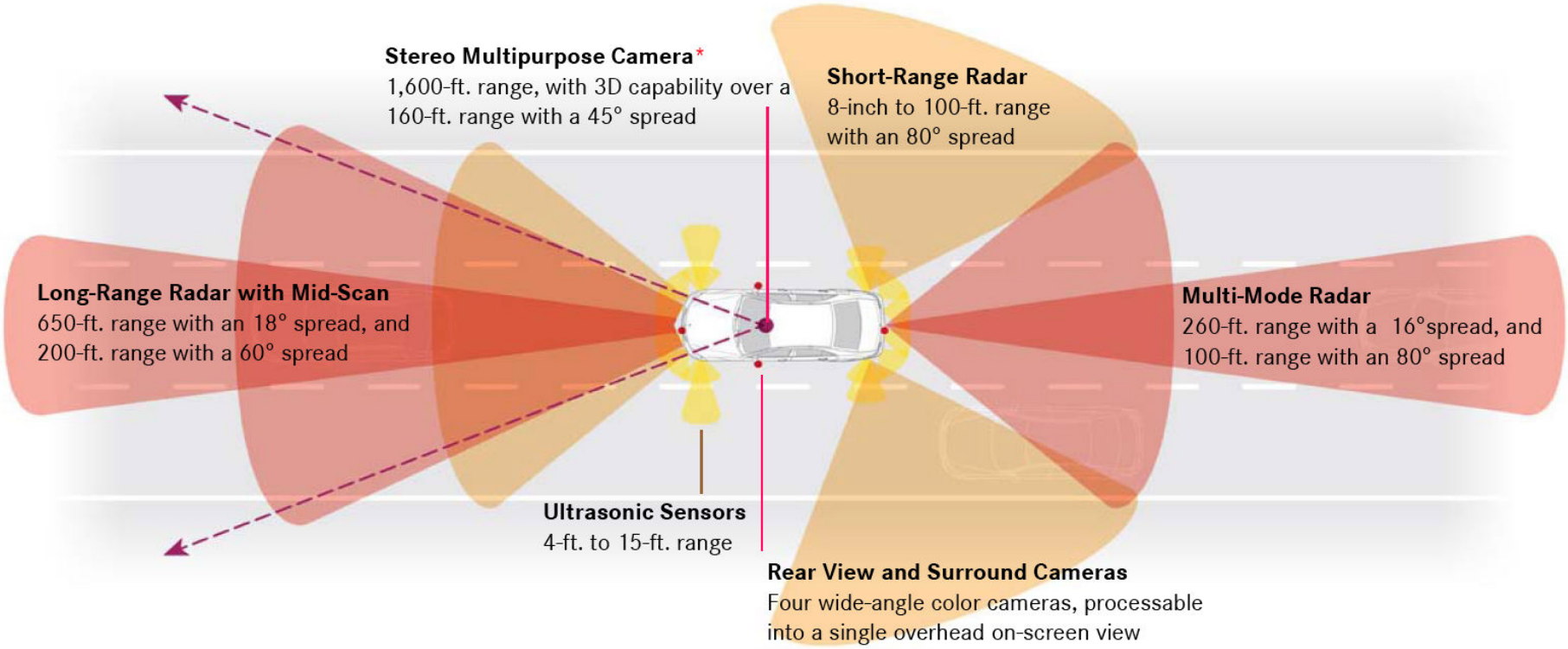
Vision for VFX

Motion Capture
Pirates of the Caribbean,
Industrial Light and Magic

Vision in Industry

- Computer Vision in finding defects in filling
- Counting, half-filled/overfilled bottles or incorrectly printed labels can be detected by a computer vision system located in the process line

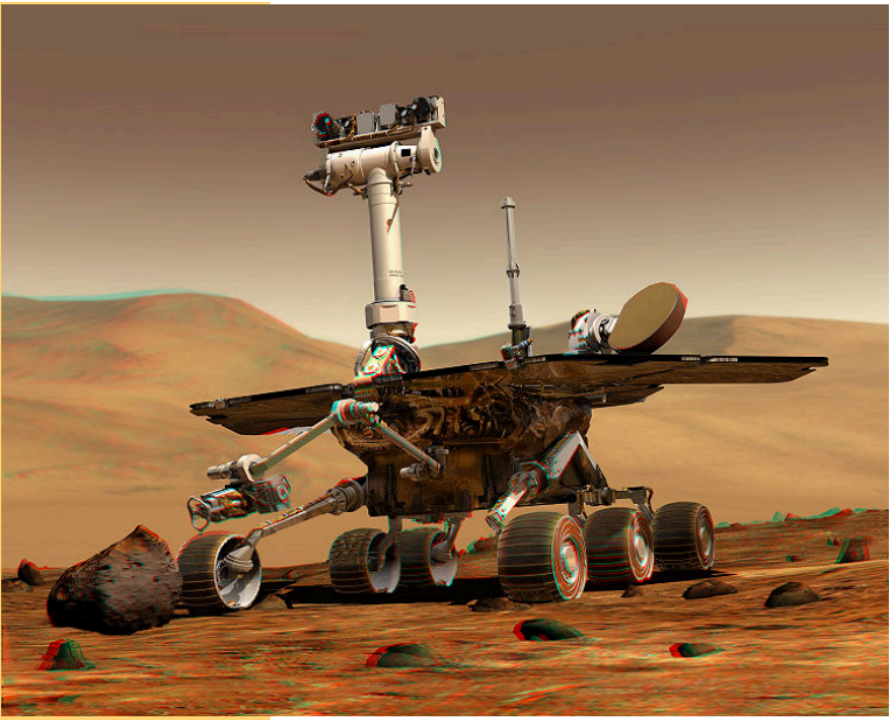
Vision in Autonomous Cars



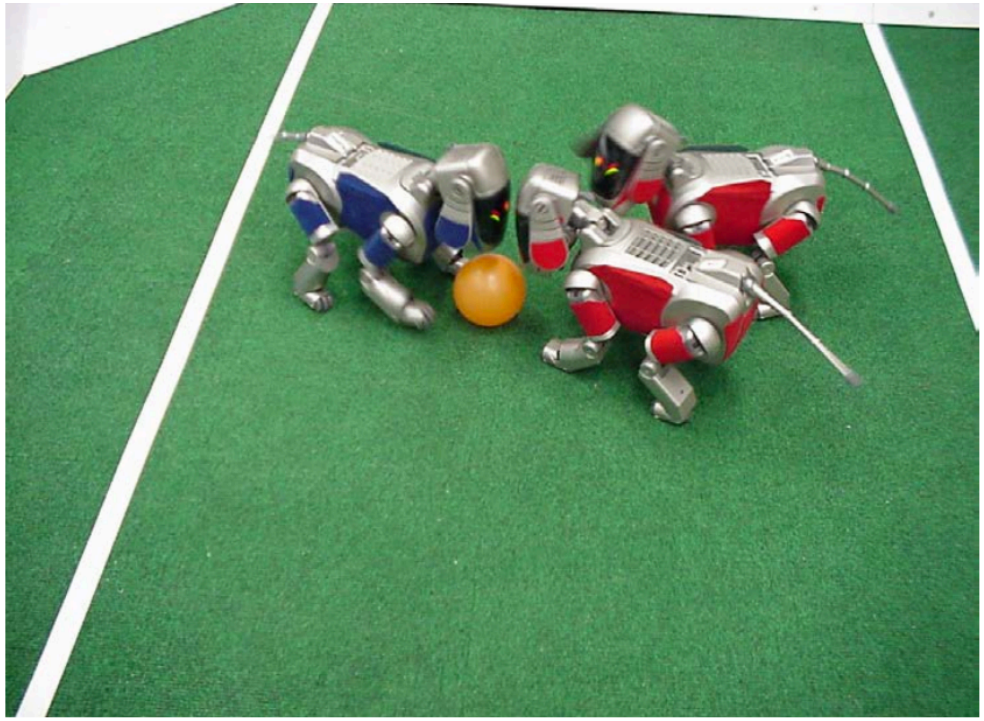
AutoCars - Uber bought CMU's lab

Vision Augments Reality (AR)

Vision in Robotics



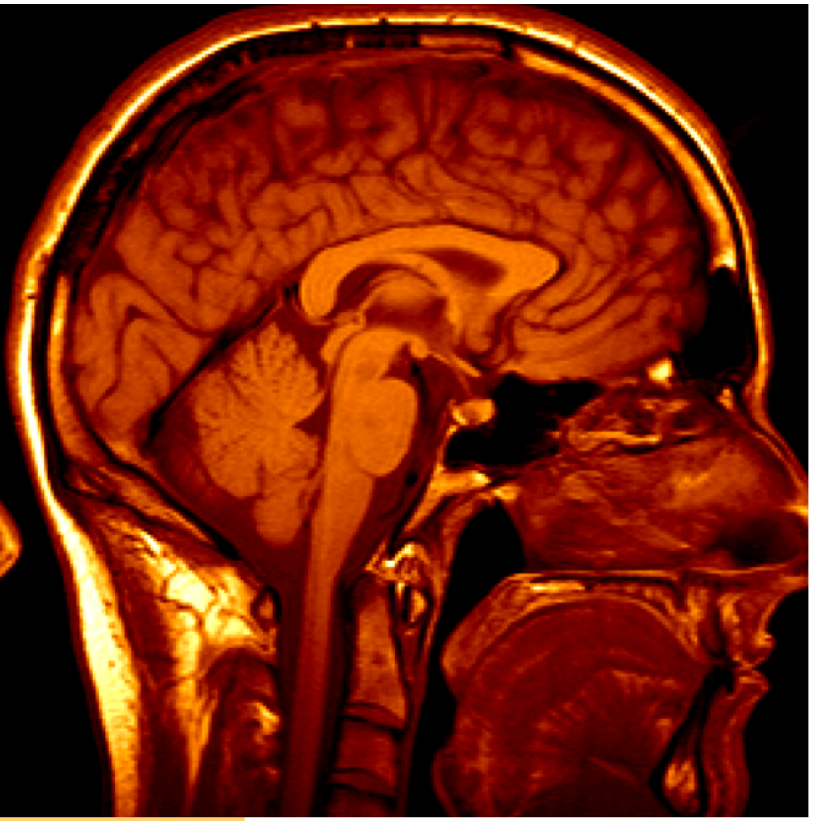
NASA's Mars Spirit Rover
http://en.wikipedia.org/wiki/Spirit_rover



http://www.robocup.org/

Vision in Agriculture

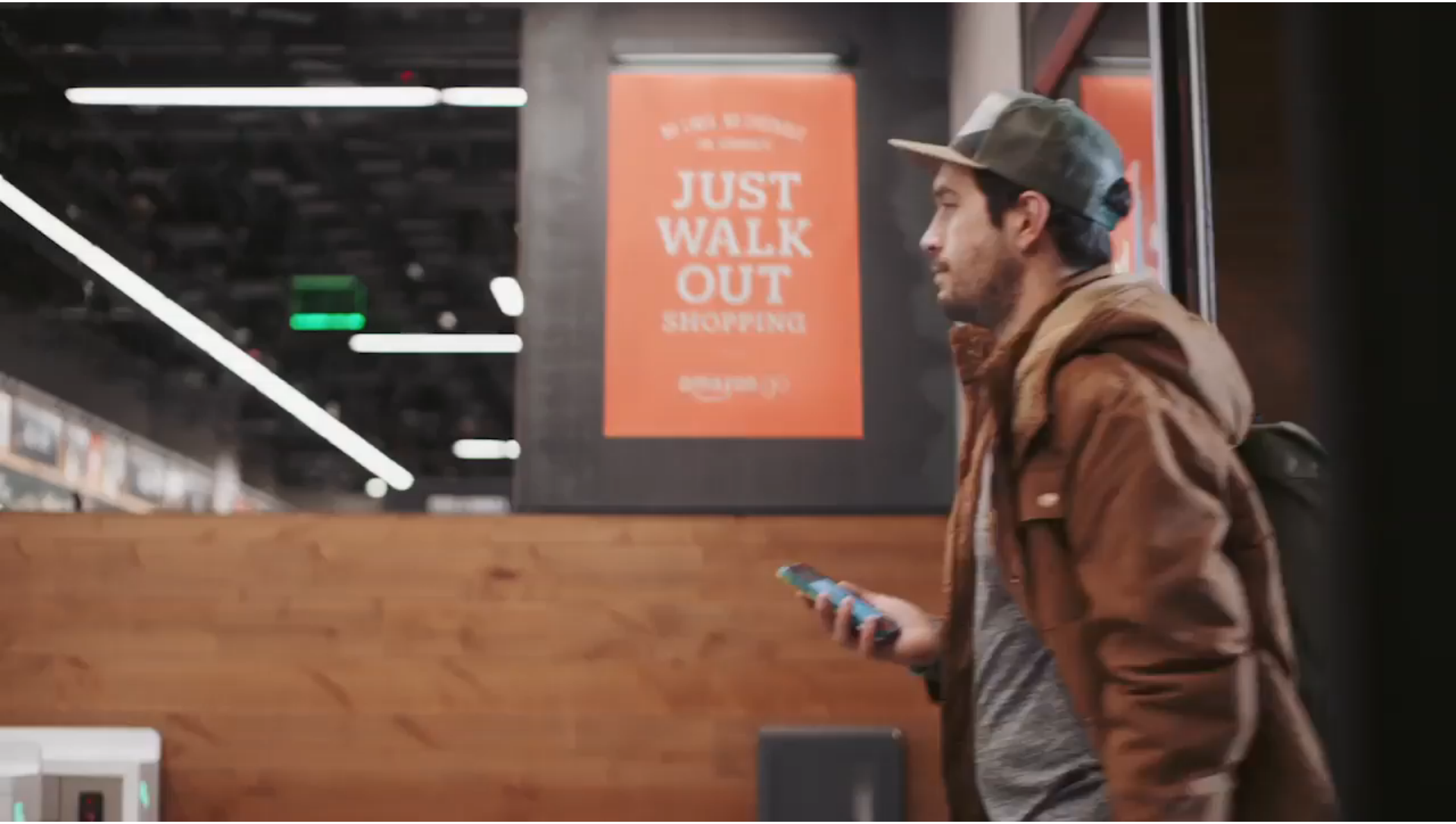
Vision in Medicine



3D imaging
MRI, CT

Image guided surgery
Grimson et al., MIT

Vision in Retail



Vision for Emotional Analytics

emotient

Emotient Analytics In Action

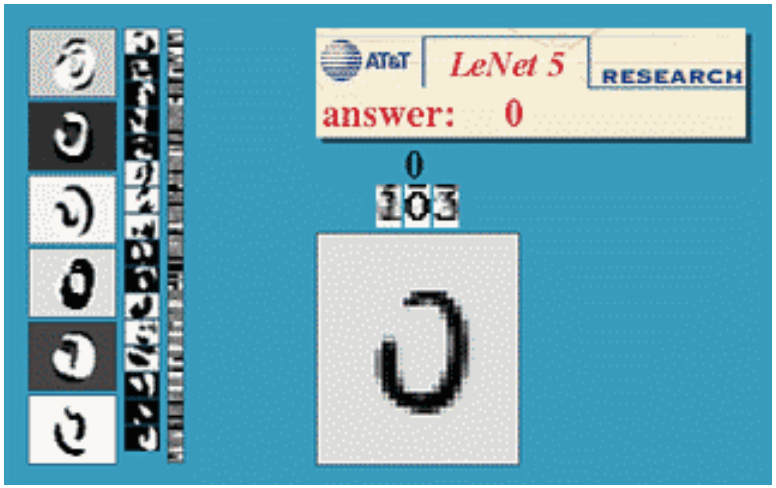
Visit analytics.emotient.com/signup for a free trial.

Vision in Security and Intelligence

Optical character recognition

Technology to convert scanned docs to text

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

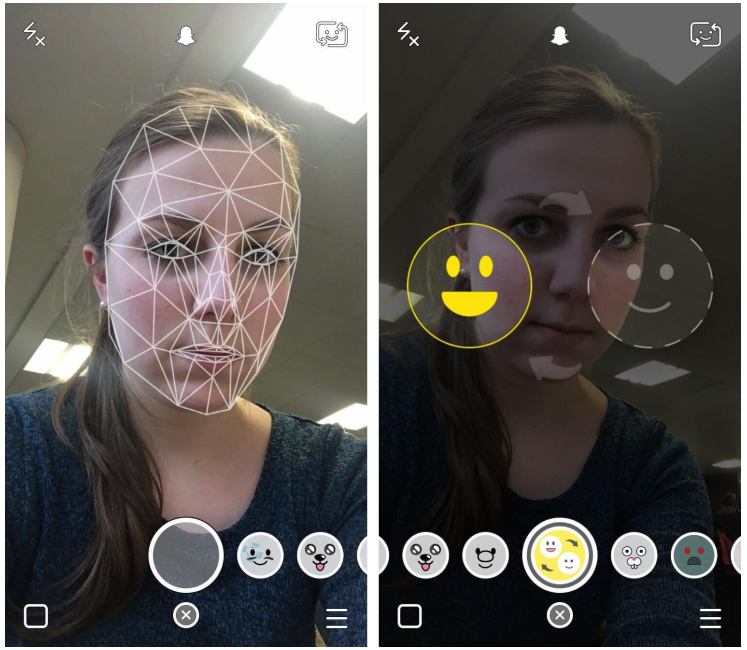


Digit recognition, AT&T labs
http://www.research.att.com/~yann/

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

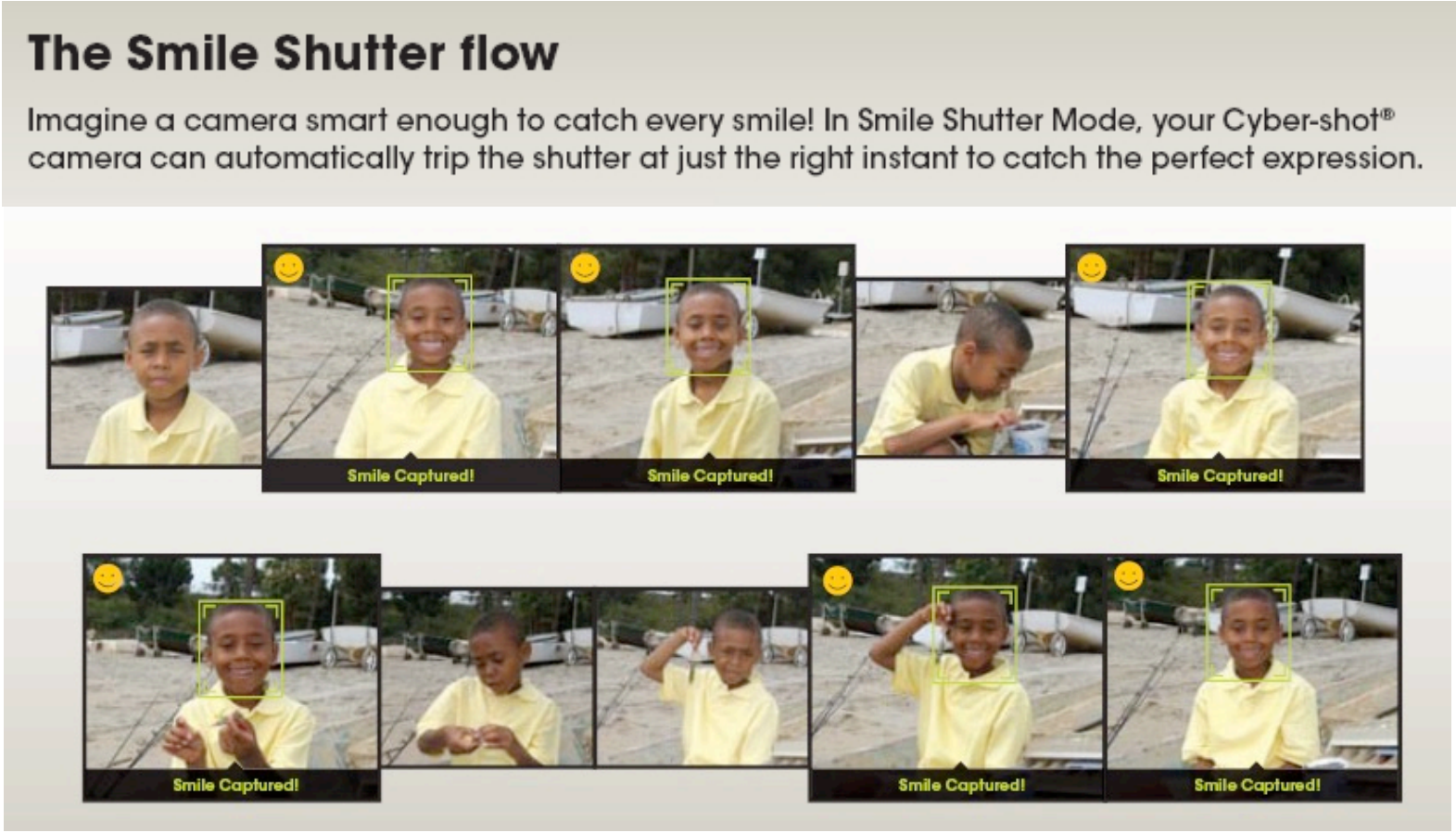
JH

Face detection



- Almost all digital cameras detect faces
- Snapchat face filters

Smile detection

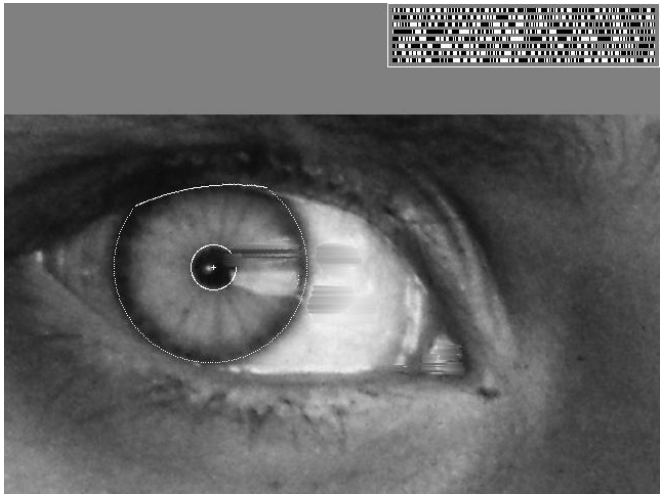


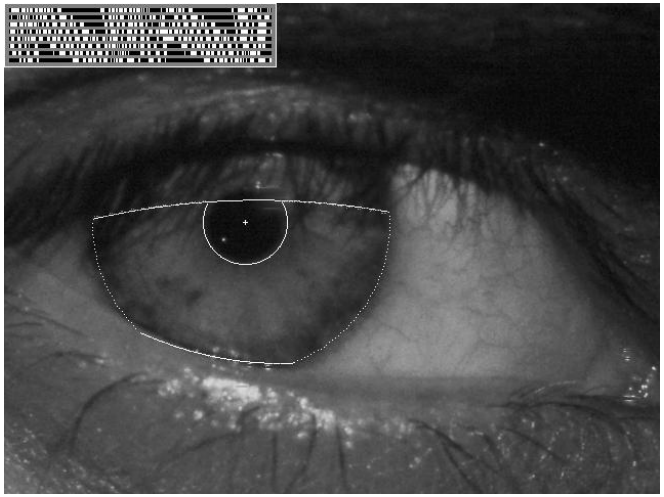
[Sony Cyber-shot® T70 Digital Still Camera]

JH

Vision-based biometrics

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

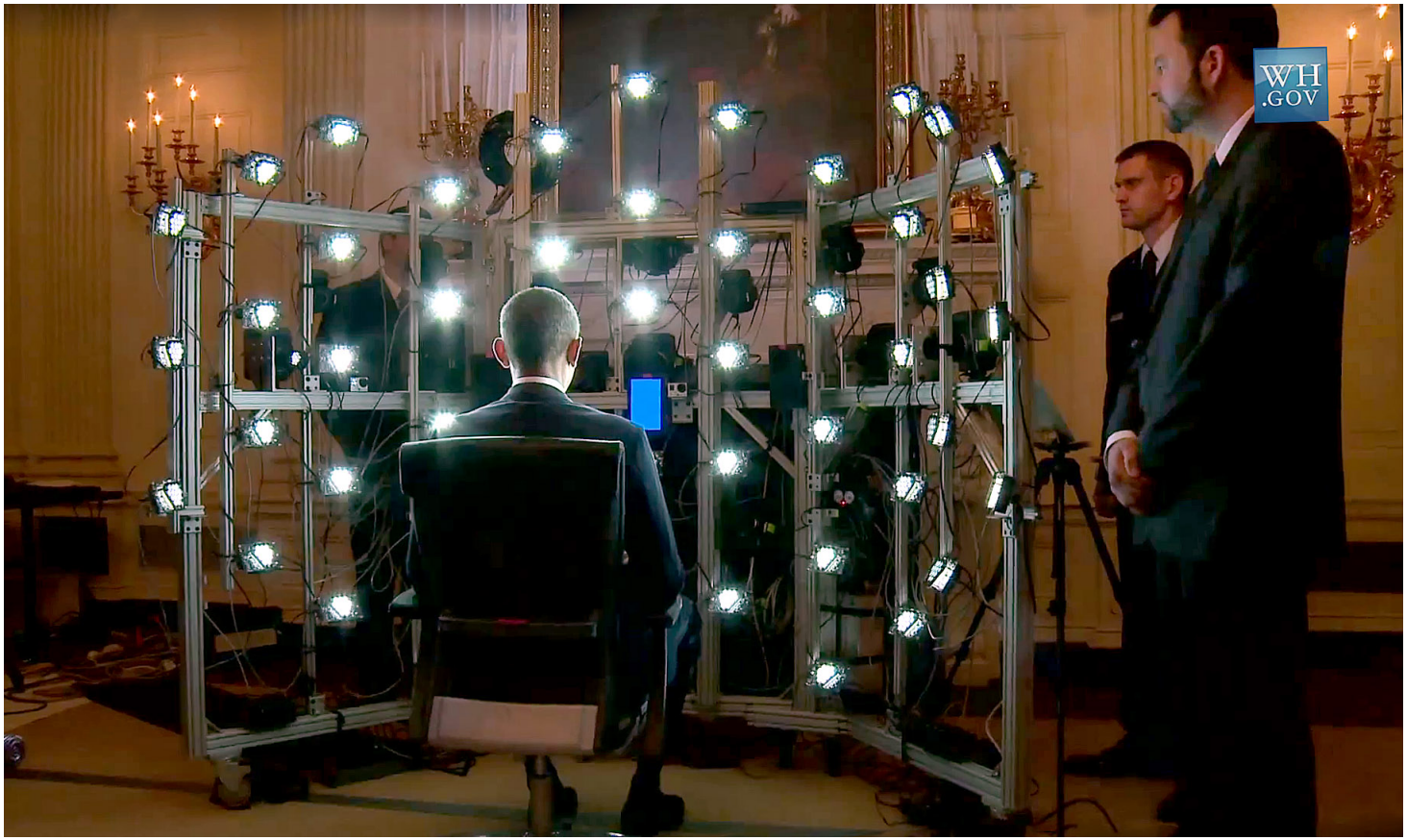




JH

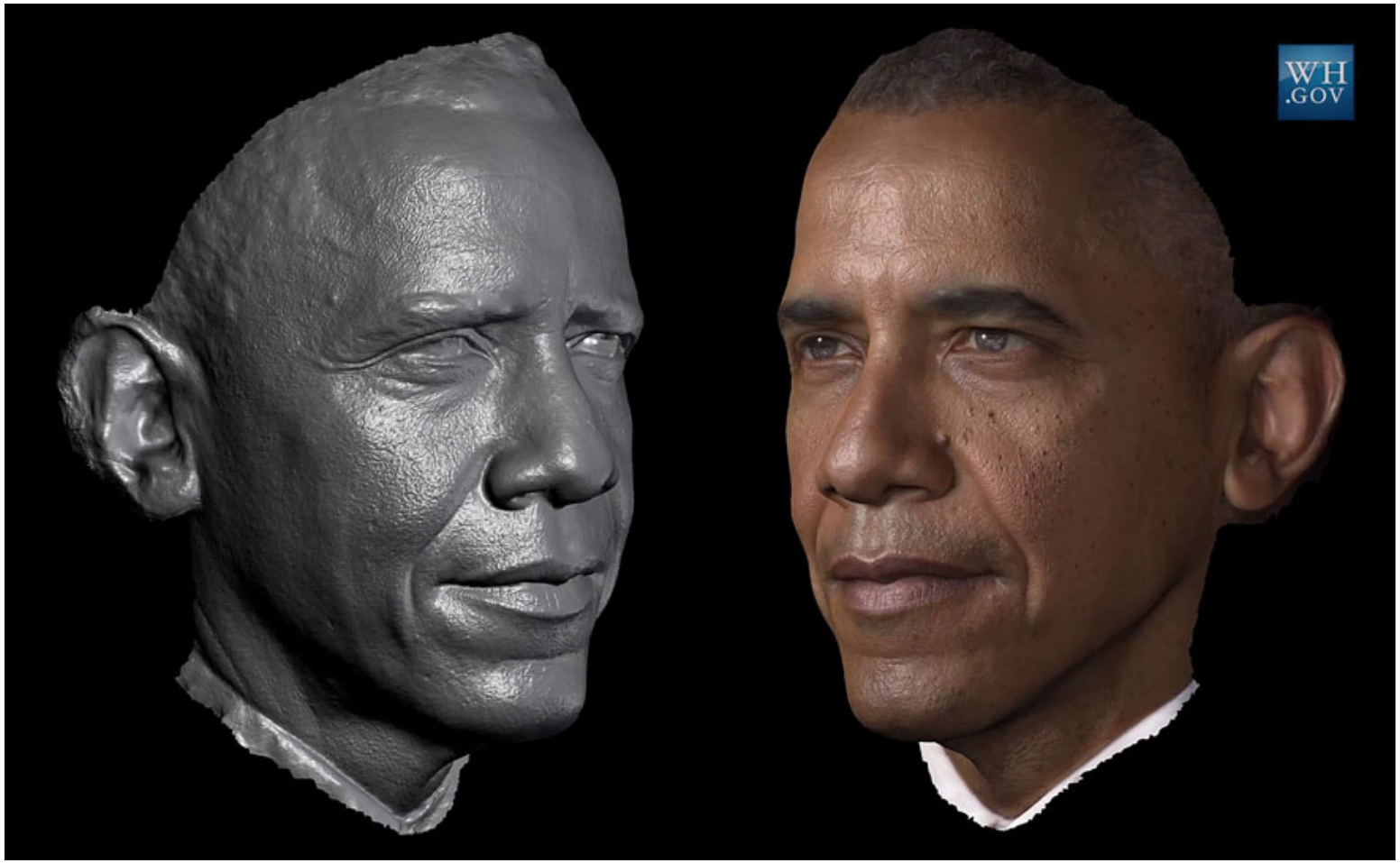
Login without a password…

Human shape capture



Human shape capture

Human shape capture



Human shape capture

Current State-of-the-Art

- Many of these are less than 5 years old
- This is a very active research area, and **rapidly** changing!
- Many new apps in the next 5 years

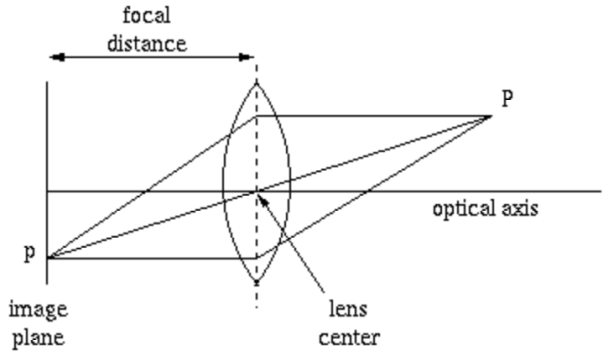
Coming Up in this Course

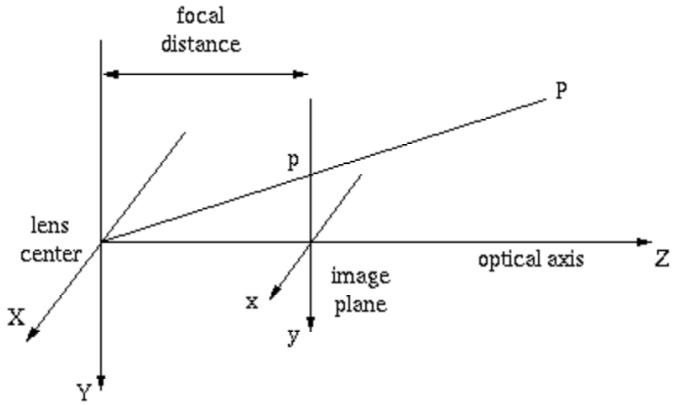
1. Camera geometry
2. Shape from X
3. Motion Estimation
4. Machine learning in computer vision

Coming Up in this Course

1. Camera geometry
2. Shape from X
3. Motion Estimation
4. Machine learning in computer vision

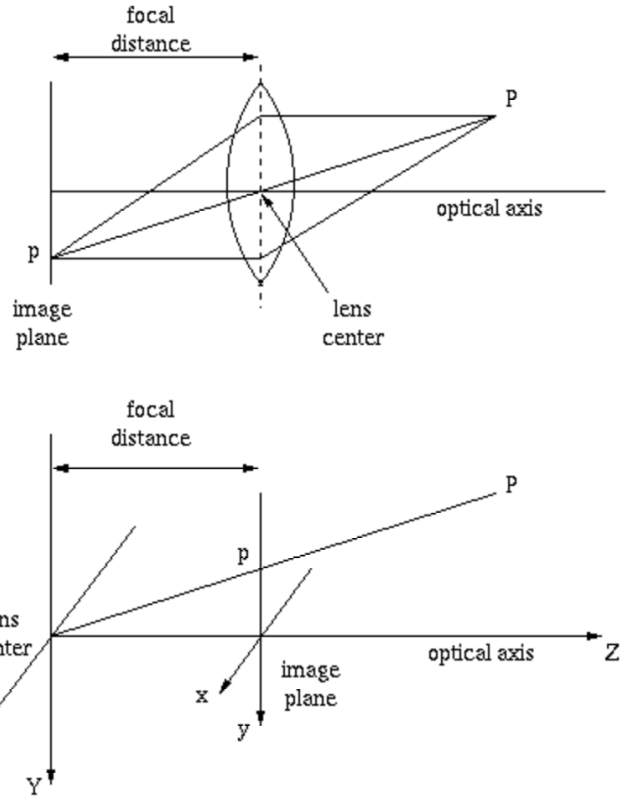
1. Camera Geometry





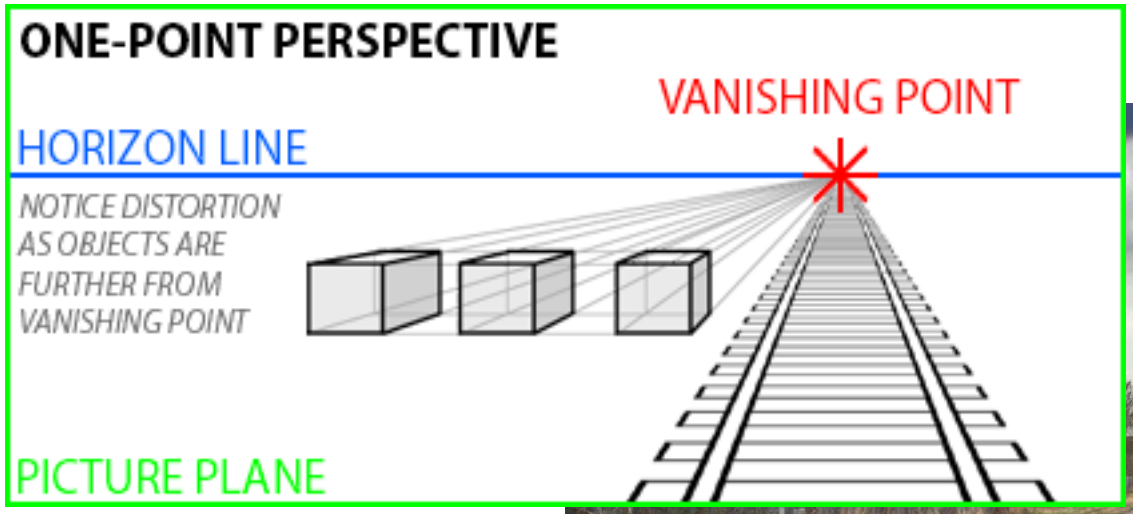
- Relationship between object coordinates (given by a vector **P** in 3D) and image coordinates (given by vector **p** in 2D)
- Effect of various intrinsic camera parameters (focal length of lens, nature of the lens, aspect ratio of sensor array, etc.) on image formation
- Effect of various extrinsic camera parameters on image formation

1. Camera Geometry

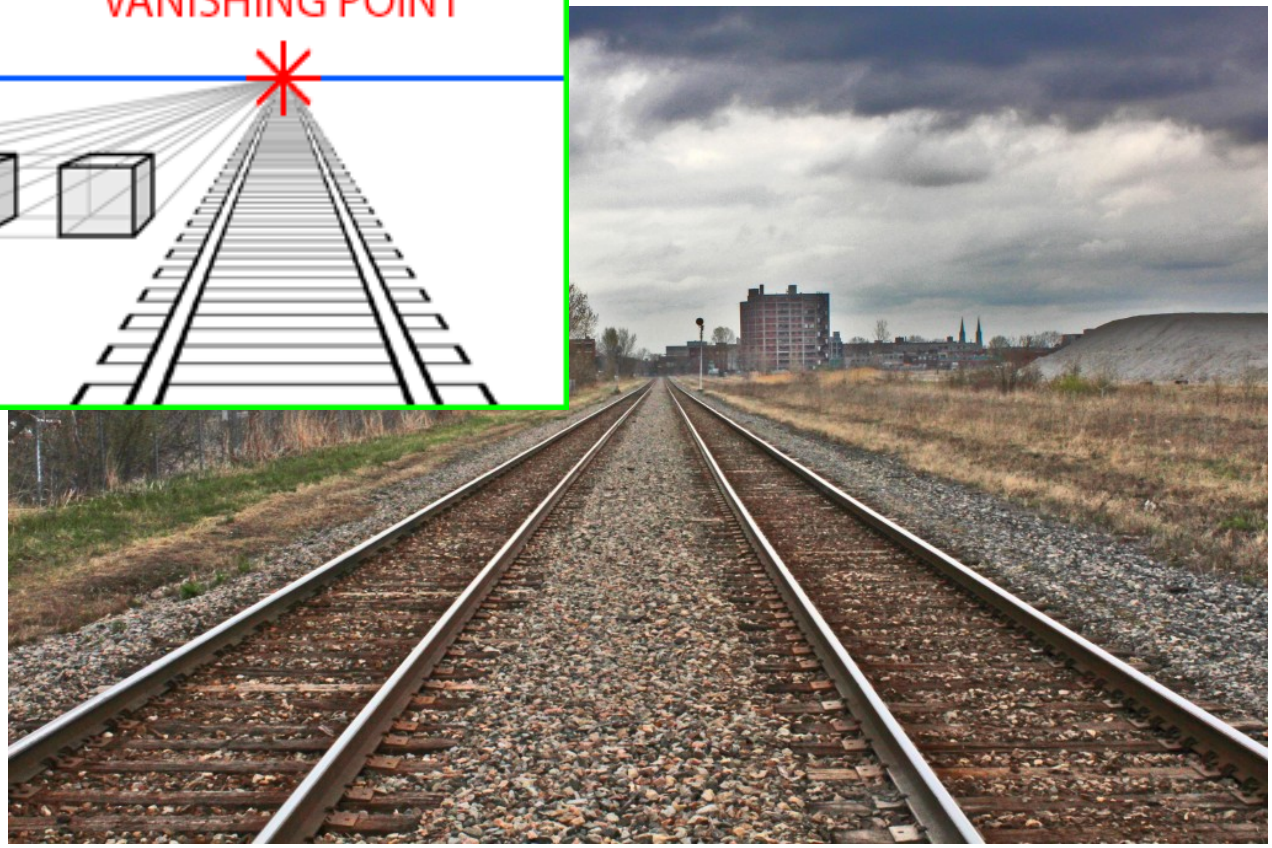


- Let's say we take a picture of a simple object of known geometry (example: chessboard, cube, etc.).
- Given the 3D coordinates of N points on the object, and their corresponding 2D coordinates in the image plane, can you determine the camera parameters such as focal length?
- Answer is **YES** we can! This process is called as camera calibration.

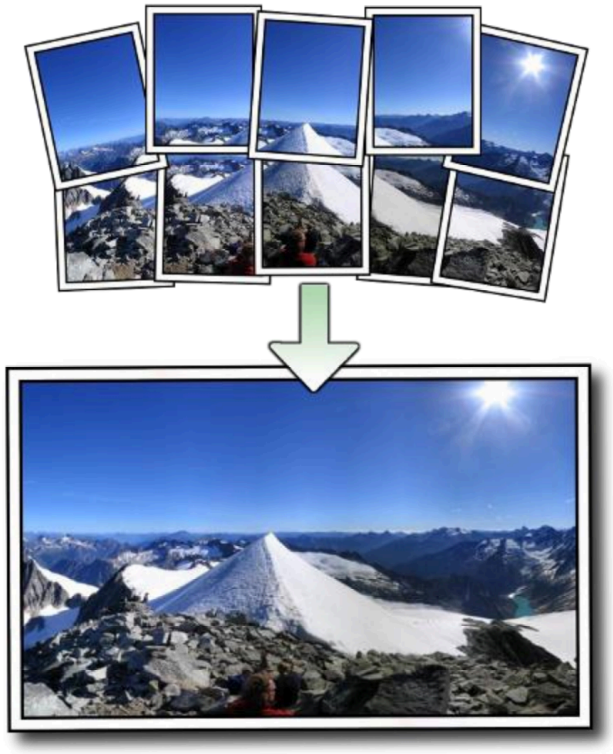
1. Camera Geometry (via Vanishing Points)



http://www.atpm.com/9.09/design.shtml



1. Camera Geometry - Image Mosaicing/Panoramas



We will study an end-to-end technique for generating a panorama out of a series of pictures of a scene from different viewpoints.

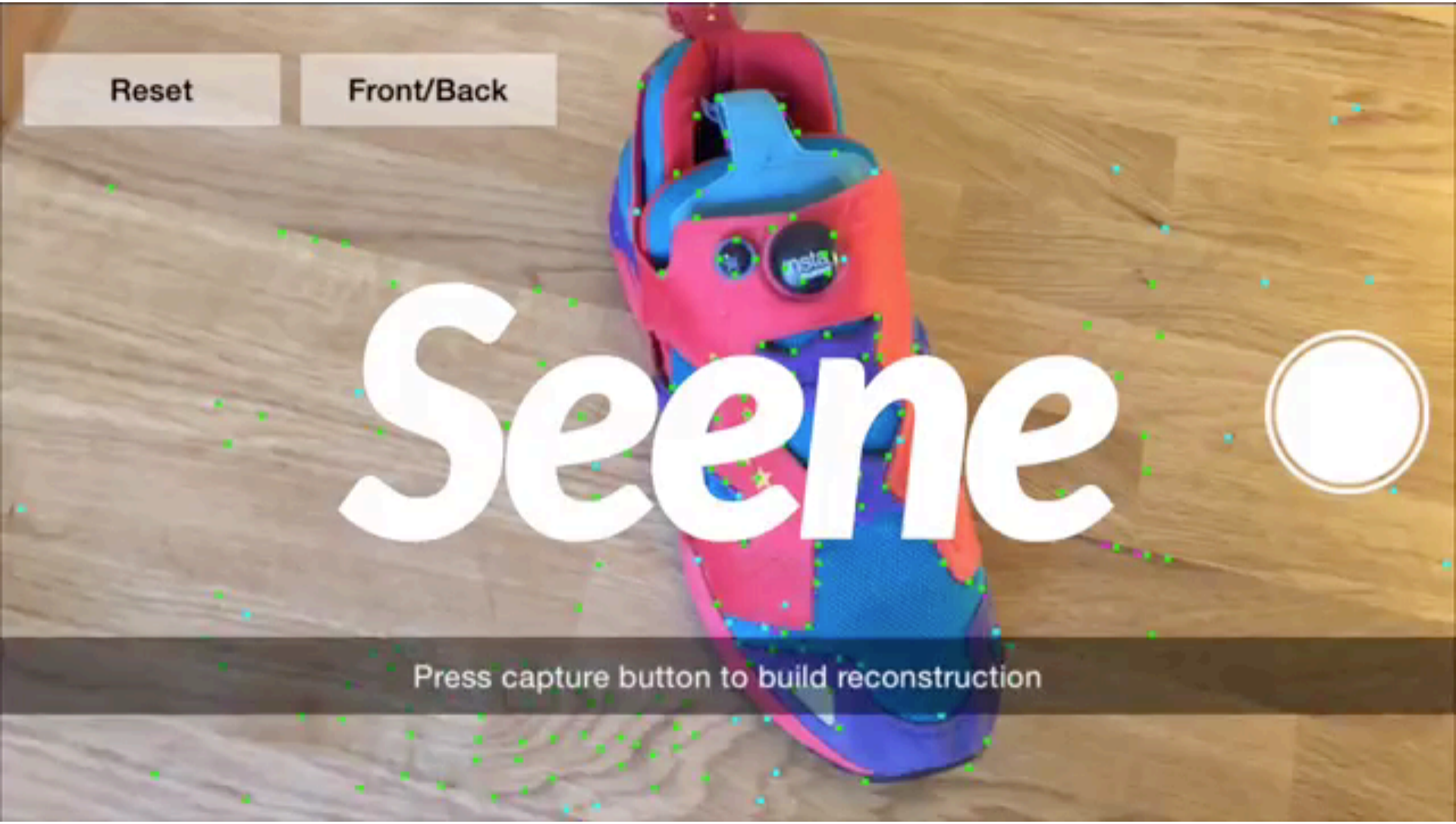
http://cs.bath.ac.uk/brown/autostitch/autostitch.html

Coming Up in this Course

1. Camera geometry
2. Shape from X
3. Motion Estimation
4. Machine learning in computer vision

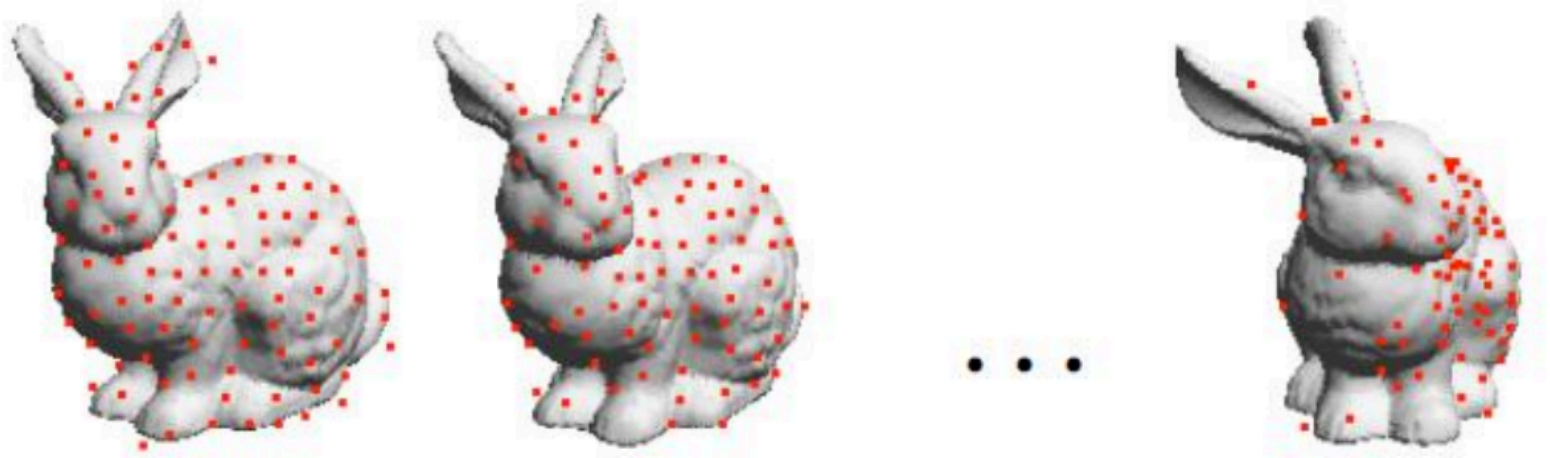
2. Shape from “X”

Structure from Motion



2. Shape from "X"

Structure from Motion



1. Input: Video sequence of moving (translating + rotating) object taken from a still camera
2. Solve: Tracks of some N 2D salient points from each frame of the video sequence (correspondence problem)
3. Outputs: 3D coordinates of each of those N points in each frame + 3D motion of the object!

2. Shape from "X"

Depth from Stereo and Disparity

http://3dstereophoto.blogspot.in/2015/06/

Left Image

Right Image

Output Depth Map

2. Shape from “X”:

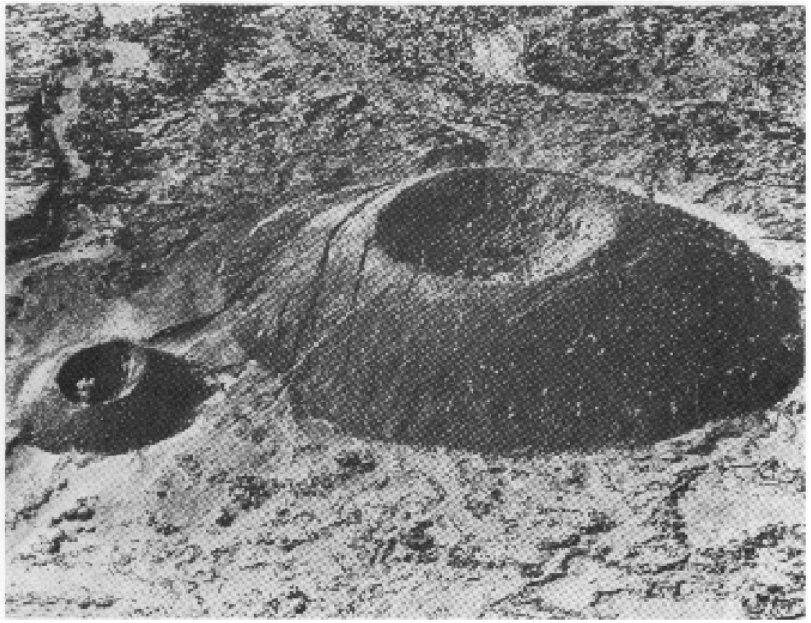
Shape from Shading

- An image is 2D. But most underlying objects are 3D.
- Can you guess something about the 3D structure of the underlying object just given the 2D image?
- The human visual system does this all the time!
- We want to reproduce this effect computationally (the “holy grail” of computer vision)

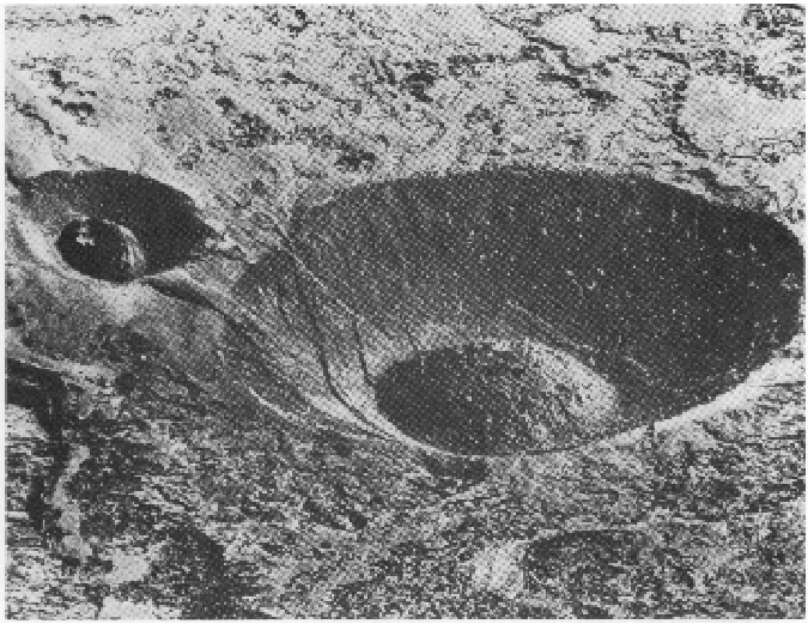
2. Shape from "X"

Shape from Shading

(a)



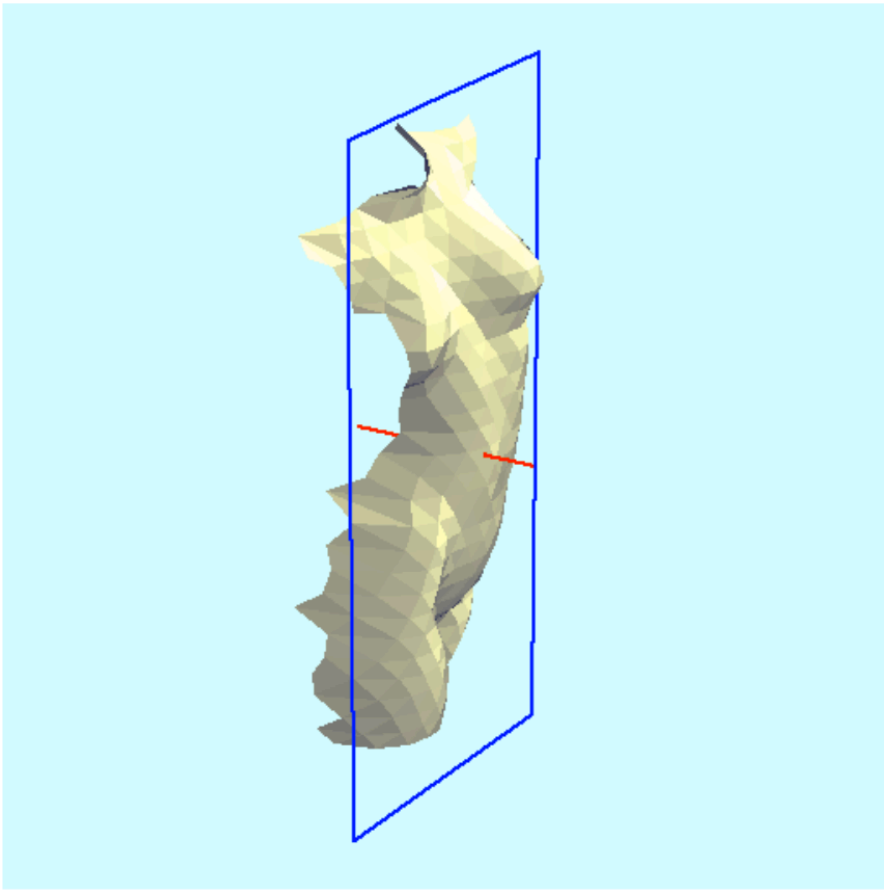
(b)



Shading influences shape. The image in (a) has the appearance of mound of dirt with a small indentation. The image in (b) appears to contain a crater with a mound at the top. Yet, the two images are the same except for an up-down flip

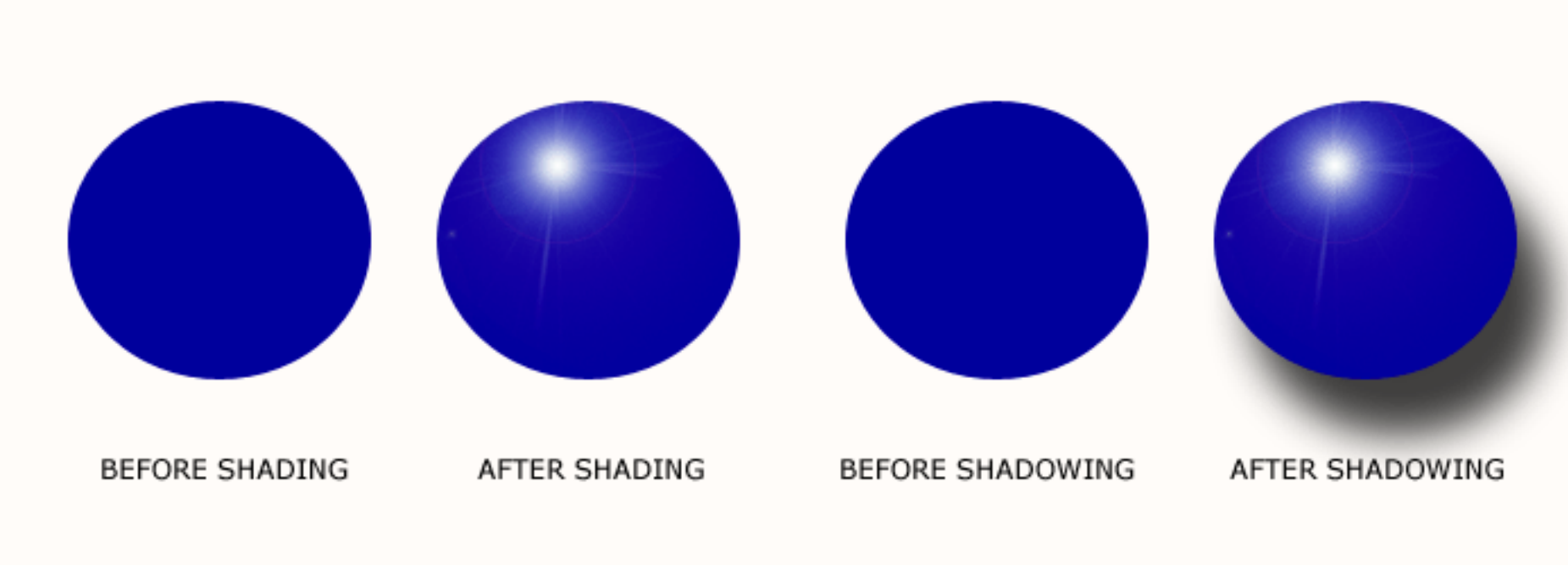
2. Shape from “X”

Shape from Shading



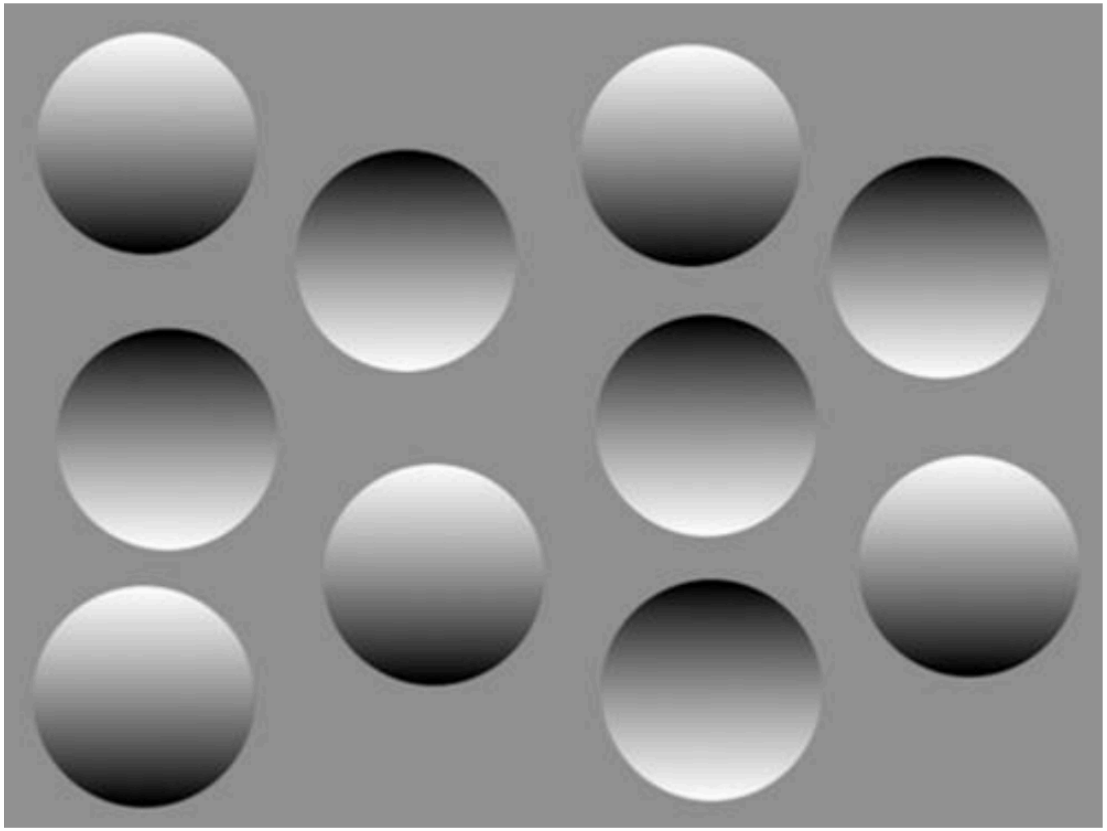
2. Shape from “X”

Shape from Shading



2. Shape from “X”

Shape from Shading



Crater vs mound?

Coming Up in this Course

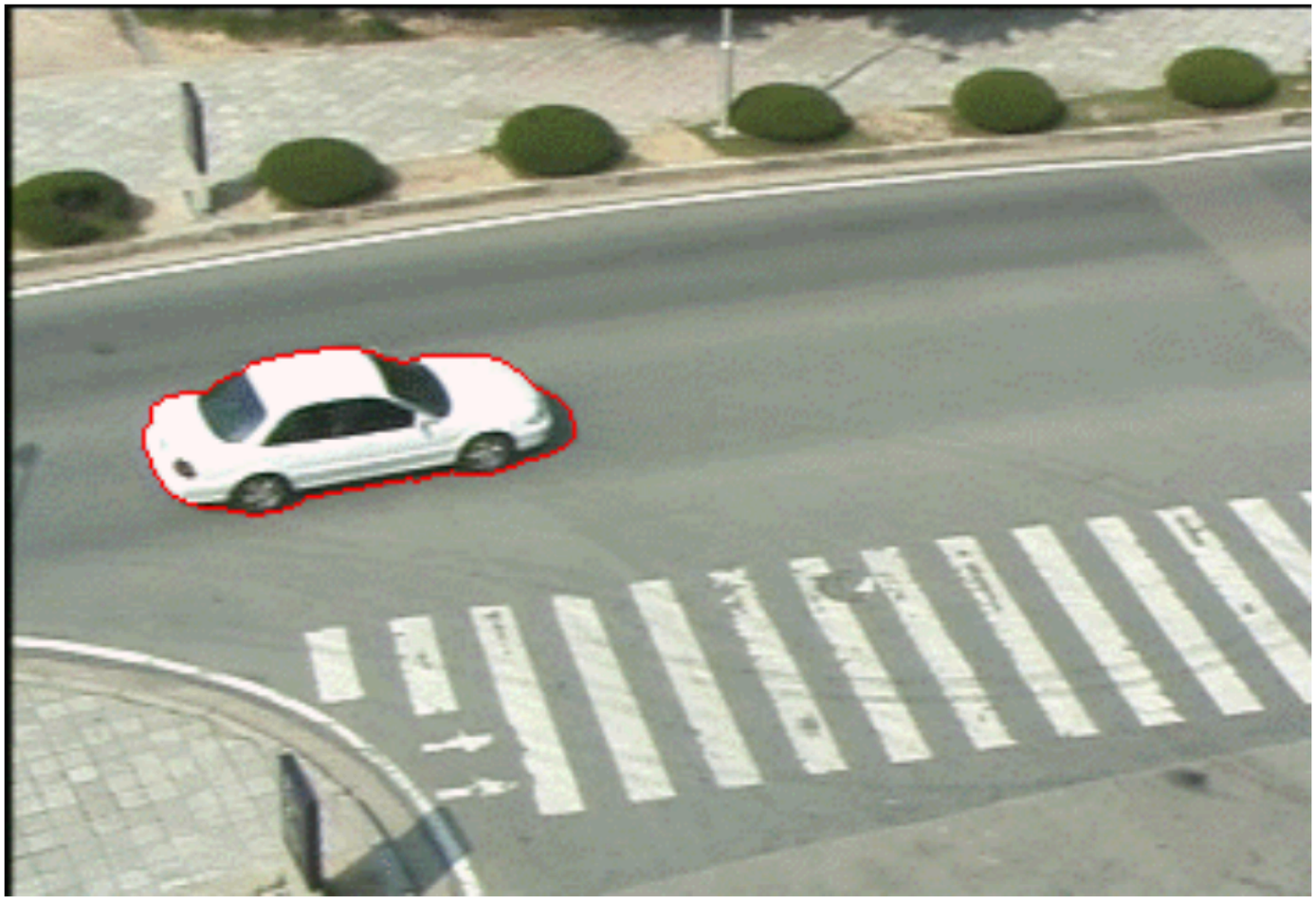
1. Camera geometry
2. Shape from X
3. Motion Estimation
4. Machine learning in computer vision

3. Motion Estimation

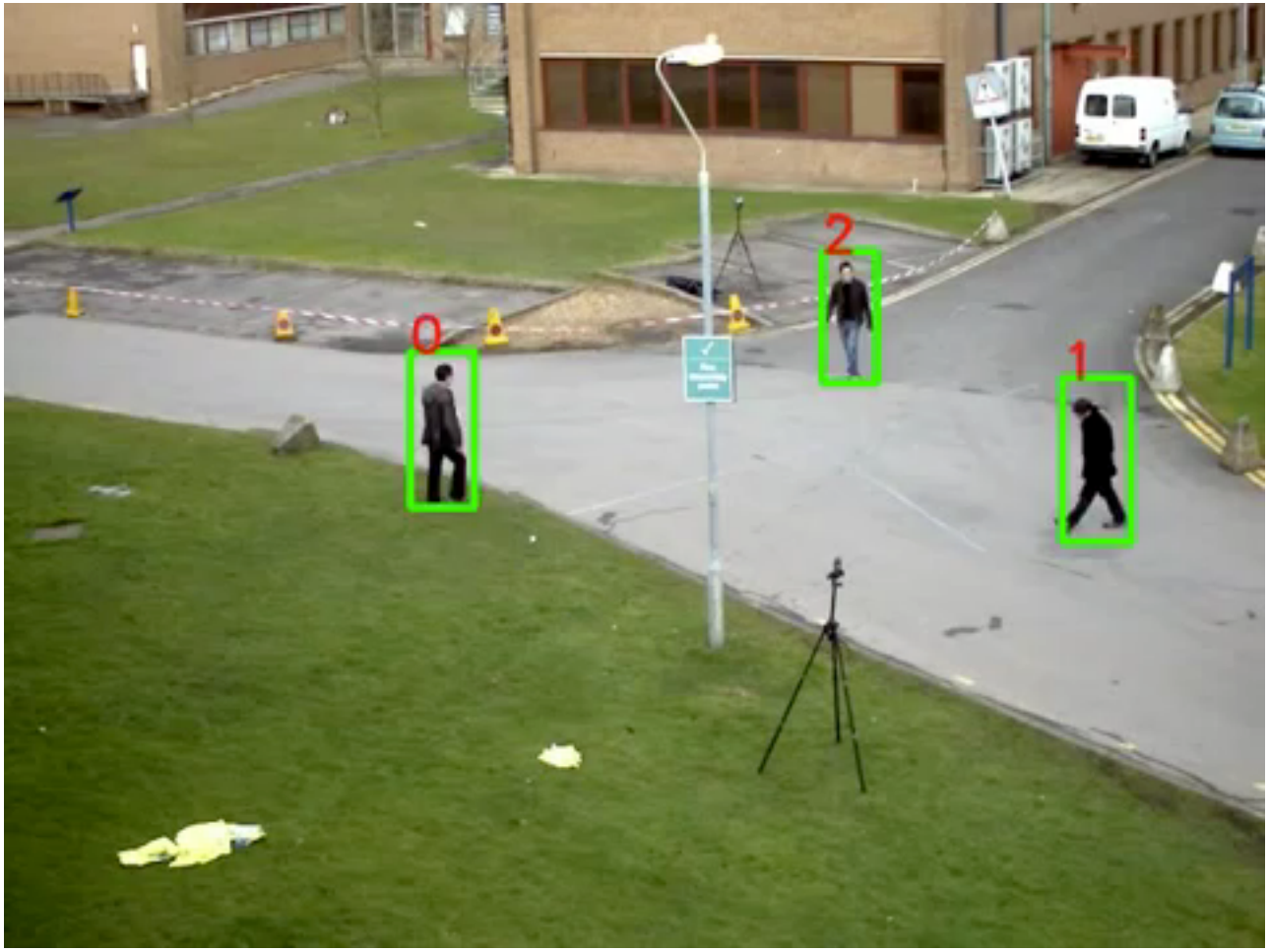
- A video sequence is very rich in information content
- Movement brings in most of this information
  - Movement allows objects identification
  - Image characteristics are coherent along motion trajectories
- Motion detection: binary decision (motion or no motion)
- Motion estimation: measure the movement

3. Motion Estimation

Other types of Motion Estimation:
1. Object Tracking

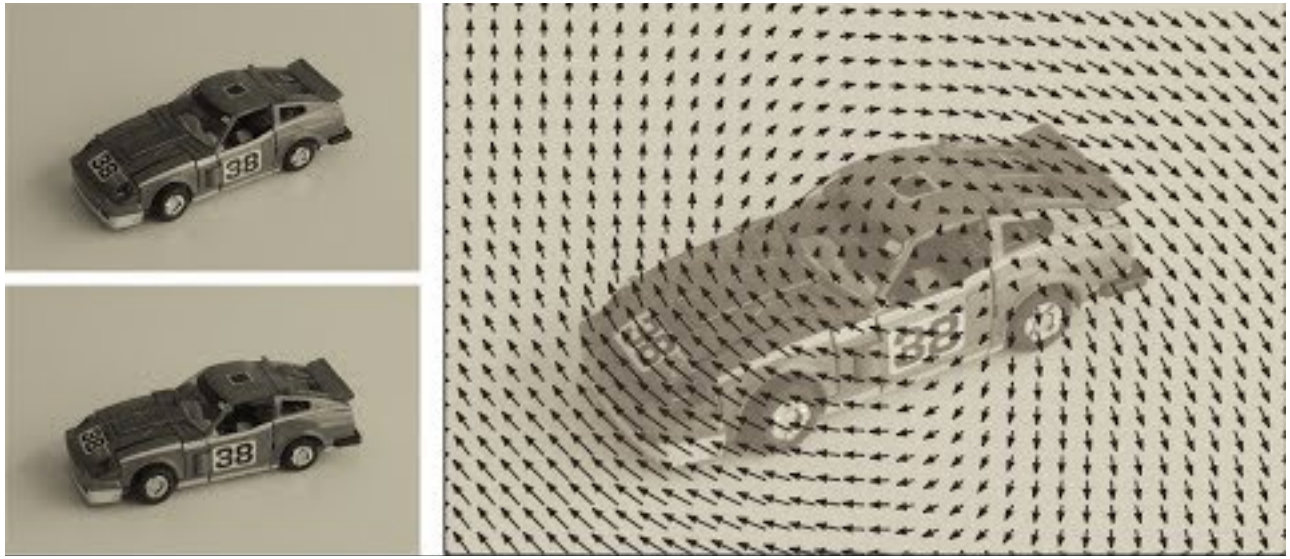


Other types of Motion Estimation: 2. Multiple Object Tracking



3. Motion Estimation

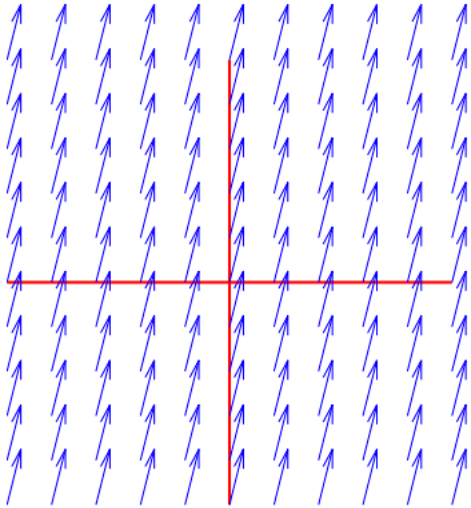
- **Input**: a video sequence
- **Target**: an estimate of the motion(2D) at all pixels in all frames
- Applications of such an algorithm: object tracking, video stabilization, etc.
- Typical assumptions: small motion between consecutive frames



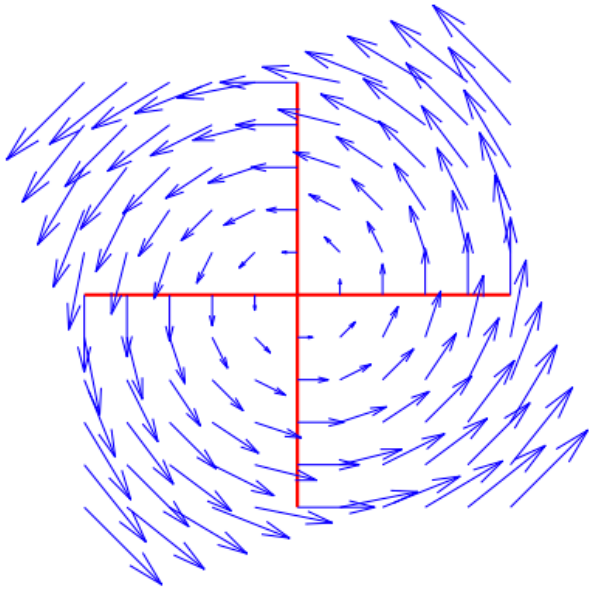
https://www.youtube.com/watch?v=KoMTYnlNNnc

3. Motion Estimation

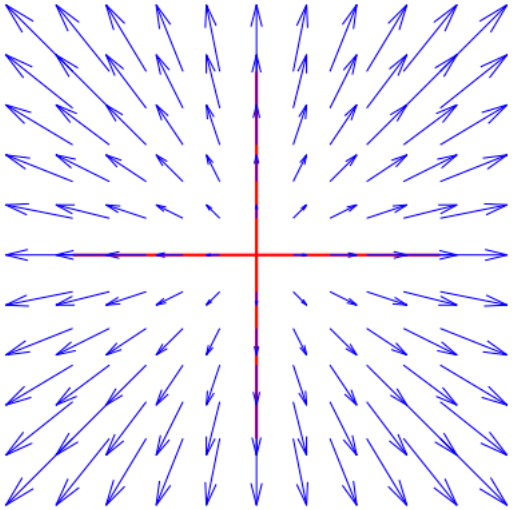
- Sometimes the motion between two images can be represented more compactly – e.g.: rotation, scaling, translation, etc.
- We will look at methods to estimate such “parametric motion”



Translation



Rotation



Scaling

Coming Up in this Course

1. Camera geometry
2. Shape from X
3. Motion Estimation
4. Machine learning in computer vision

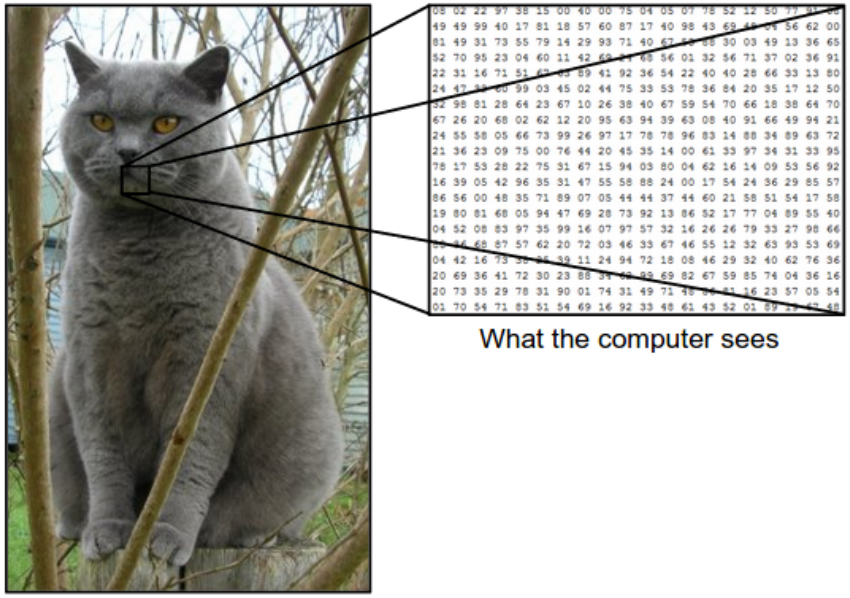
4. Machine Learning in Computer Vision

- Why do we need to do machine learning?

Images are represented as 3D arrays of numbers, with integers between [0, 255].

E.g.
300 x 100 x 3

(3 for 3 color channels RGB)



What the computer sees

- Why do we need to do machine learning? **No way to hand code it!**

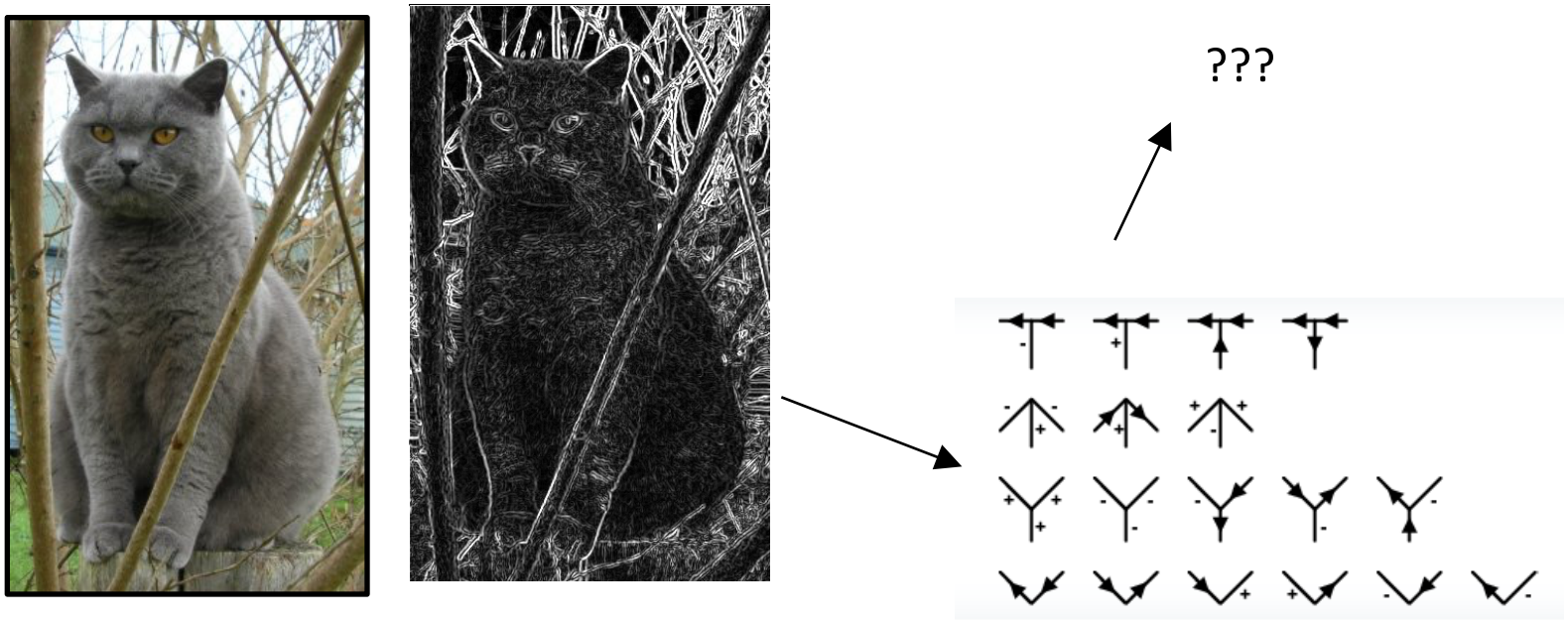
Image classification:

```
function predict(image)
  -- ????
  return class_label
end
```

- Unlike e.g. sorting a list of numbers
- No obvious way to hard-code the algorithm for recognizing a cat, or other classes

- People have attempted

- Image classification:



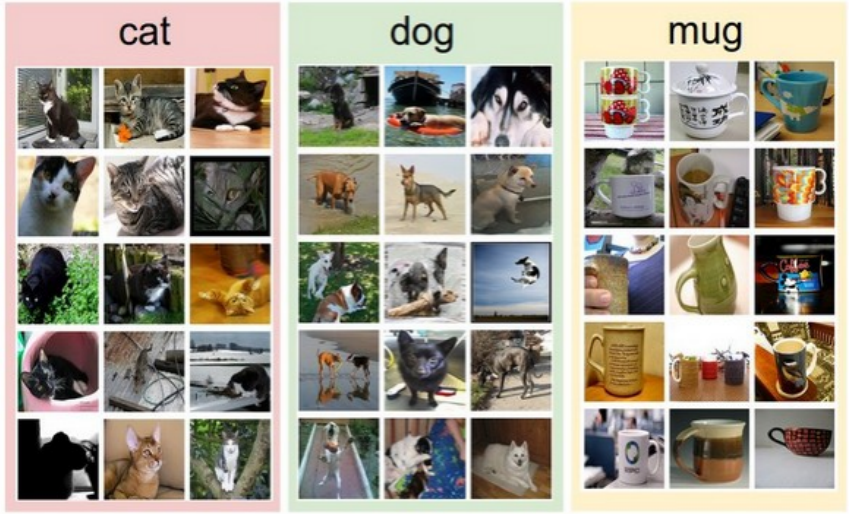
The Data Driven Paradigm

1. Collect a dataset of images and labels
2. Use Machine Learning to train an image classifier
3. Evaluate the classifier on a withheld set of test images

```
function train(train_images, train_labels)
  -- Build model: images -> labels
  return model
end


function predict(model, test_images)
  -- Predict test_labels using the model
  return test_labels
end
```

Example Training Set

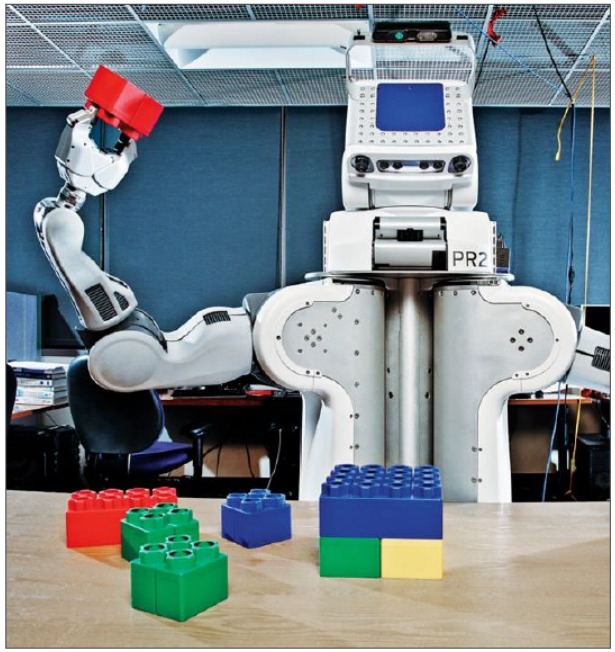


4. Machine Learning in Computer Vision (Deep Learning)

- We will cover basics of Neural Nets, MLPs, backpropagation, stochastic gradient descent
- Will cover different architectures: convolutions neural nets, siamese nets, triplet nets, GANs, capsules (if time permits)
- Compression of neural network architectures (for real-time or low-power applications)
- Applications in human pose estimation, finding point correspondences in images for 3D point cloud reconstruction, neural art, etc.

4. Machine Learning in Computer Vision (Deep Learning)

- Deep Learning == AI



4. Machine Learning in Computer Vision (Deep Learning)

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

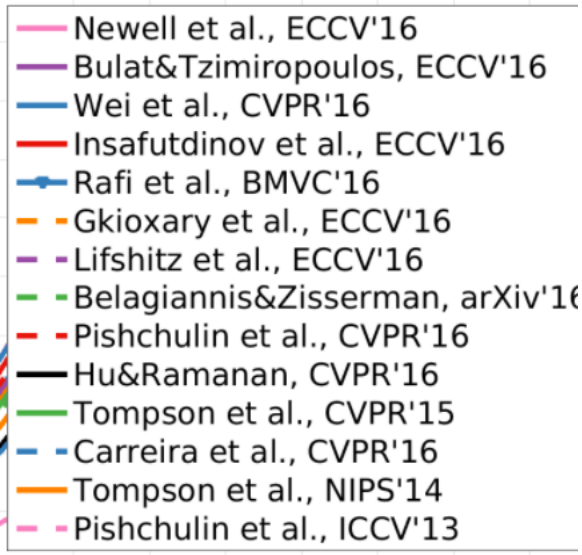
"two young girls are playing with lego toy."

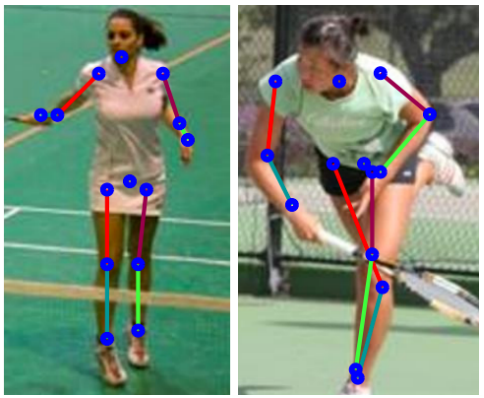
"girl in pink dress is jumping in air."

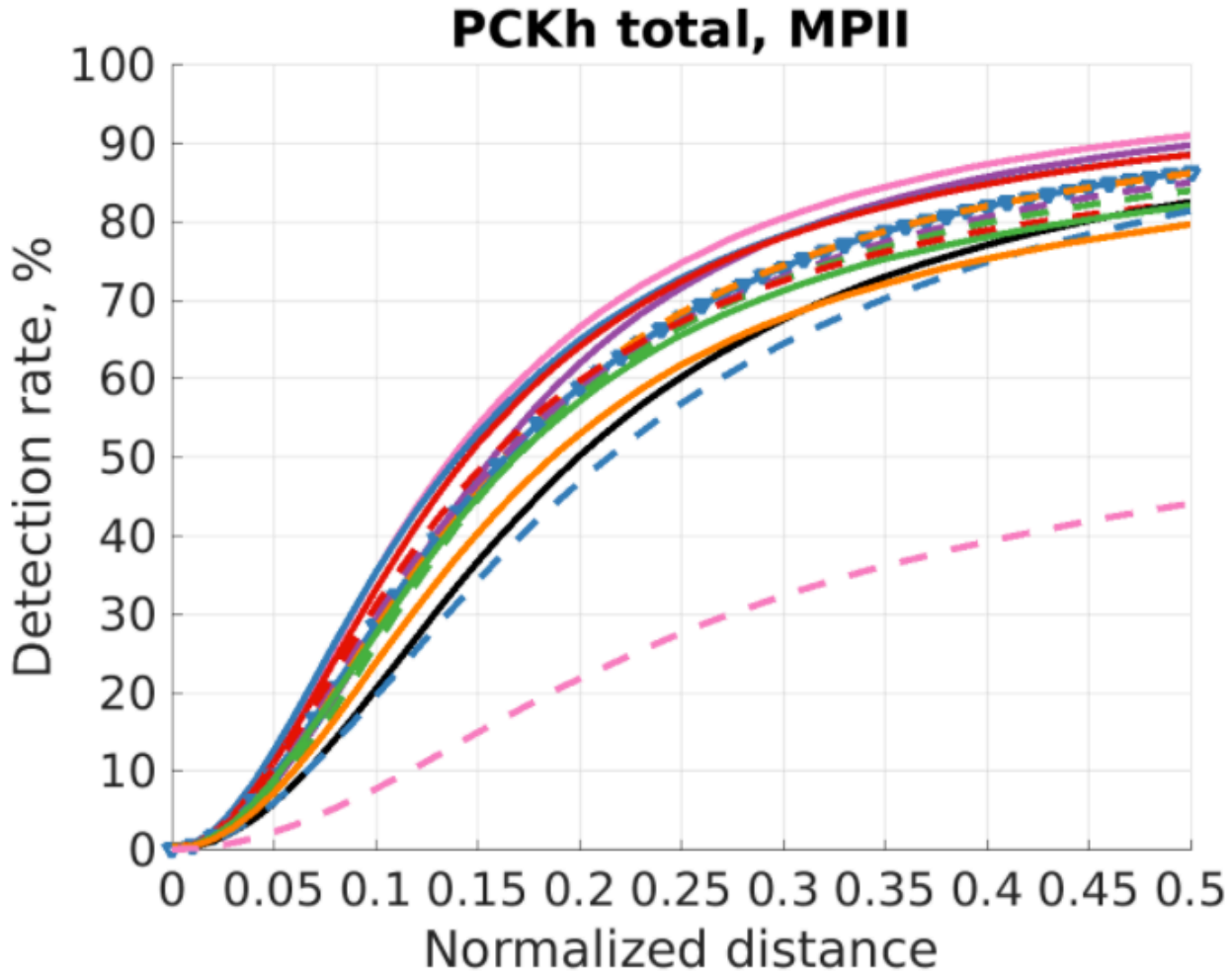
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

4. Machine Learning in Computer Vision (Deep Learning)







3D Pose Estimation from Video

Rishabh Dabral

Anurag Mundhada

Uday Kusupati

Safeer Afaque

Arjun Jain

Instructor: Arjun Jain

IIT Bombay
Dept. of Computer Science and Engineering

Max Planck Institute
Germany


Apple SPG

TA Introduction

Rishabh Dabral

Rishabh is a PhD candidate at IIT Bombay, and his thesis is on automatic human sensing. Prior to joining IIT, he has worked at Canon Inc., on performance modelling of SoCs. He holds a B.Tech in Computer Science from IIIT Jabalpur.

Safeer Afaque

Safeer is currently pursuing his PhD at IIT Bombay. He has served as an Assistant Professor at SRM University and has undertaken several projects on surveillance, self-driving cars and robotics. He holds an M.Tech in Robotics from IIIT Allahabad.

Administrative

- Classes: Thursdays and Fridays 7:00-8:30pm, CC 103
- Course policies: details on course website.
- Instructors: Arjun Jain Room 216 CSE New Building
- Assistants: Rishabh Dabral, Safeer Afaque
- Prerequisite: Visual Computing (or equivalent)
- Reference Textbook: Computer Vision: Algorithms and Applications by Rick Szeliski http://szeliski.org/Book/
- Webpage: https://github.com/cs763/Spring2018/ (slides, assignments, extra reading, grading policy, …)
- **CS663 a hard prerequisite**

Grading Policy

- Mid-sem exam: 20%
- Final exam (cumulative): 20%
- Assignments (five or six): 35% (all to be done in groups of 2-3 students)
- Course project: 20% (to be done in the same group of 2-3 students)
- Class participation: 5%
- Course project work will be presented by the student group during a viva at the end of the course. During this viva, each student in the group will be separately questioned, not only on the project work, but also the assignments. Each student is expected to contribute to each and every assignment and the course project.
- Audit requirements: You must write both exams, submit all assignments and the project, and score at least 40% to get an AU.

Pre-requisites: Math Tools

- Numerical linear algebra (eigenvectors and eigenvalues, SVD, matrix inverse and pseudoinverse)
- Vector calculus: Gradients and Jacobians, Laplacian Operation
- Signal processing concepts: Convolution, Correlation
- Optimization basics: gradient descent, regularization, etc. (we will cover some of this in class)

Pre-requisites: Programming Tools

- MATLAB and associated toolboxes, maybe some python
- For deep learning: Torch7 (lua)

Thanks to:

- A lot of material from Ajit Rajwade's CS763 course
- Thanks to Marc Pollefeys and JamesTompkin for some slides
- In case I made a mistake or missed acknowledging someone, please let me know.

Arjun Jain
ajain@cse.iitb.ac.in