Introduction à la vision artificielle

Jean Ponce and Josef Sivic
Email: Jean.Ponce@ens.fr and Josef.Sivic@ens.fr
Web: http://www.di.ens.fr/willow/

Planches après les cours sur :
http://www.di.ens.fr/~ponce/introvis/lect1.pptx
http://www.di.ens.fr/~ponce/introvis/lect1.pdf

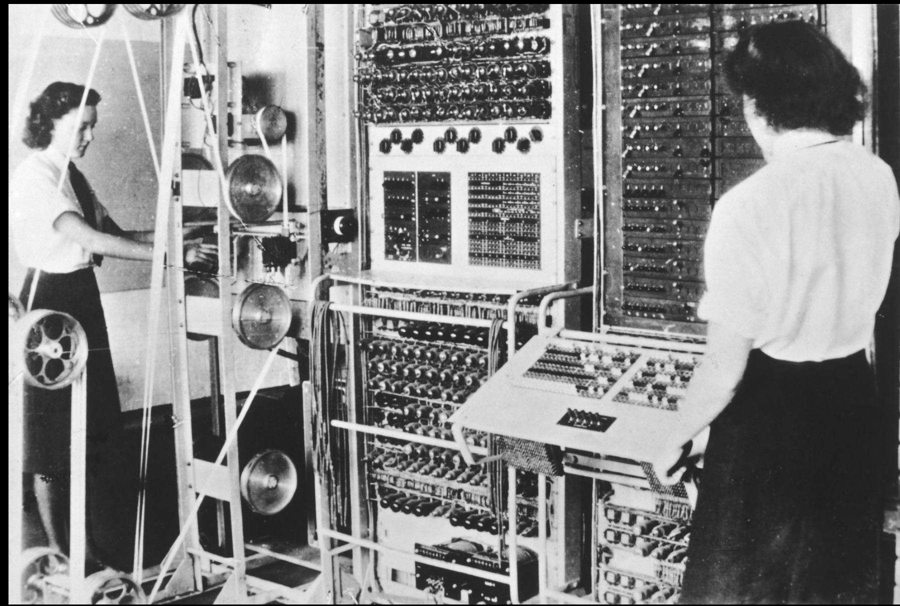
Today with some slides from J. Hays and S. Seitz

Change of time for the class

From next week the class is on:

Wednesdays 14:00-17:00 in the same room S16

Next class: Wednesday 21/9 14:00-17:00 in S16



Description:

- Street scene
- Bar
- Chairs
- People drinking coffee
- Ashtray, etc.

Computer vision

… extracting information from images and video

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 really hard

WHY IS VISION DIFFICULT ?

Too much information:

- 1000x1000x24xN bits;
- matching n features against n features costs n!;
- shadows, highlights, texture..

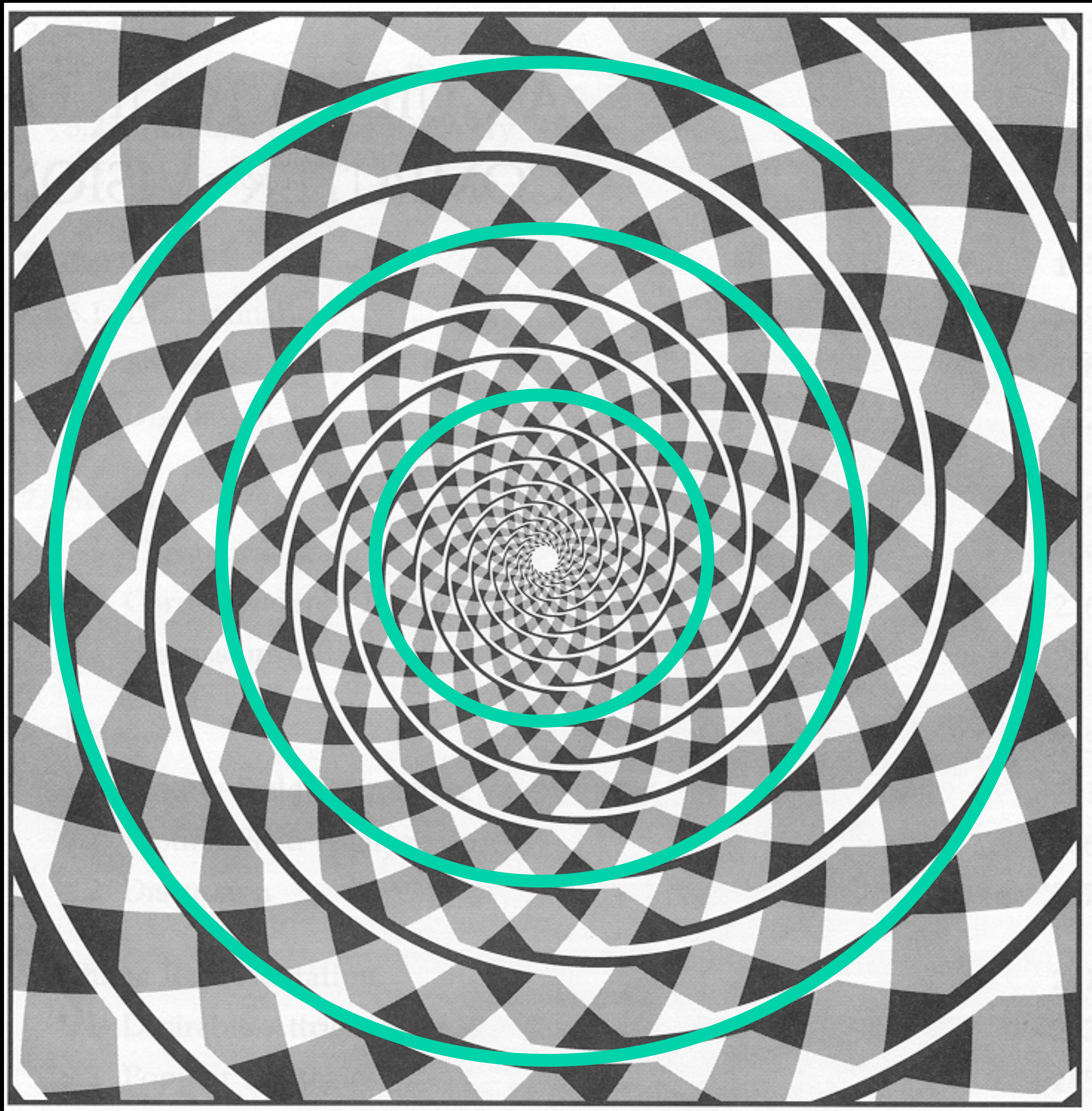
Too little information:

- Physical properties (depth, orientation, reflectance..) of the world are not directly observable.

What are appropriate representations?

- for images, object instances, object classes, video content and the interpretation process..

What are appropriate algorithms and architectures?



(Nalwa, 1993)

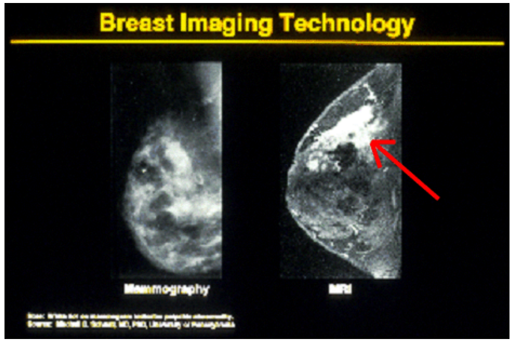
J.J. Koenderink, www.gestaltrevision.be/en/resources/clootcrans-press

COMPUTER VISION IS INTERESTING.

- We know it is possible.
- We know it is difficult.
- We don't know how to do it.

Why computer vision matters

Safety

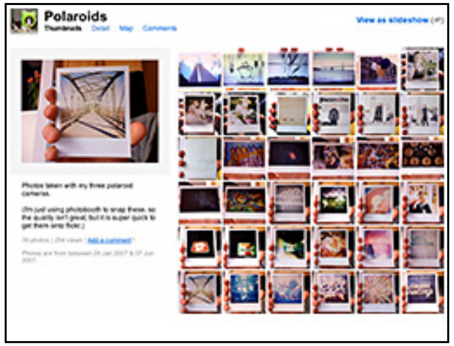


Health

Security

Comfort

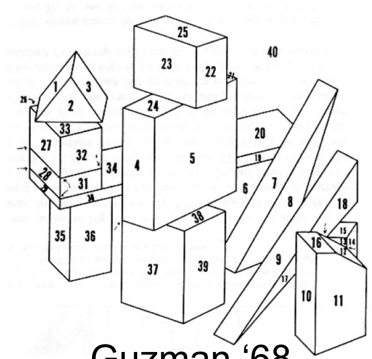
Fun



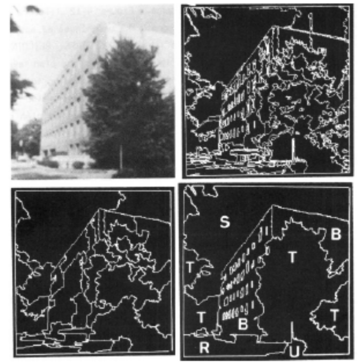
Access

Ridiculously brief history of computer vision

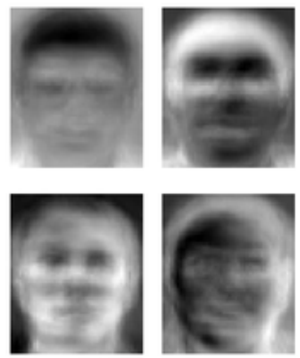
- 1966: Minsky assigns computer vision as an undergrad summer project
- 1960's: interpretation of synthetic worlds
- 1970's: some progress on interpreting selected images
- 1980's: ANNs come and go; shift toward geometry and increased mathematical rigor
- 1990's: face recognition; statistical analysis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



Guzman '68



Ohta Kanade '78



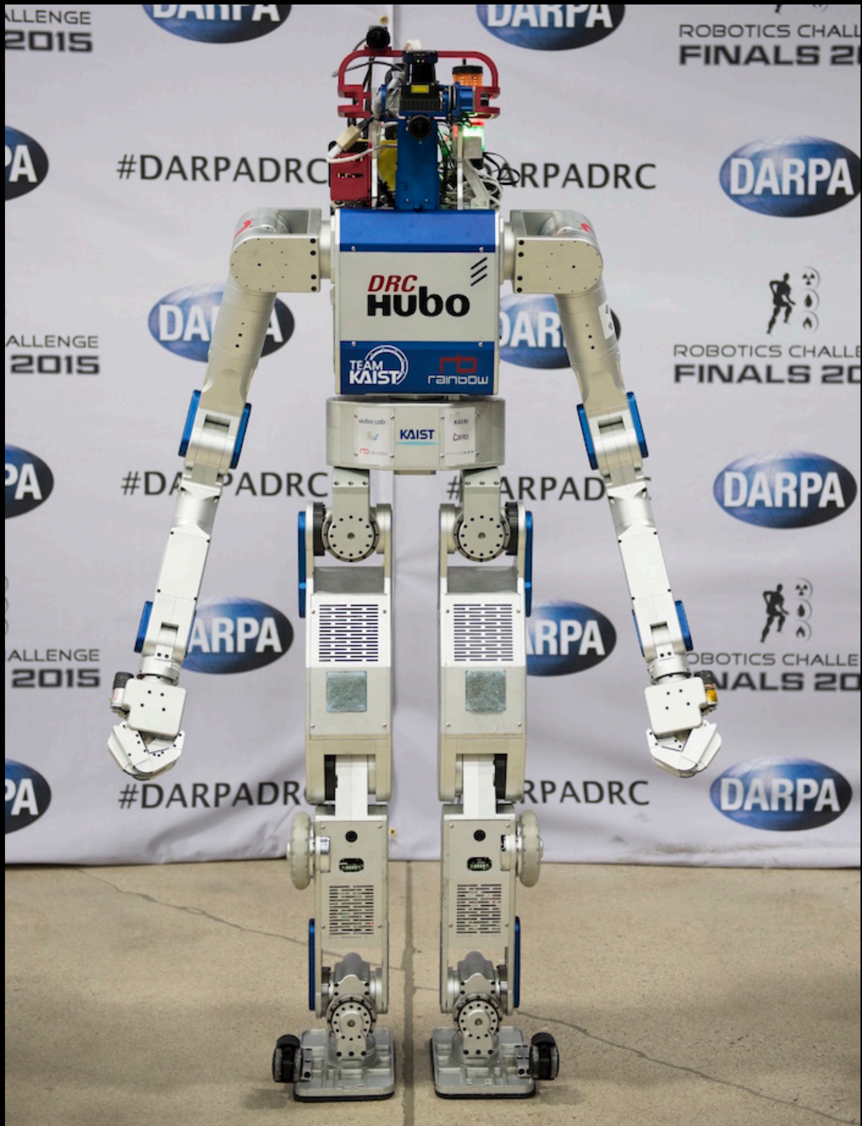
Turk and Pentland '91

WHAT IS COMPUTER VISION GOOD FOR?

Traditionally:
- Manufacturing: inspection, bin picking;
- Defense: ATR, photogrammetry, surveillance;
- Robotics: navigation, visual servoing.

Recently:
- Computer graphics, medical imaging, HCI
- 3D vision and recognition
- The Web, Internet, social networks.
- Robotics again.

Really:
- Understanding the principles of object recognition;
- Building the robots of tomorrow, for home and space;
- Understanding how people tick;
- It is just difficult, fun, and interesting.



KAIST's Hubo

CMU's Chimp

How vision is used now

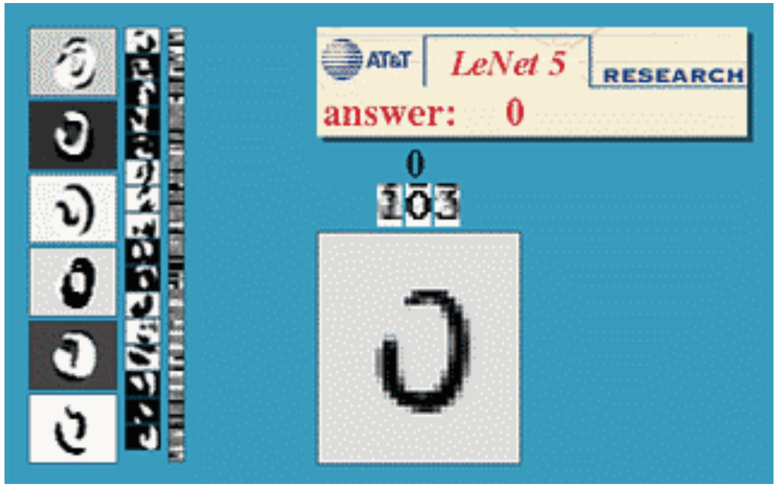
- Examples of recent real world applications

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



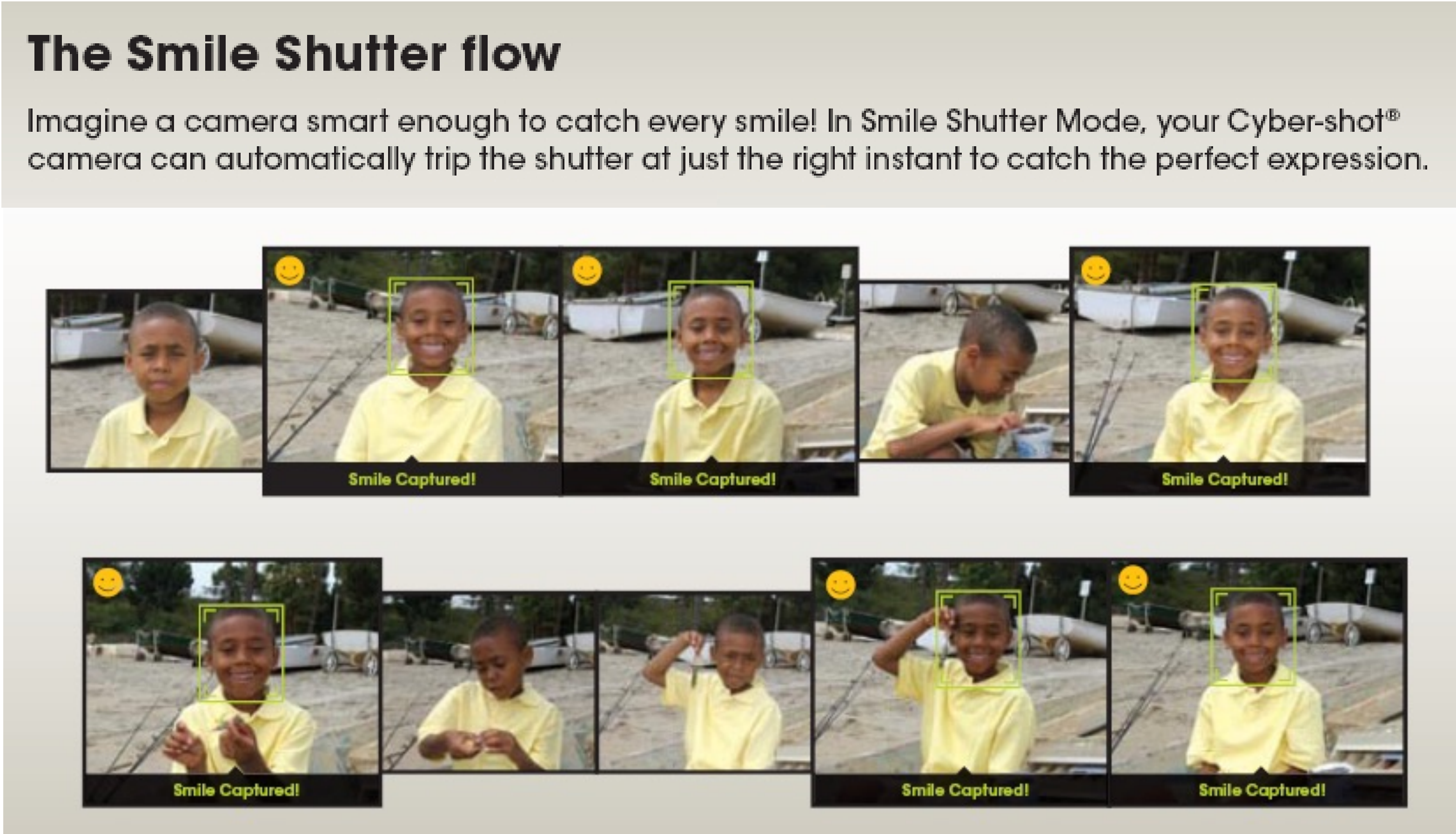
Digit recognition, AT&T labs
http://www.research.att.com/~yann/

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

Face detection

- Many new digital cameras now detect faces
  - Canon, Sony, Fuji, …

Smile detection



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

3D from thousands of images

Agarwal et al. 2009

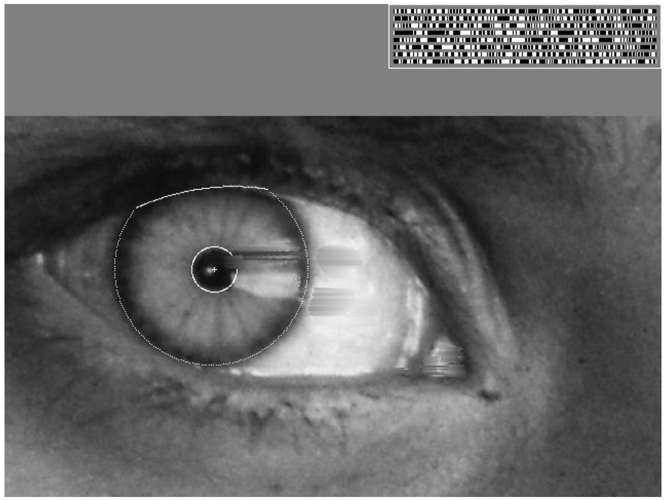
Object recognition (in supermarkets)

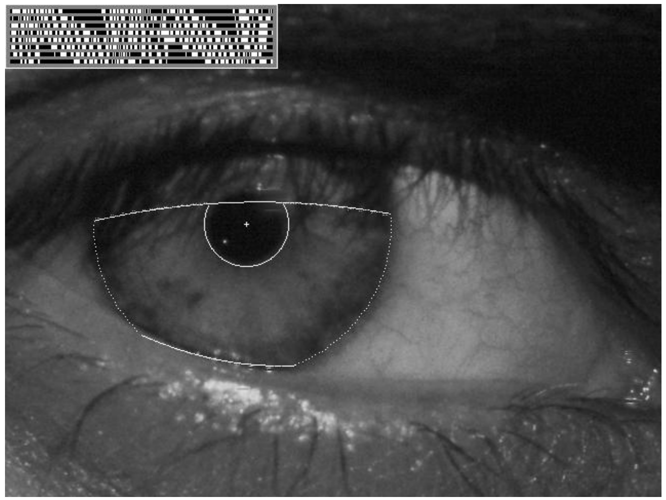
LaneHawk by EvolutionRobotics

"A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized, the cashier verifies the quantity of items that were found under the basket, and continues to close the transaction. The item can remain under the basket, and with LaneHawk,you are assured to get paid for it… "

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]()

Special effects: shape capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC

Special effects: motion capture

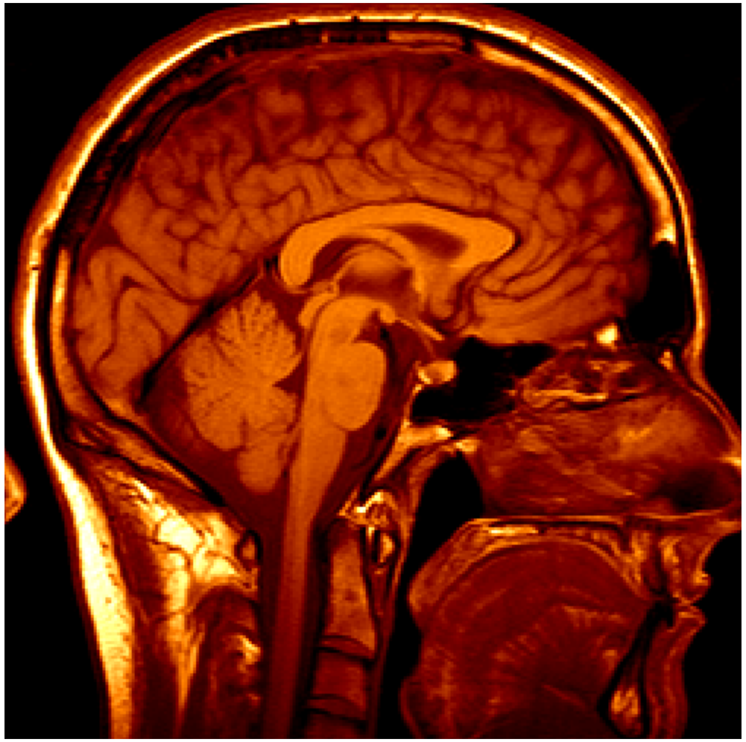
Pirates of the Carribean, Industrial Light and Magic

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

Smart cars

Slide content courtesy of Amnon Shashua

- Mobileye
  - Market Capitalization: 11 Billion dollars
  - See also CVPR 2016 keynote

Google cars

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

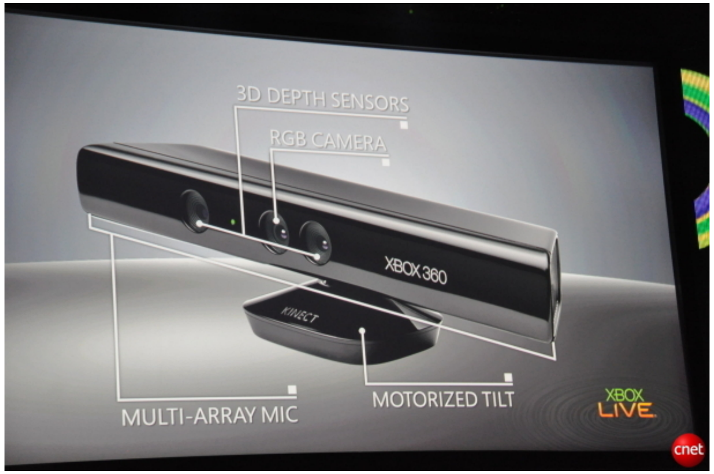
CATEGORIES
Automotive

WEBSITE
http://www.ford.com/

Full profile for Ford Motor Company

Interactive Games: Kinect

- 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

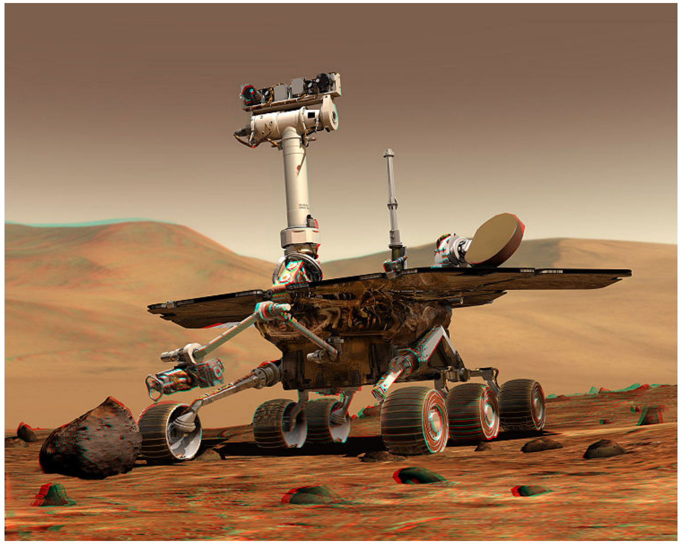
Vision in space

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

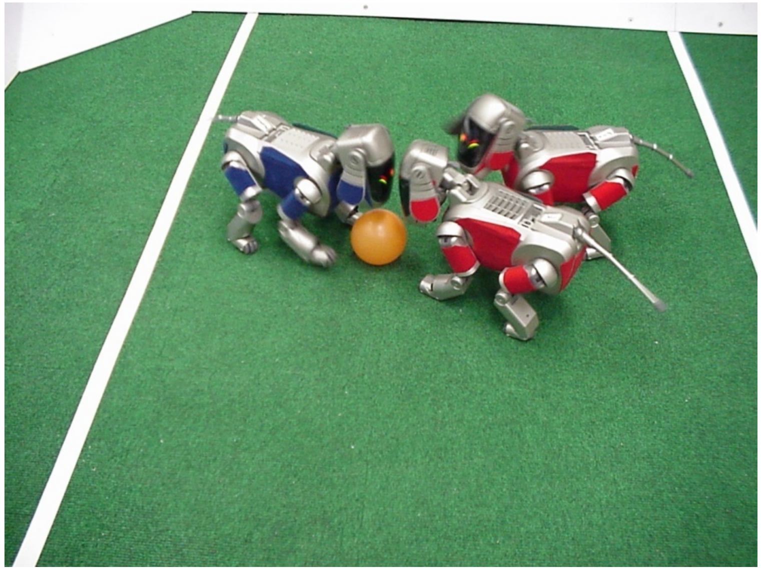
Vision systems (JPL) used for several tasks

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

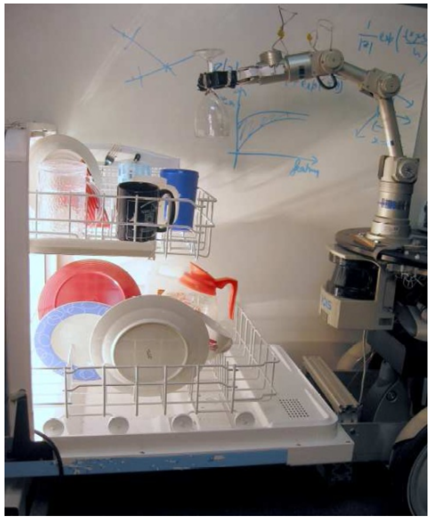
Mobile robots



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



http://www.robocup.org/

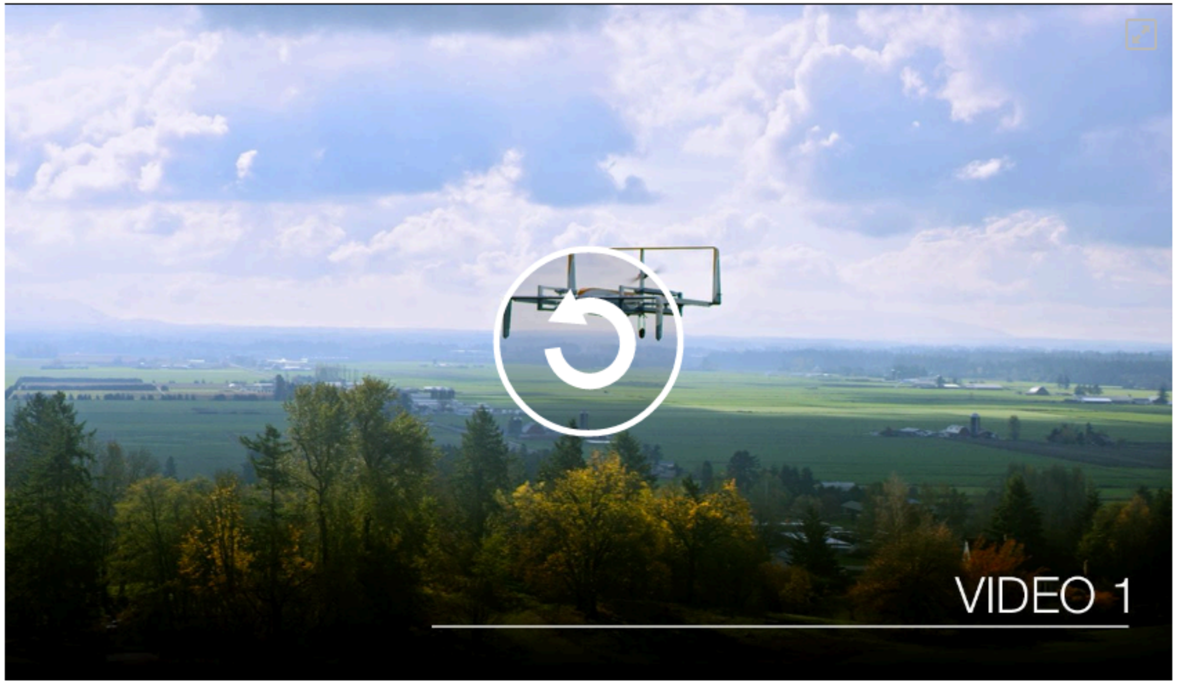


Saxena et al. 2008
STAIR at Stanford

Amazon Prime Air

amazon Prime Air

We're excited about Prime Air — a future delivery system from Amazon designed to safely get packages to customers in 30 minutes or less using small unmanned aerial vehicles, also called drones. Prime Air has great potential to enhance the services we already provide to millions of customers by providing rapid parcel delivery that will also increase the overall safety and efficiency of the transportation system. Putting Prime Air into service will take some time, but we will deploy when we have the regulatory support needed to realize our vision.



https://www.amazon.com/b?node=8037720011

Skydio

https://www.skydio.com/

Augmented Reality and Virtual Reality

Magic Leap, Oculus, Hololens, etc.

State of the art today?

With enough training data, computer vision nearly matches human vision at some recognition tasks

Deep convolutional neural networks have been a disruption to the field. More and more techniques are being “deepified”.

Major research challenges, however, remain.

Computer Vision and Nearby Fields

- Computer Graphics: Models to Images
- Comp. Photography: Images to Images
- Computer Vision: Images to Models

Steve Sullivan

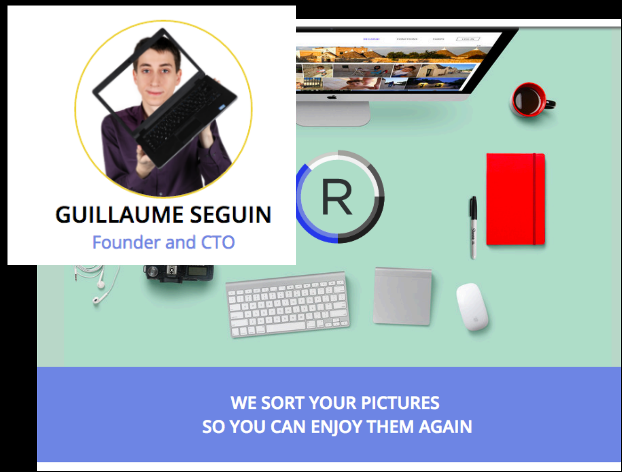
- Ph.D., UIUC, 1996
- Head of R&D, ILM, 2003
- Cover, IEEE Spectrum, 2004
- CSO, Lucasfilm, 2009-2012
- Microsoft (2013-)
- 3 Academy Awards

Computer vision as a job

Vincent Delaitre
Phd, 2015
Start-up Deepomatic.com

https://www.inria.fr/centre/paris/actualites/deepomatic-un-shazam-de-l-image-pour-le-e-commerce-et-les-medias

Guillaume Seguin
Phd, 2016
Start-up regaind.io

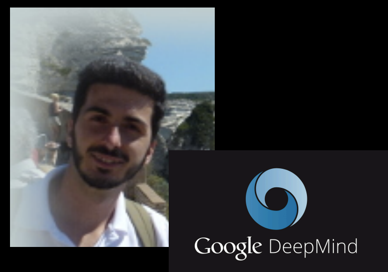


Piotr Bojanowski
Phd, 2016
Facebook AI Research

Oliver Whyte
Phd, 2012
Engineer at Microsoft

Mathieu Aubry
Phd, 2015
Faculty at ENPC

Relja Arandjelovic
Post-doc, 2016
Google DeepMind



Computer vision books

- D.A. Forsyth and J. Ponce, "Computer Vision: A Modern Approach", Prentice-Hall, 2003, 2nd edition, 2011.
- R. Szeliski, "Computer Vision: Algorithms and Applications", Springer, 2010.
- O. Faugeras, Q.T. Luong, and T. Papadopoulo, "Geometry of Multiple Images," MIT Press, 2001.
- R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2004.

Other relevant books

- J.J. Koenderink, “Solid Shape”, MIT Press, 1990.
- J.J. Koenderink, http://www.gestaltrevision.be/en/resources/clootcrans-press

- M. Berger, “Géométrie”, Nathan, 1992.
- D. Hilbert and S. Cohn-Vossen, “Geometry and the Imagination”, Chelsea, 1952.

Course outline:

1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

Reconnaissance d'objets et vision artificielle

http://www.di.ens.fr/willow/teaching/recvis16/

Ivan Laptev

Josef Sivic

Cordelia Schmid

Jean Ponce

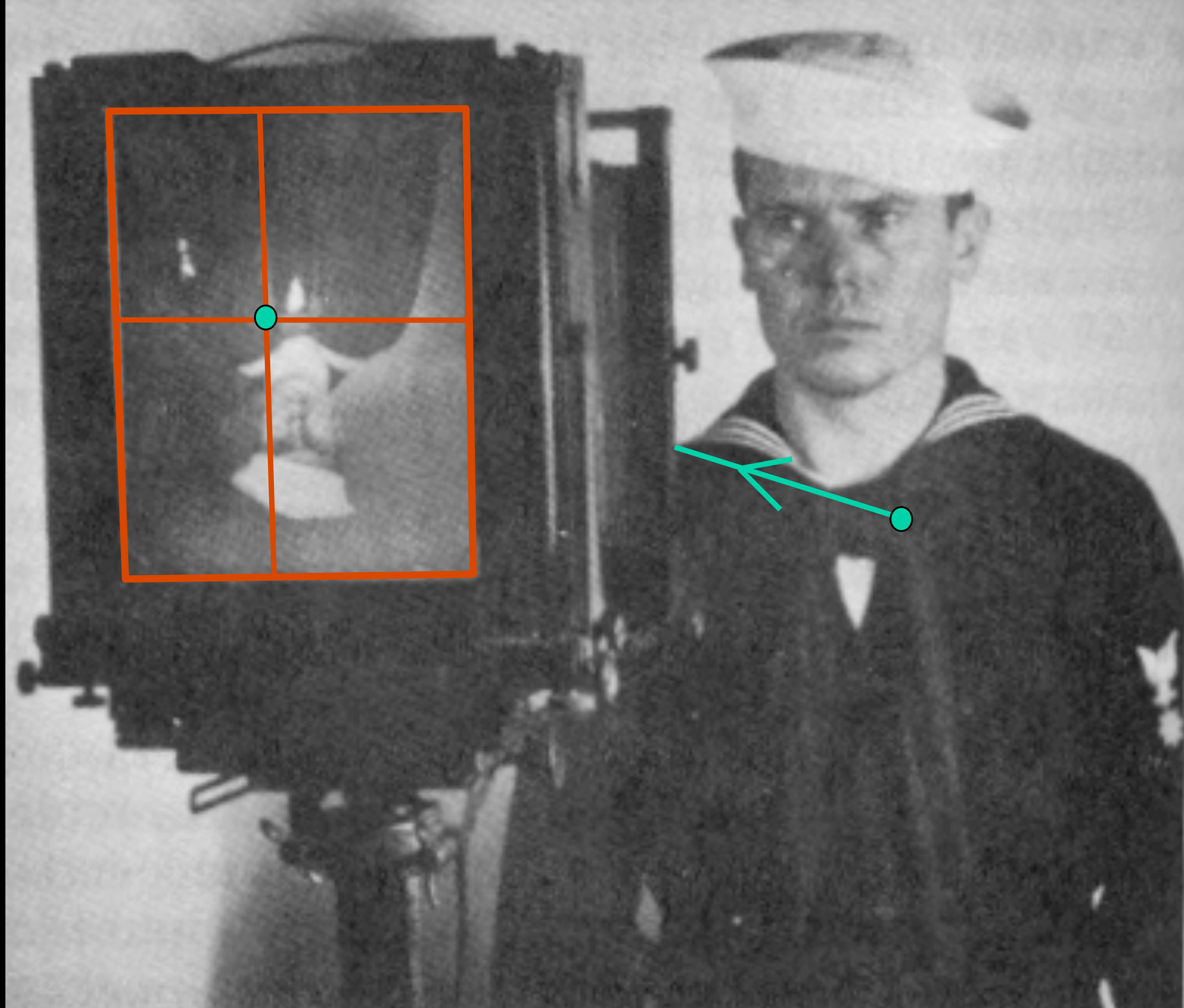
Tuesdays 16h15 salle Conference, at 46 rue d'Ulm
First class: Tuesday Oct 4th

Course outline:

1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

They are formed by the projection of 3D objects.



Images are two-dimensional patterns of brightness values.

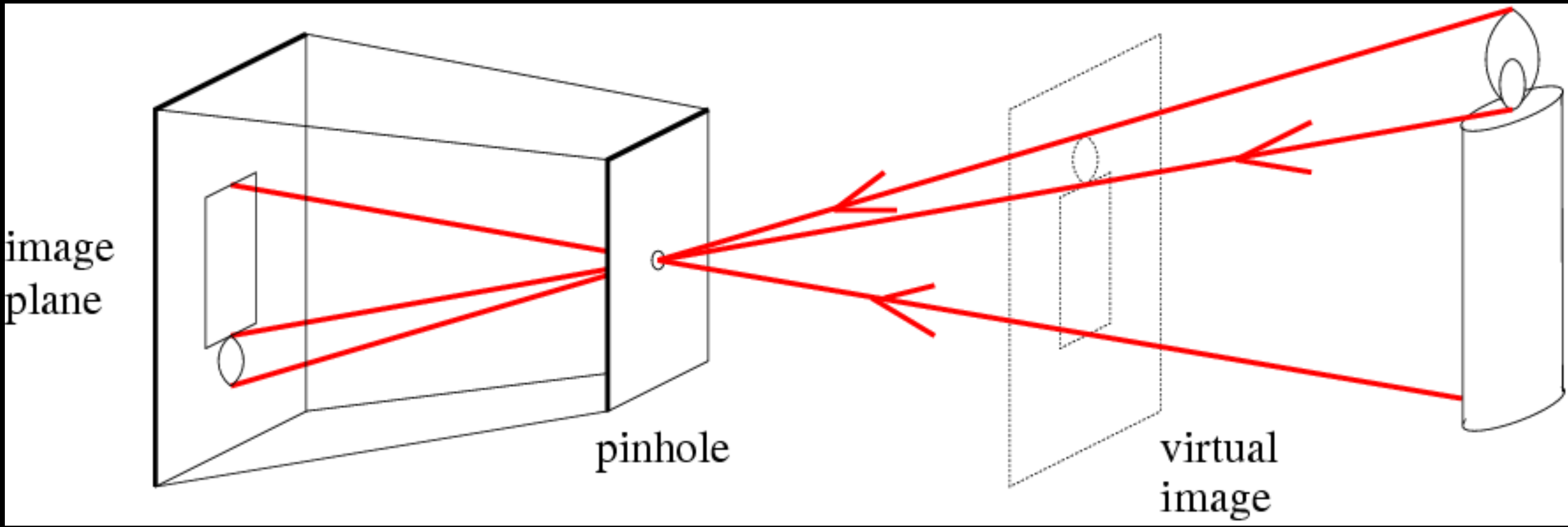
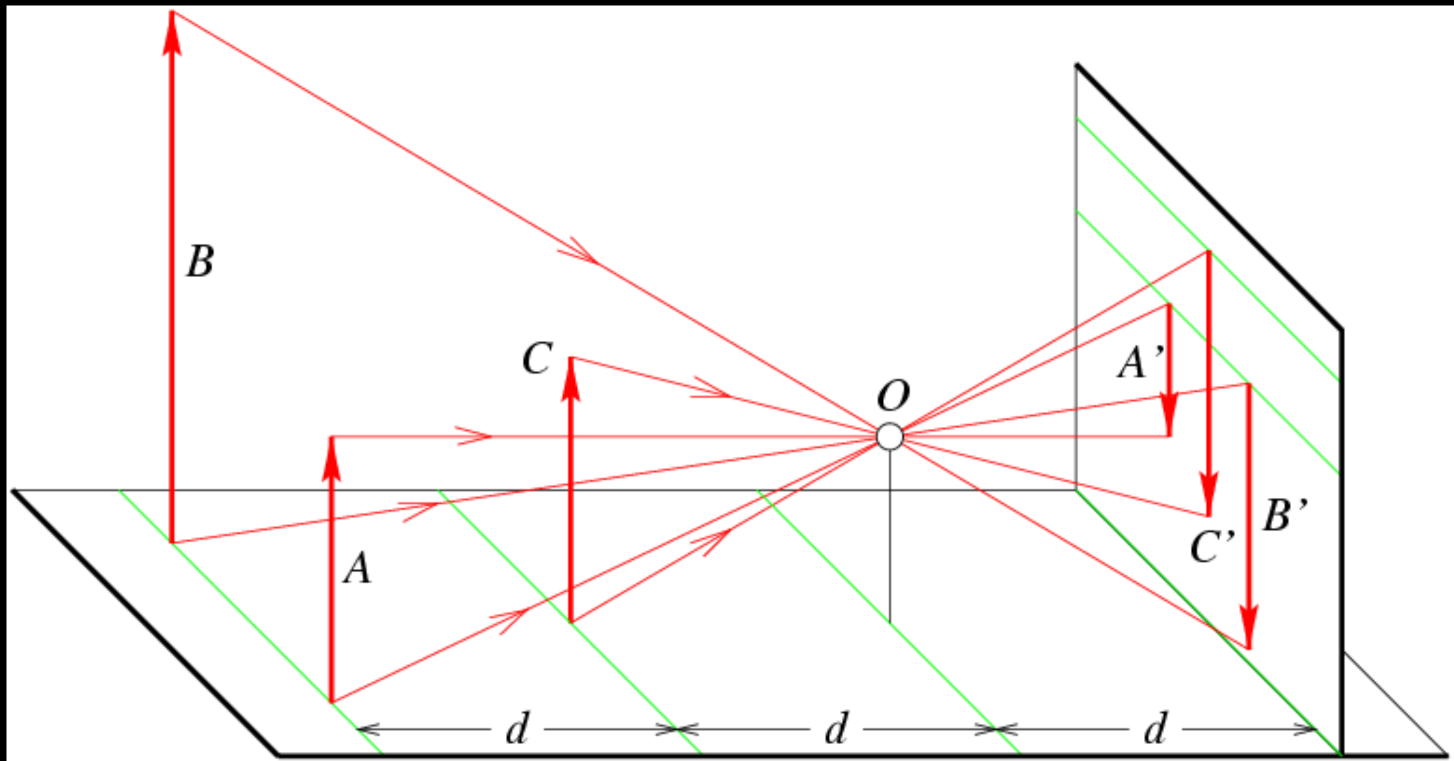
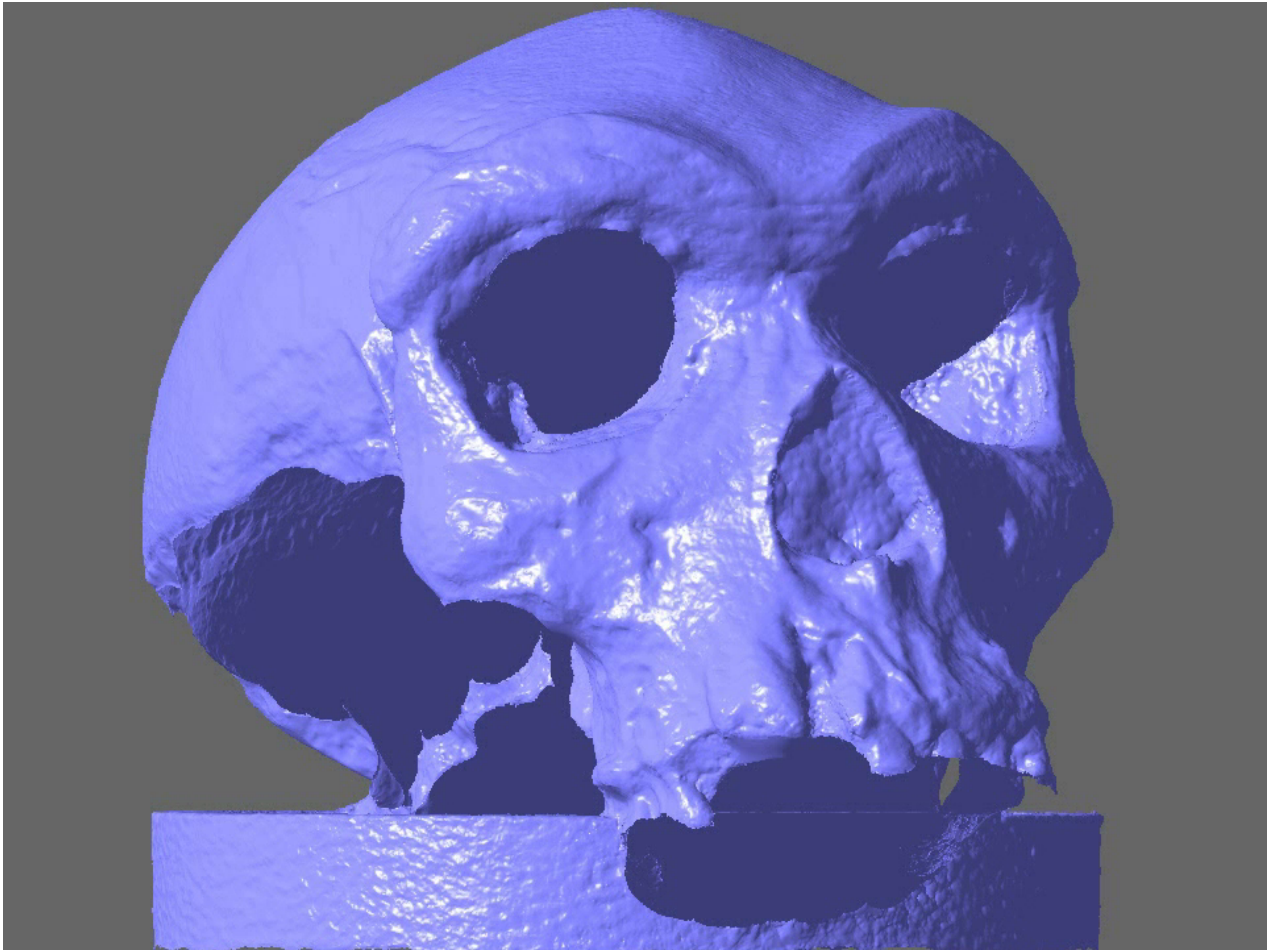


image
plane
pinhole
virtual
image

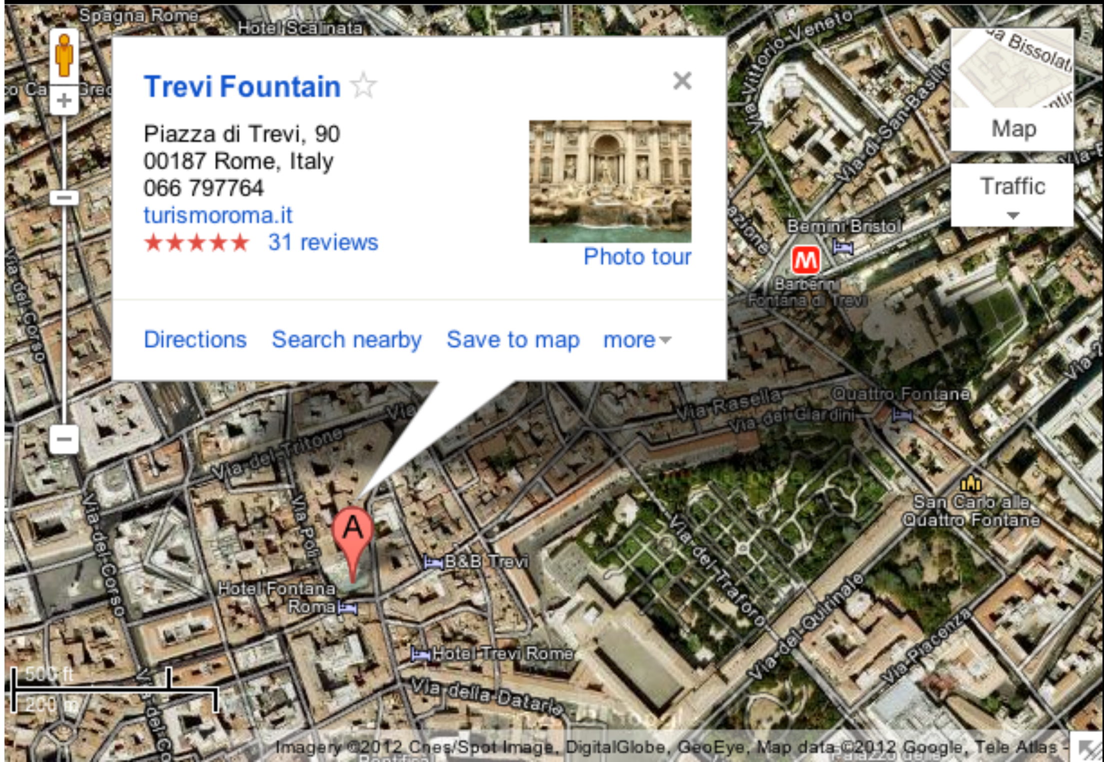

B
C
O
A'
B'
C'
A
d
d
d

High-fidelity multi-view stereopsis
(Furukawa and Ponce, CVPR'07,PAMI'10)
http://www.cs.washington.edu/homes/furukawa/research/pmvs/index.html

Data courtesy of S. Leigh, UIUC Anthropology Department. See for example (Hernandez and Schmitt, 2004; Strecha et al., 2006) for related work.



PMVS (http://www.di.ens.fr/pmvs)




(© Bath & Burke, Weta Digital, Siggraph'11)

- Google Maps Photo Tour
- Lucasfilm
- Weta Digital

Course outline:

1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

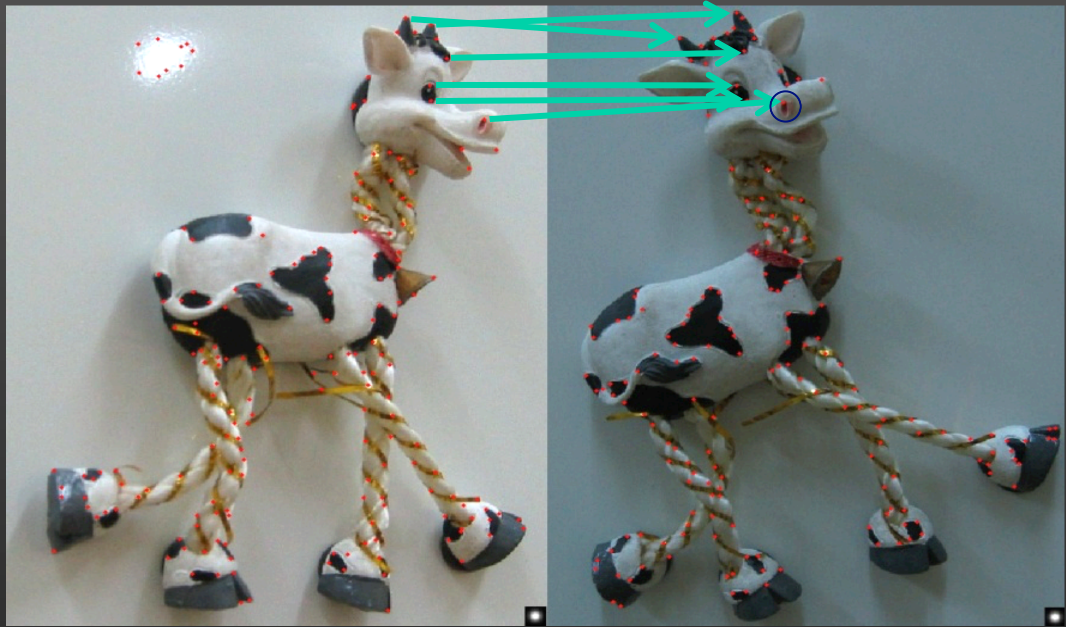
Programming assignments + final presentation

Filtering

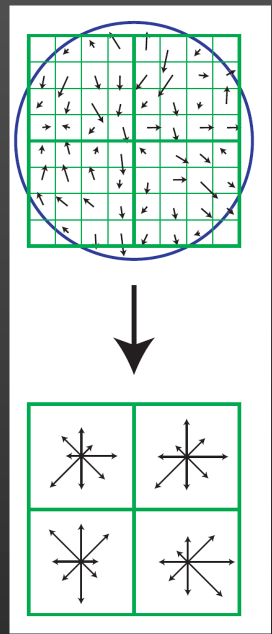
Edge Detection

Edge Detection

Interest points and local appearance models



(Image courtesy of C. Schmid)



(Lowe 2004)

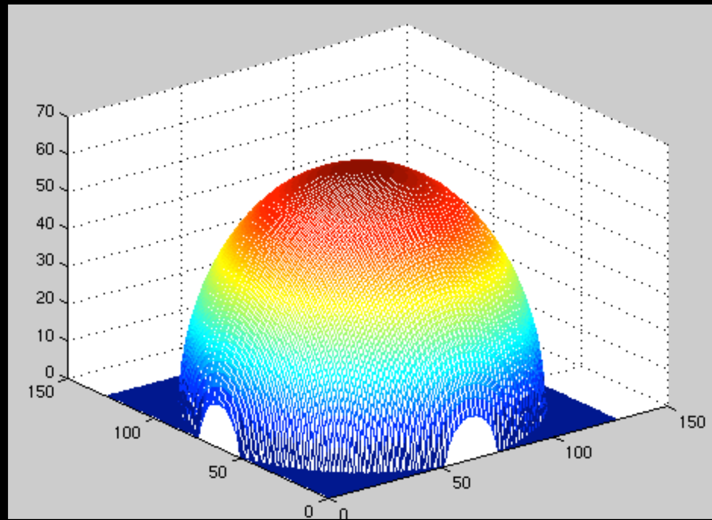
- Find features (interest points)
- Match them using local invariant descriptors (jets, SIFT)

Course outline:

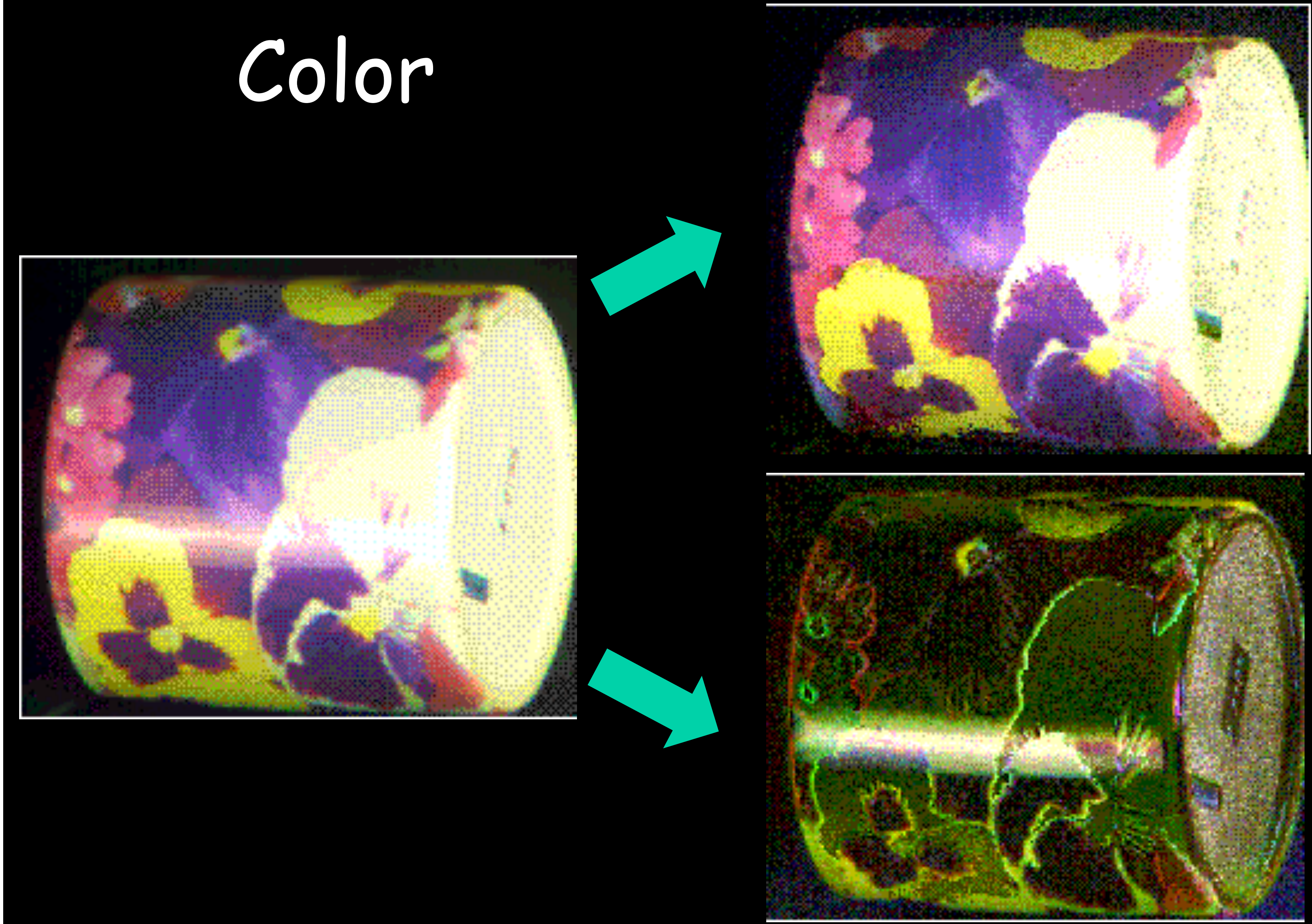
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

Radiometry/Shading



Color

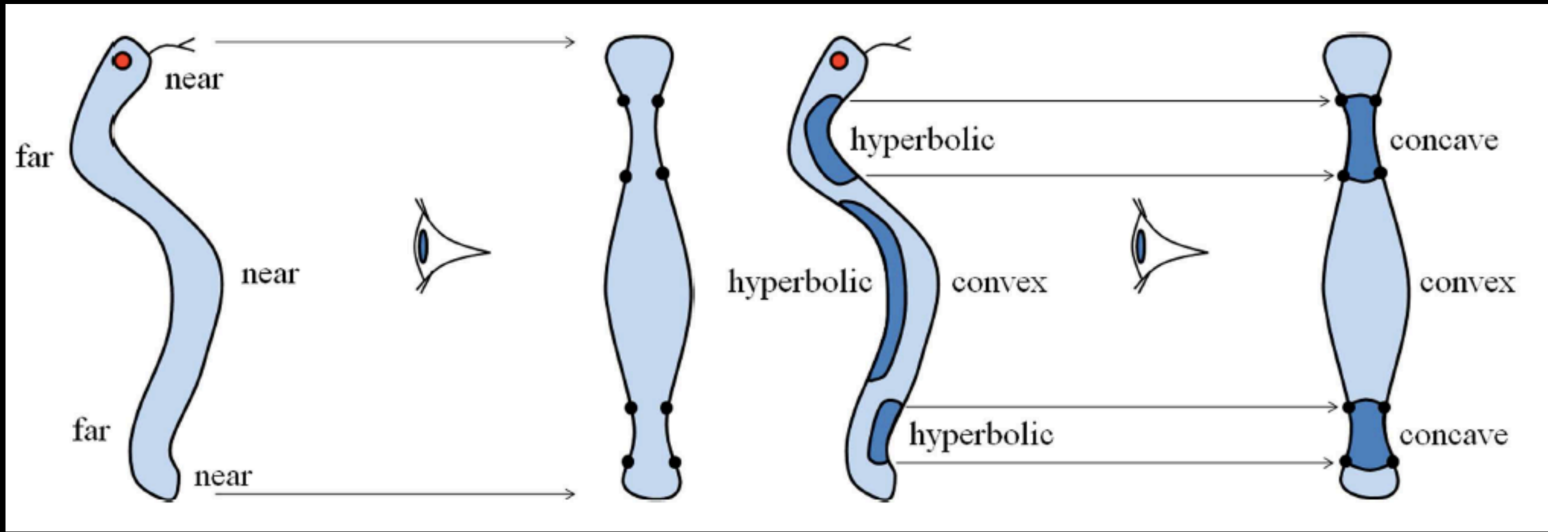


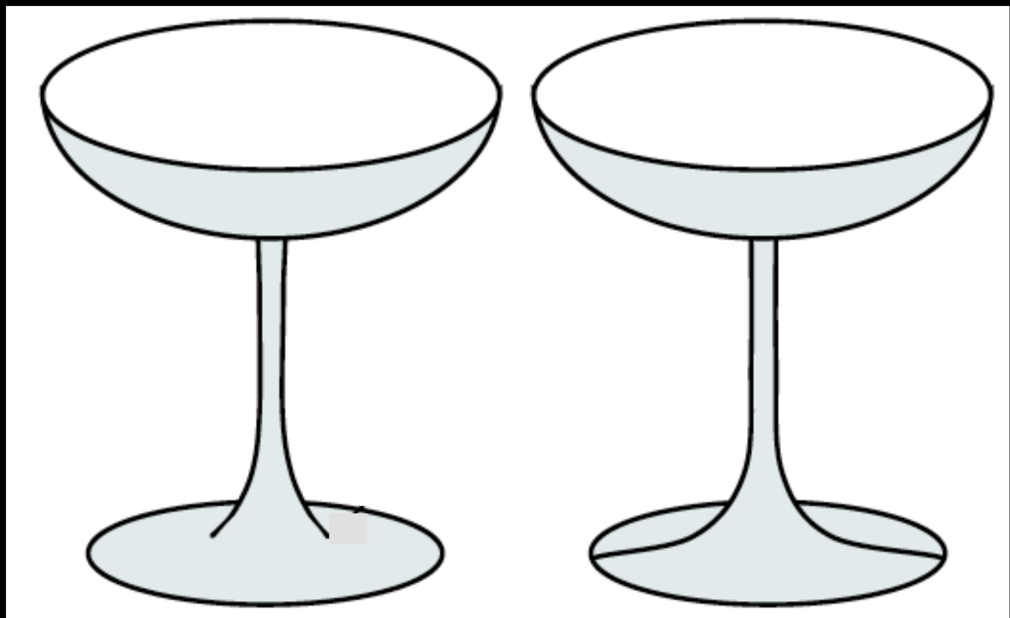
Course outline:

1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

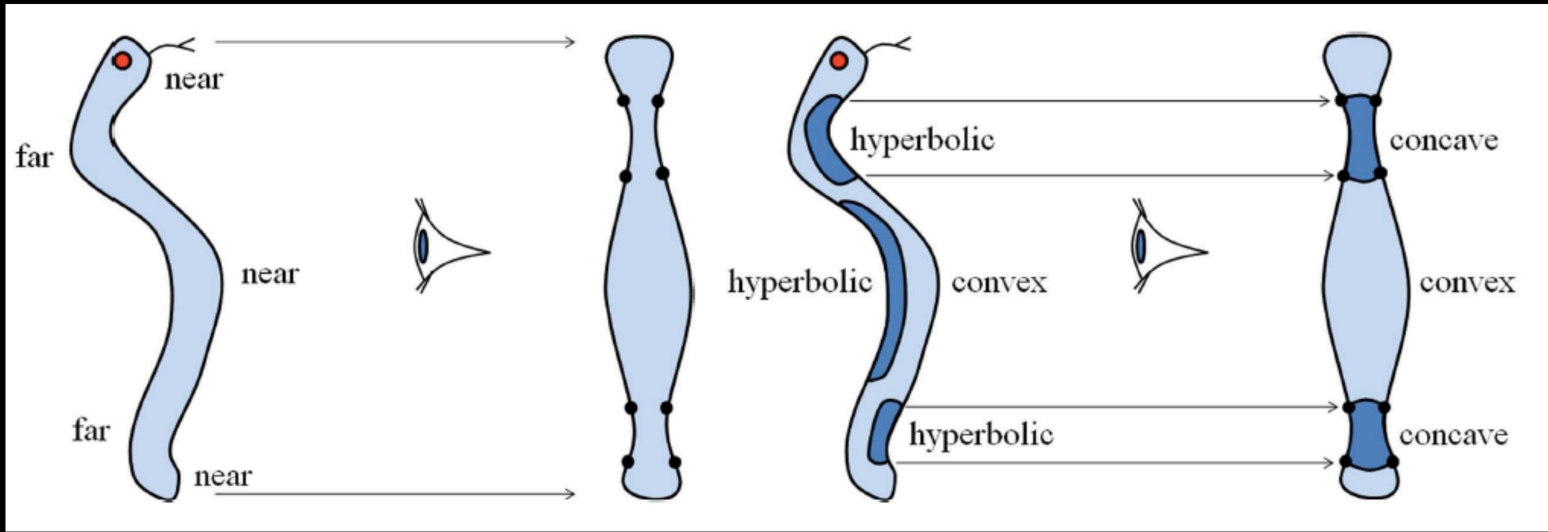
One-view (differential) geometry

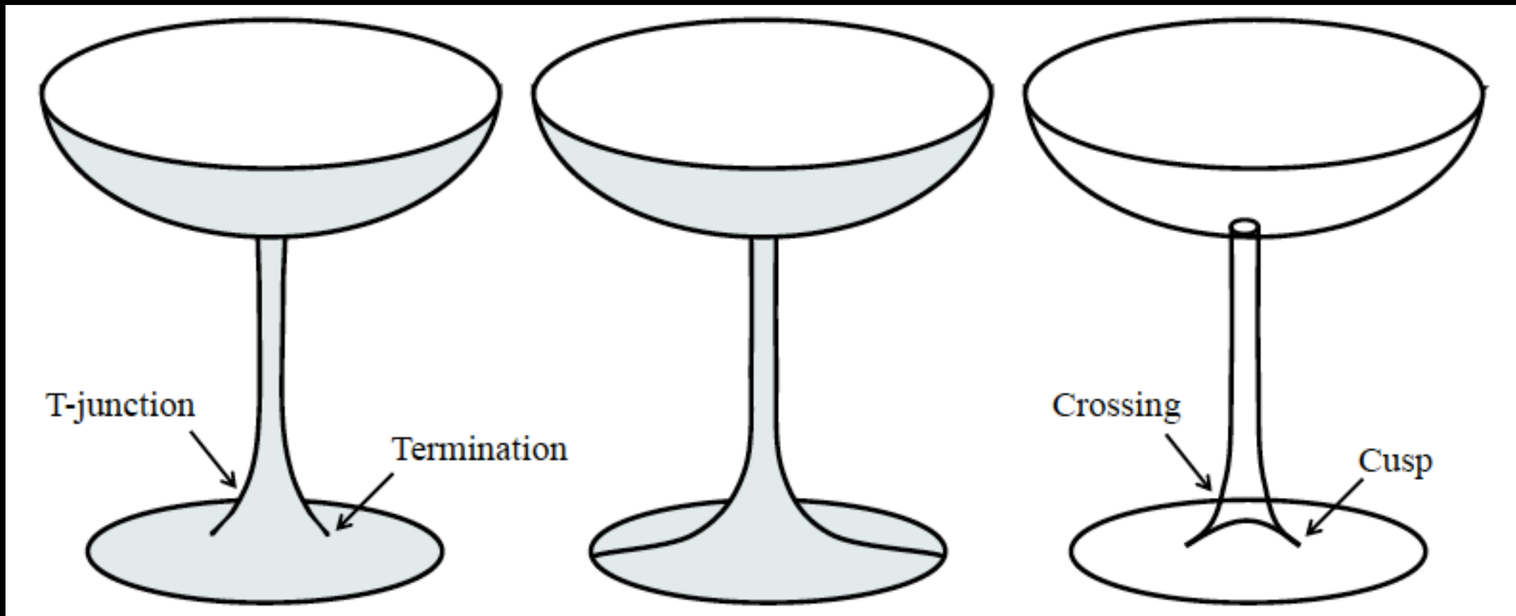




(Marr & Nishihara, 1978; Koenderink, 1984)

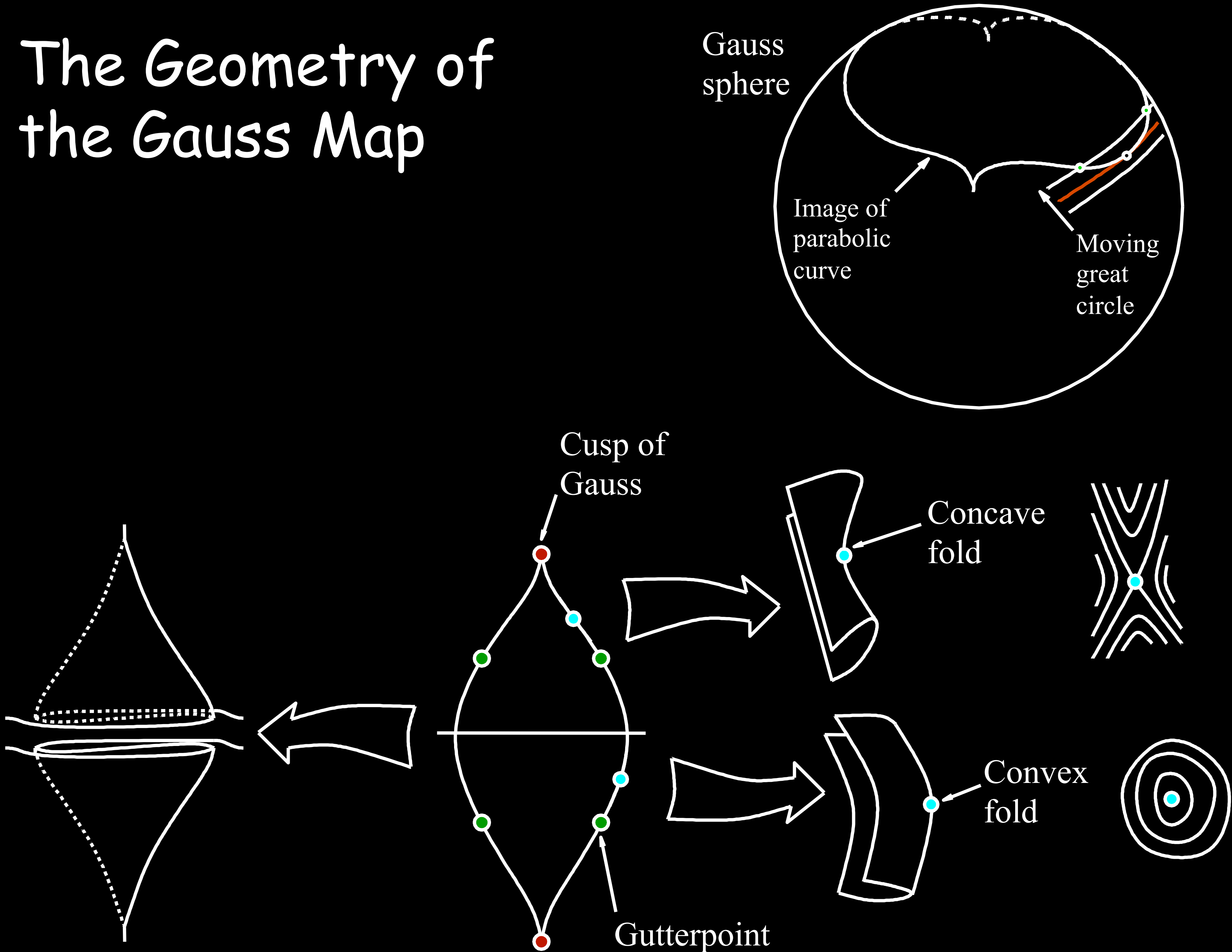
One-view (differential) geometry





(Marr & Nishihara, 1978; Koenderink, 1984)

The Geometry of the Gauss Map

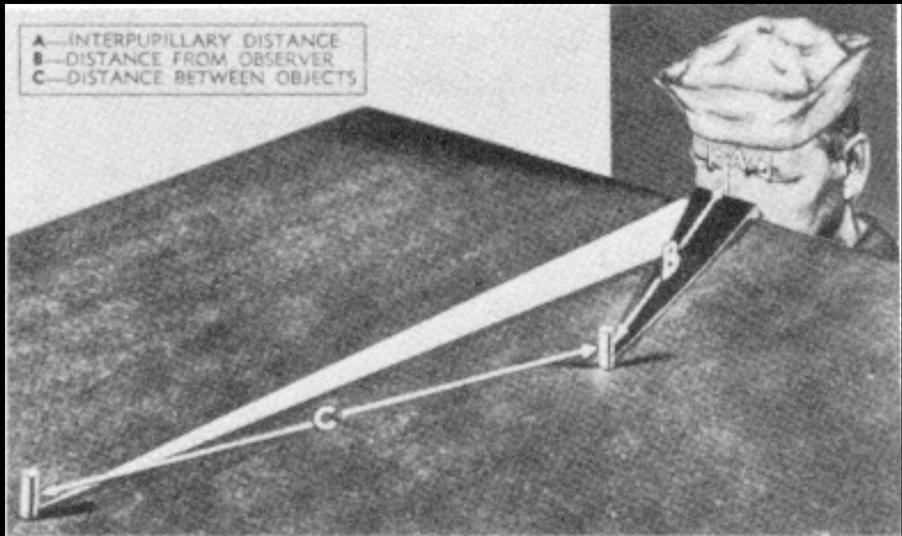


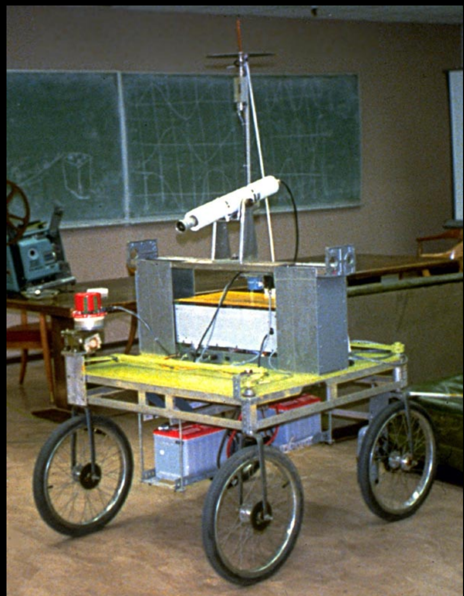
Course outline:

1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

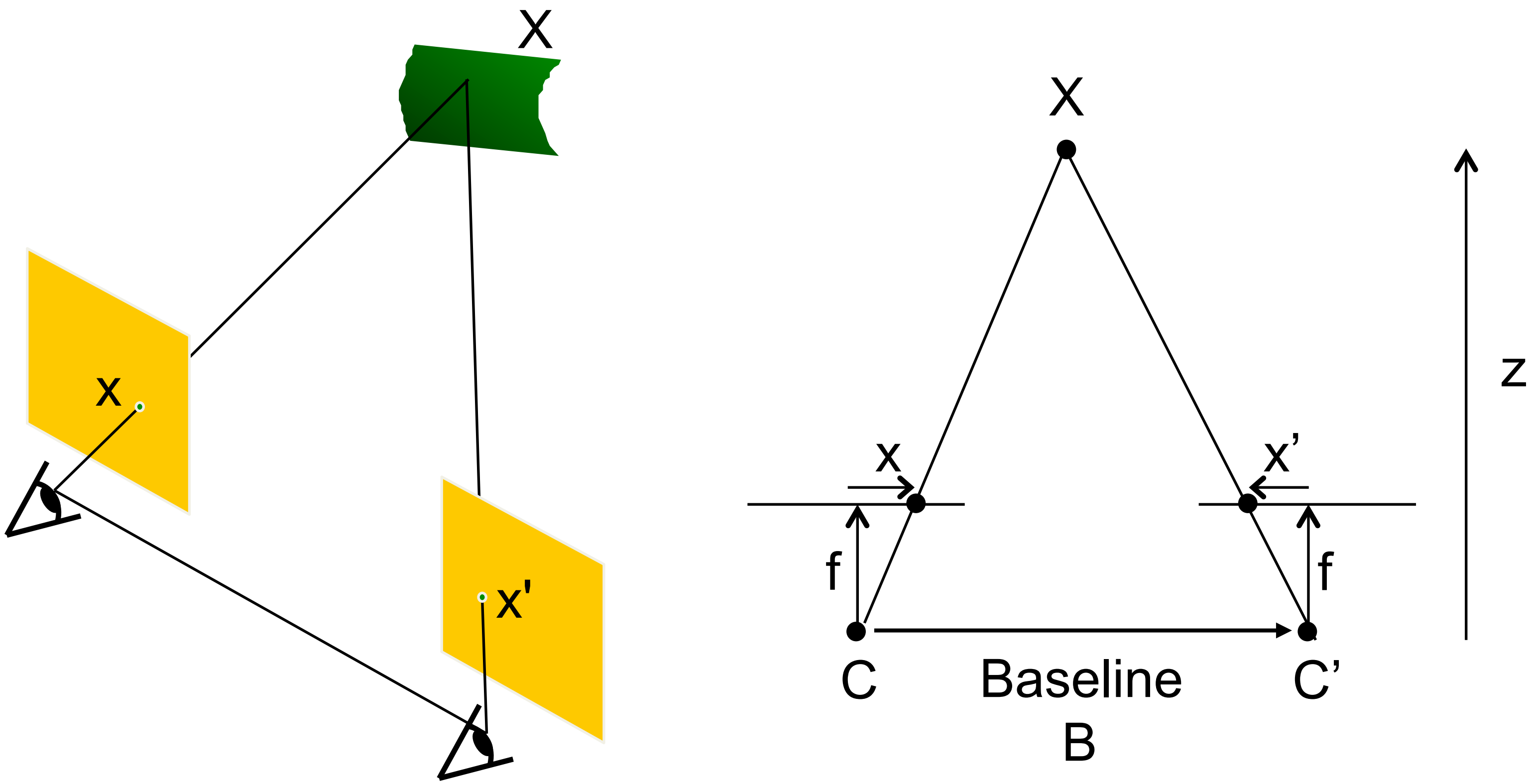
How do we perceive depth?





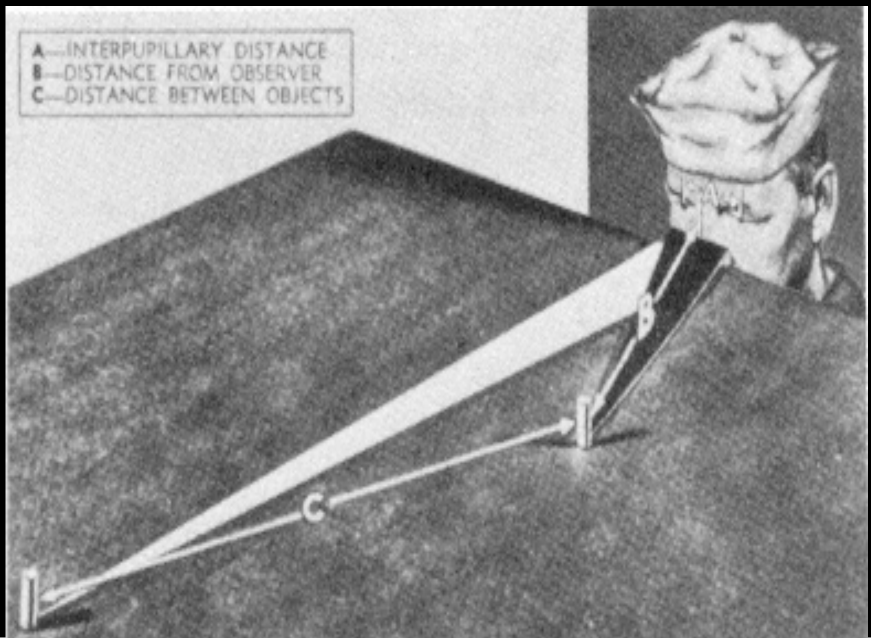
Depth from Stereo

- Goal: recover depth by finding image coordinate x’ that corresponds to x

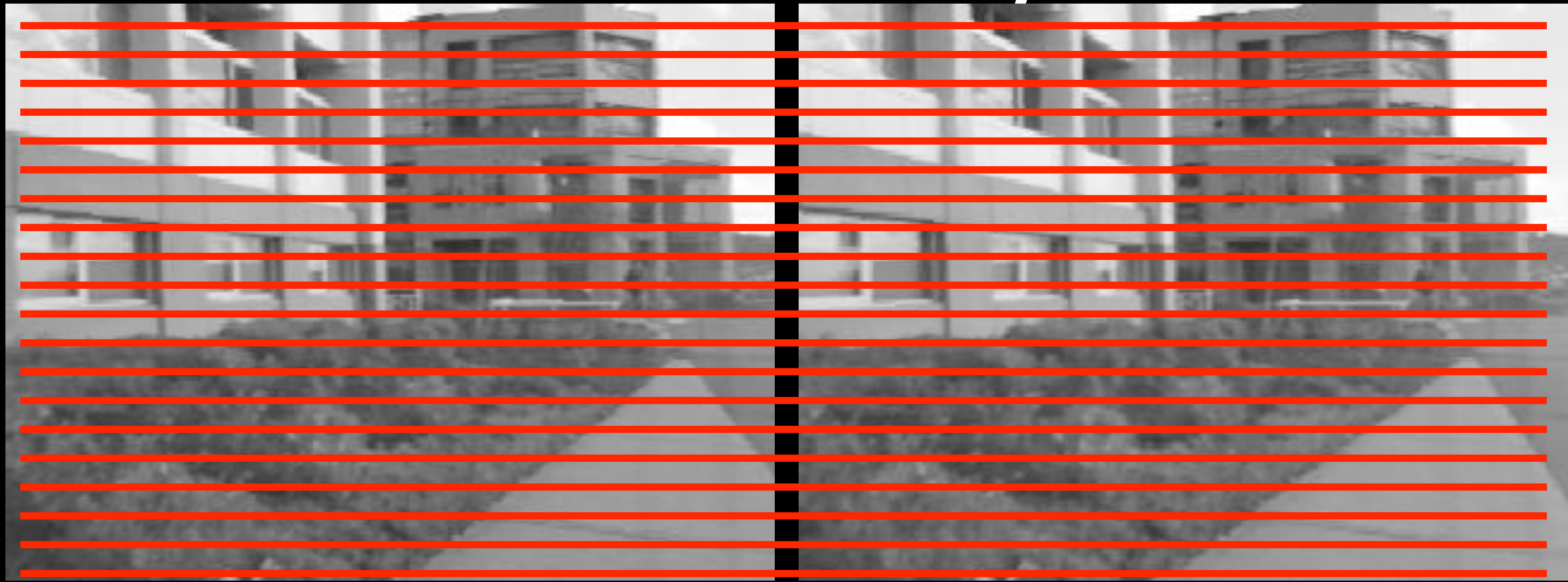


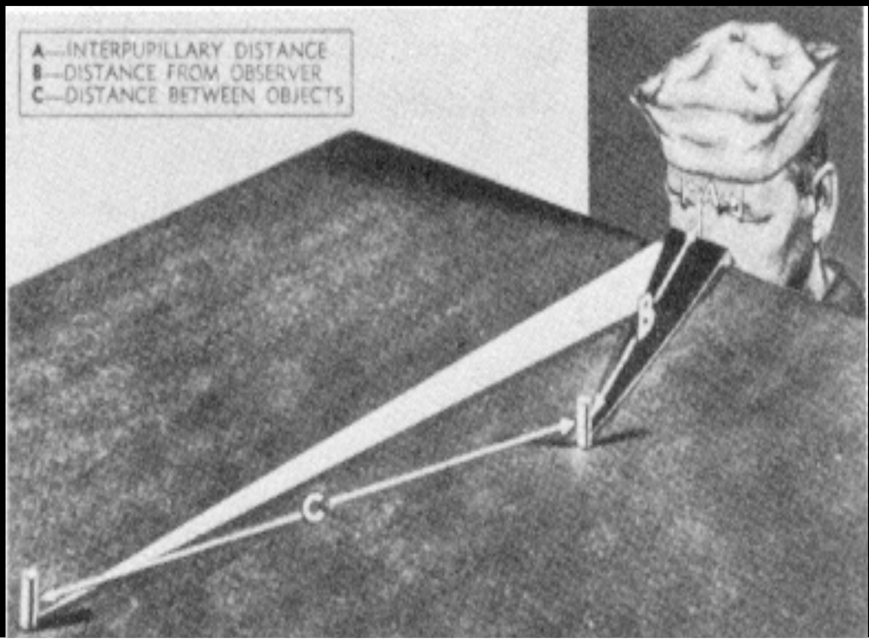
Slide: D. Hoiem

Two-View Geometry: Stereo

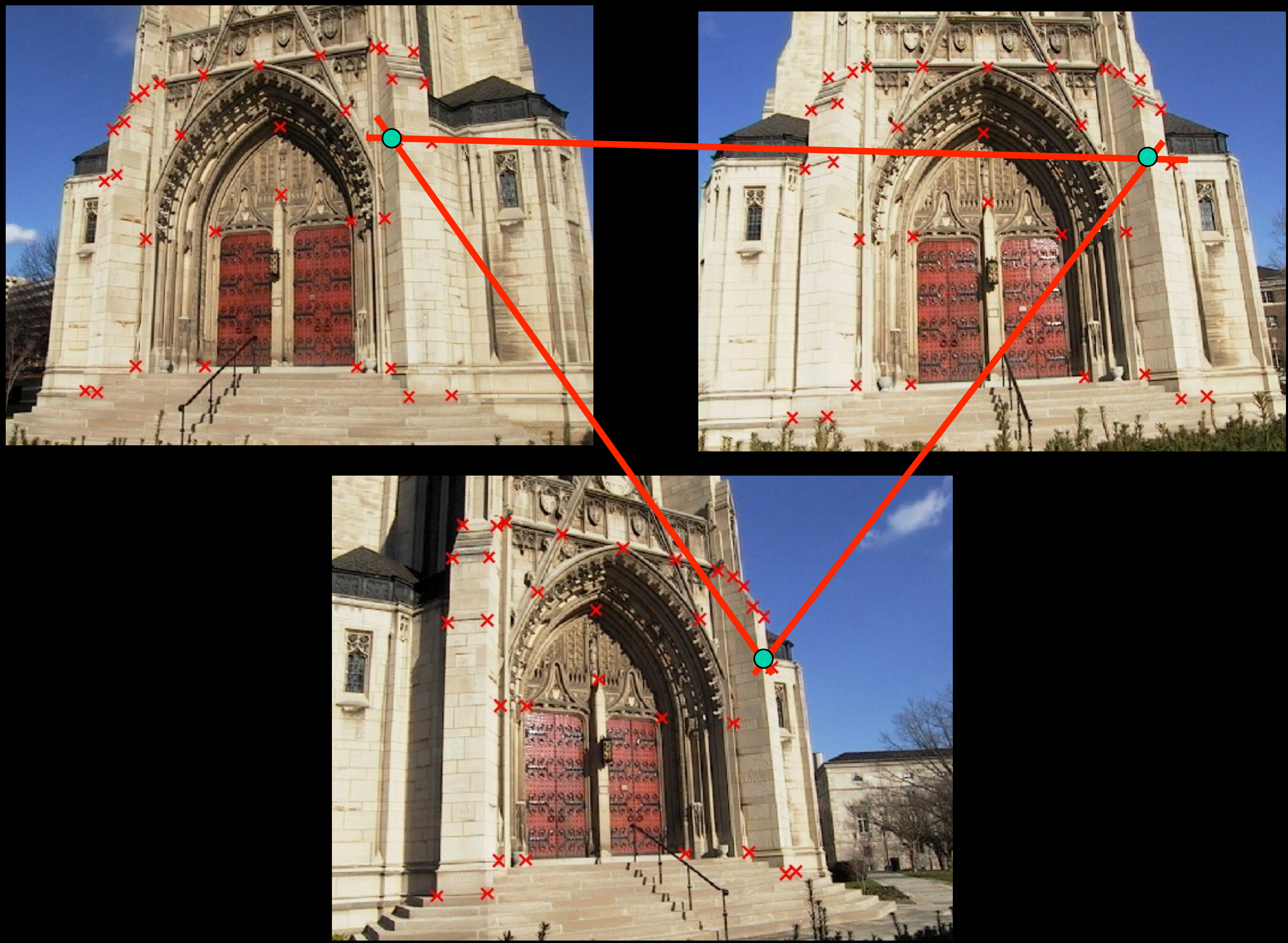


Two-View Geometry: Stereo



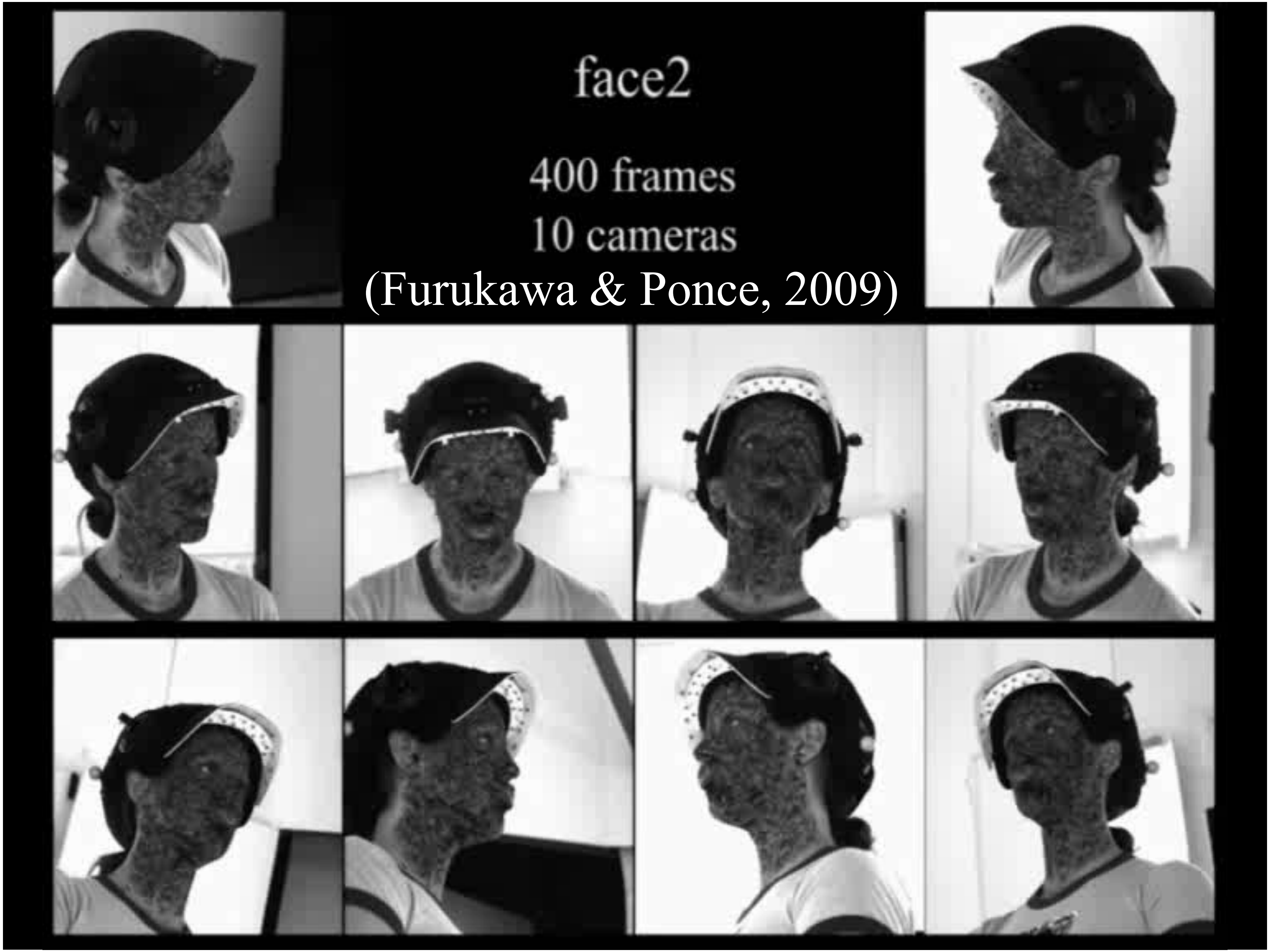


Multi-Camera Geometry



Phototourism

(Snavely, Seitz, Szeliski, 2006)
http://phototour.cs.washington.edu/

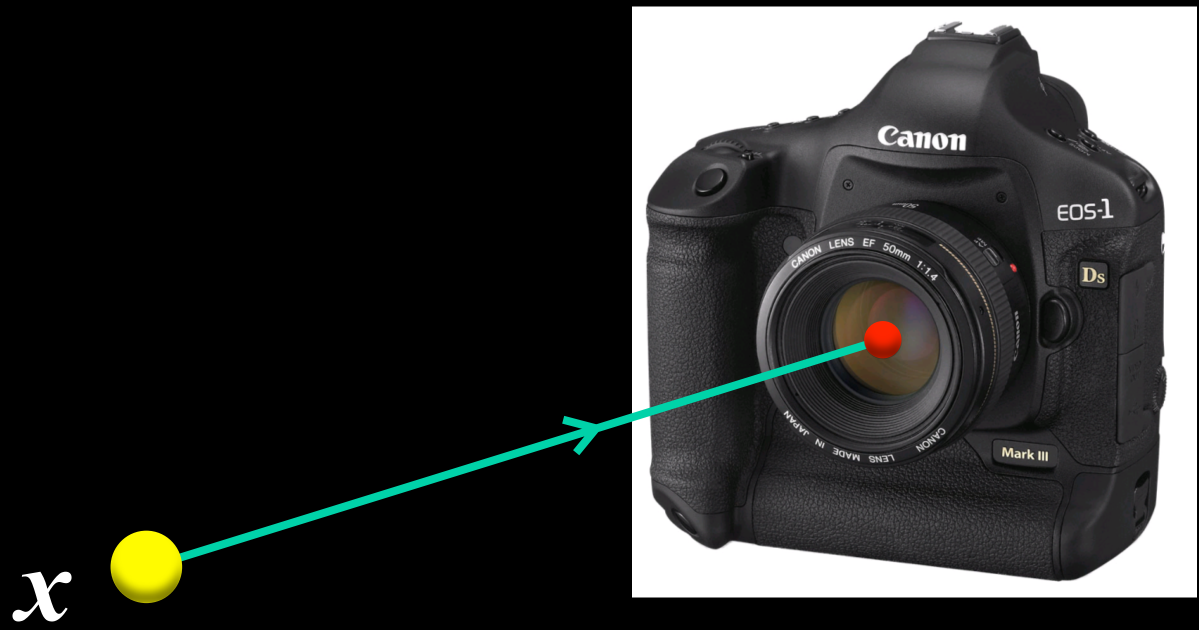

face2
400 frames
10 cameras
(Furukawa & Ponce, 2009)

Course outline:

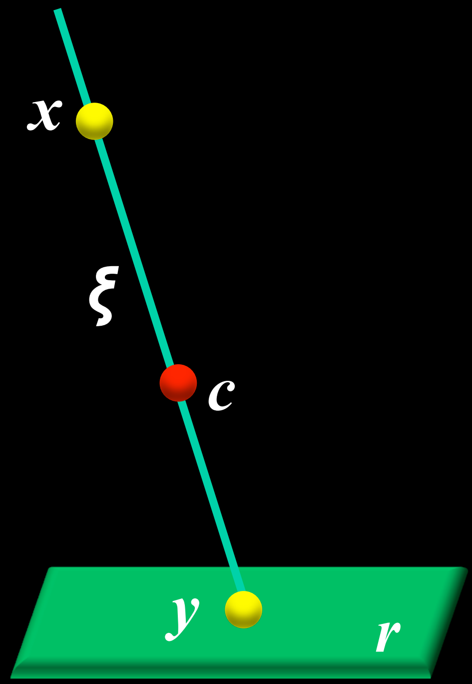
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

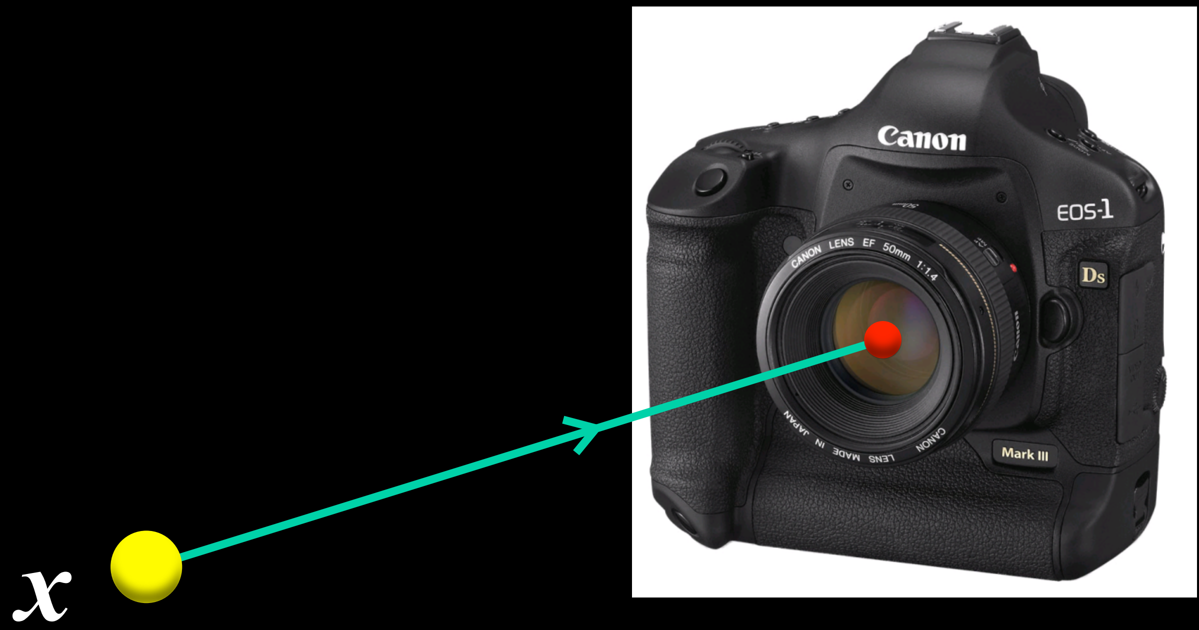
Programming assignments + final presentation

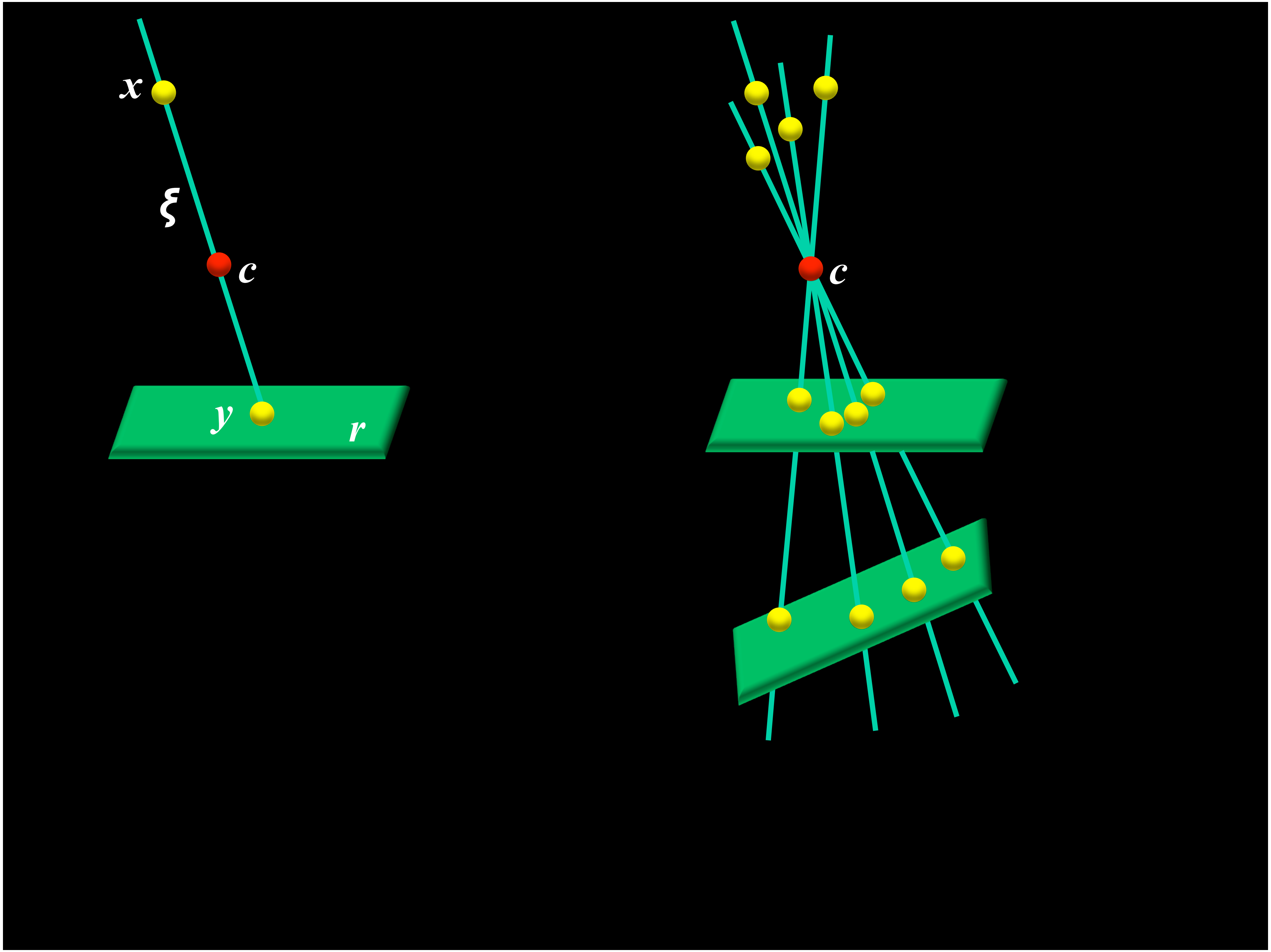
What is a camera?

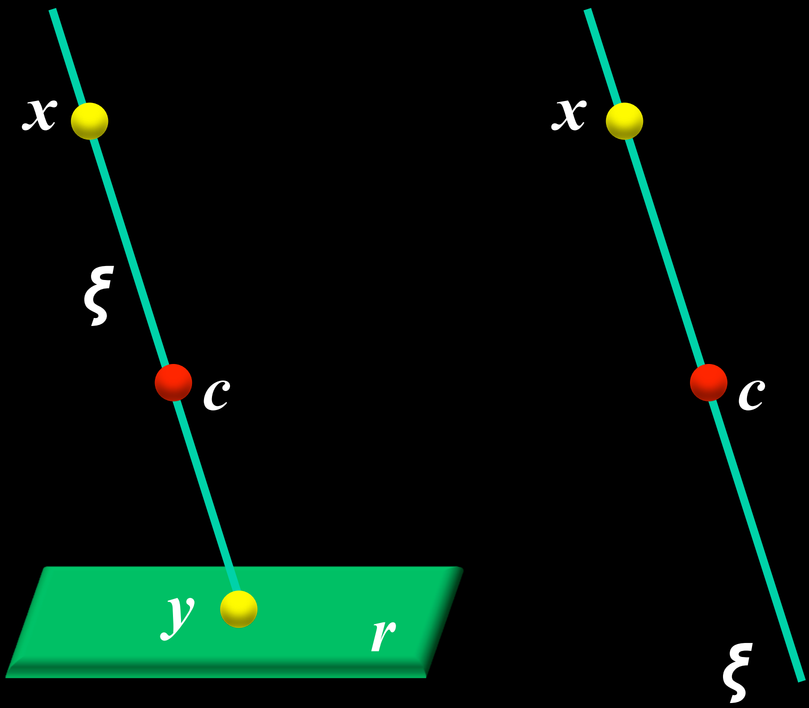


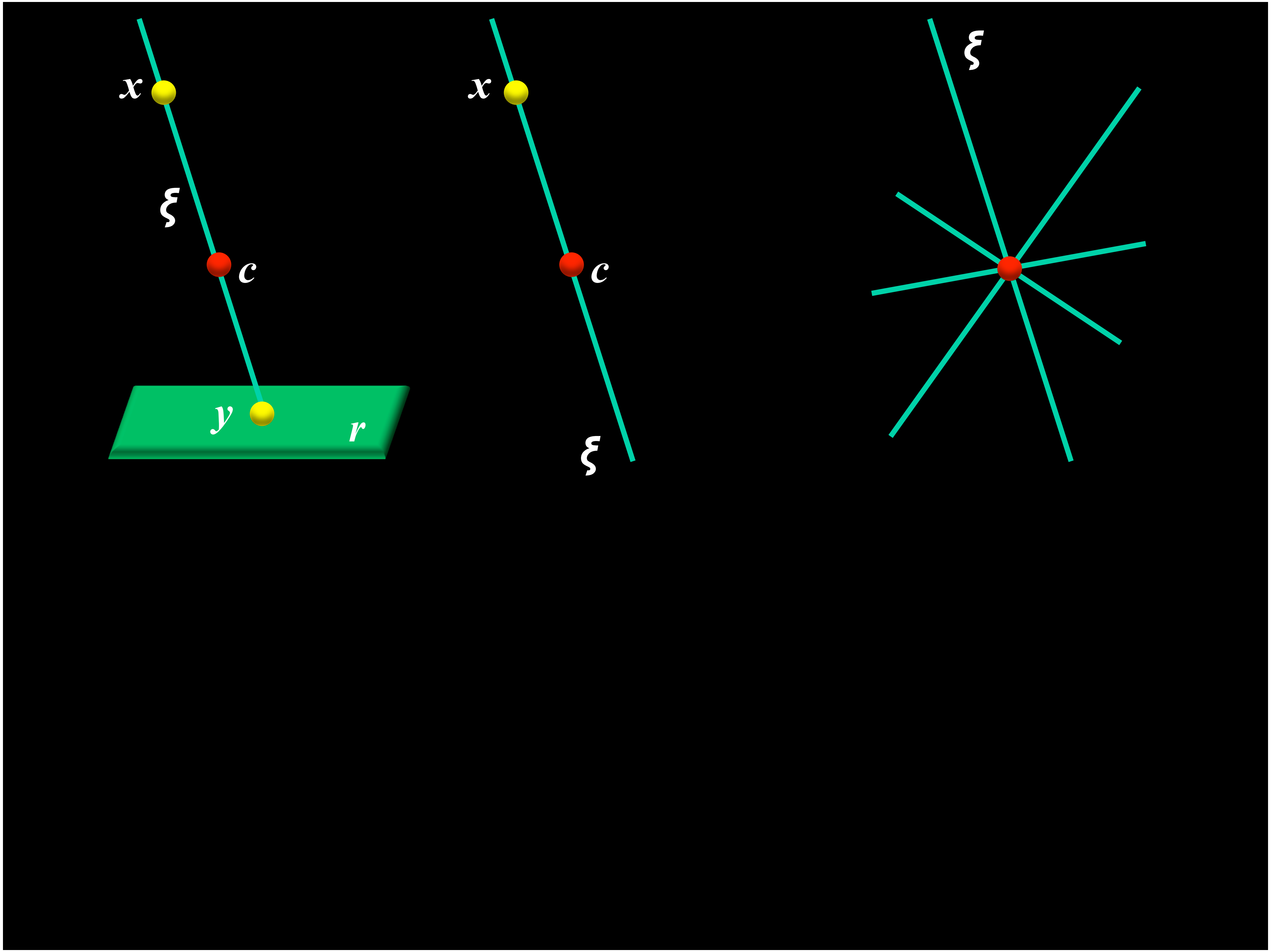
What is a camera?



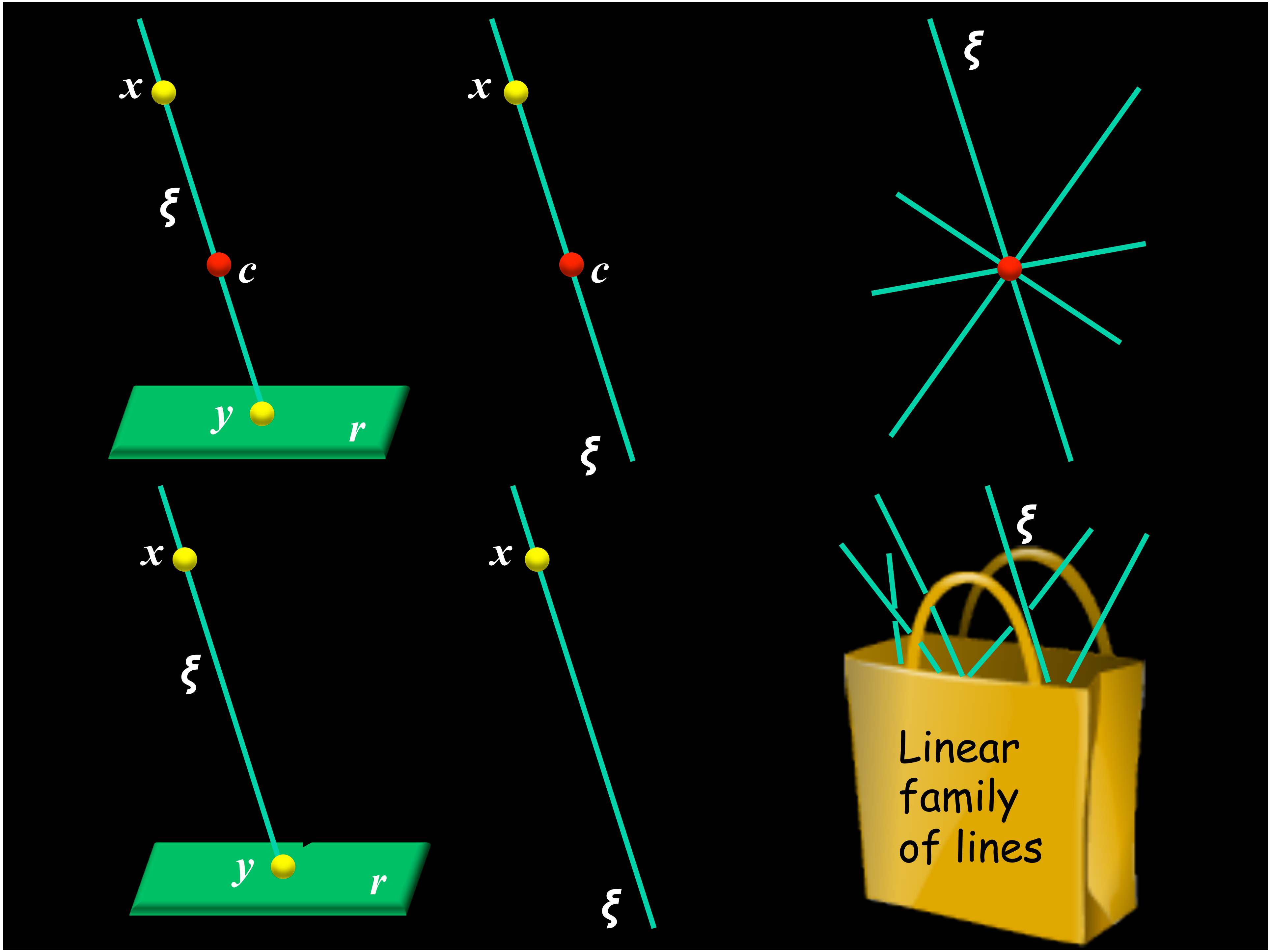



x
ξ
c
y
r
c


x
ξ
c
y
r
x
c
ξ



x
ξ
c
y
r
x
c
ξ
ξ

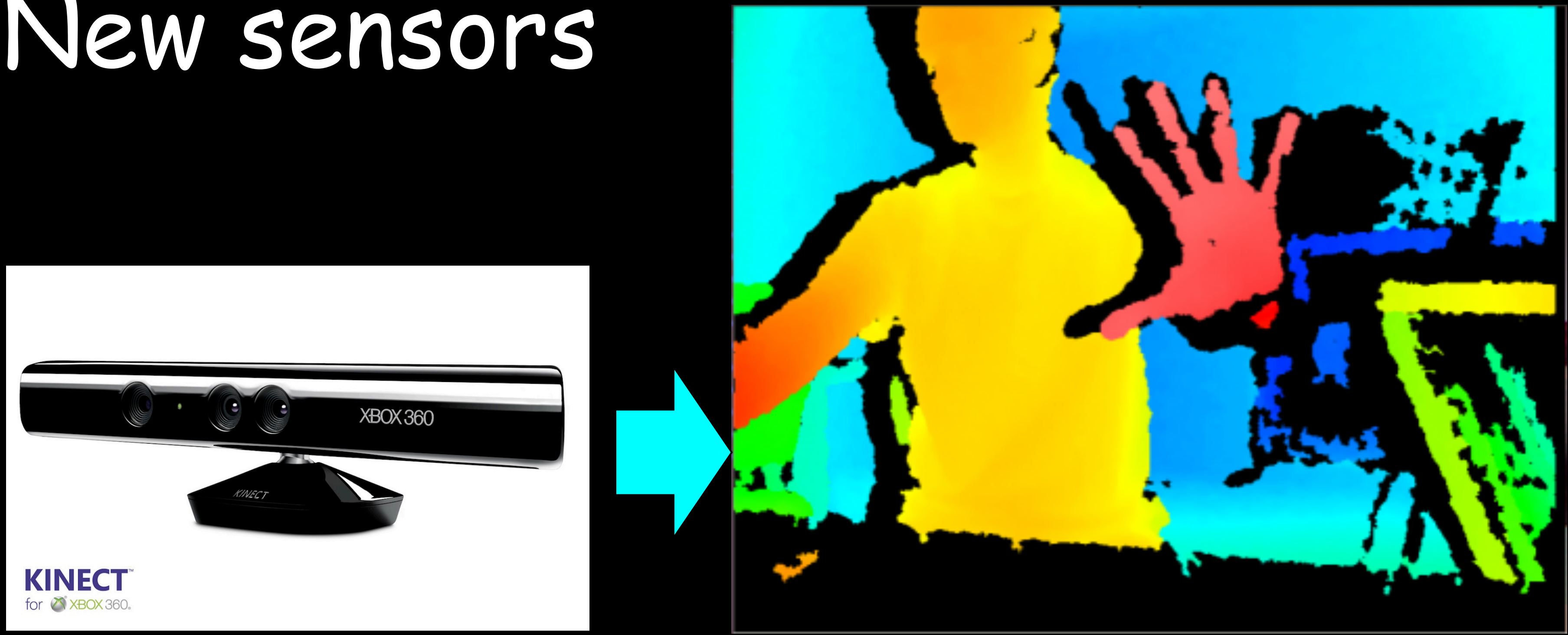

x
ξ
c
y
r
x
c
ξ
ξ
x
ξ
y
r
x
ξ
ξ
Linear family of lines

Course outline:

1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

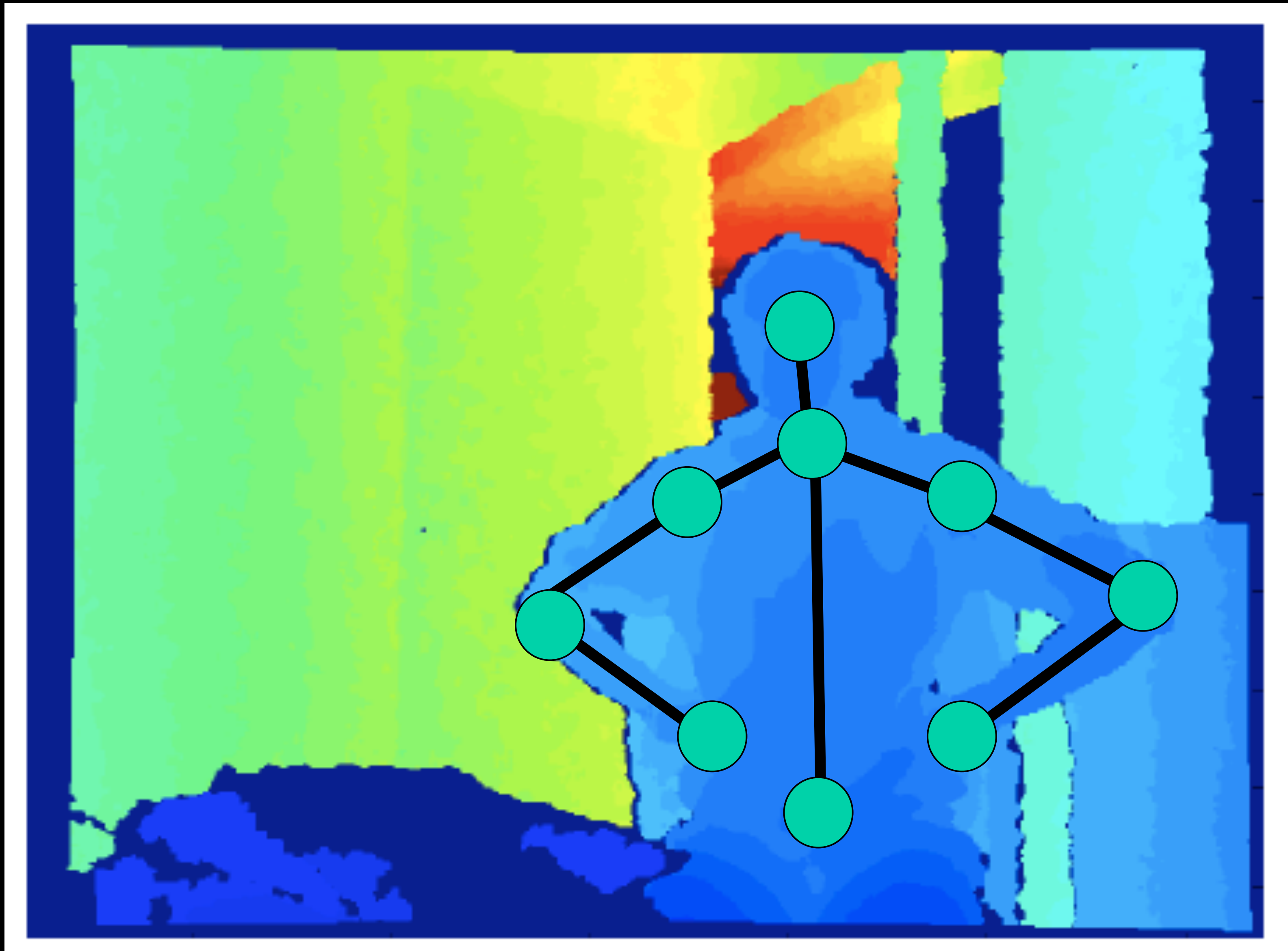
Programming assignments + final presentation

New sensors



Problem: find the 3D skeleton of people

Solution: Use random forest to classify pixels as belonging to some body part



(Shotton et al., 2011)

Course outline:

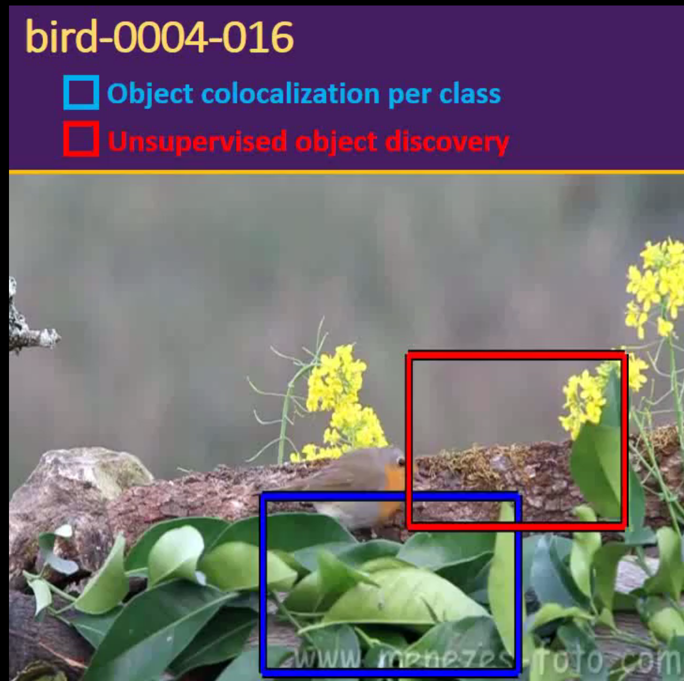
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

Segmentation

(Joulin, Bach, Ponce, CVPR'12)

Unsupervised object discovery



(Suha et al., 2015)

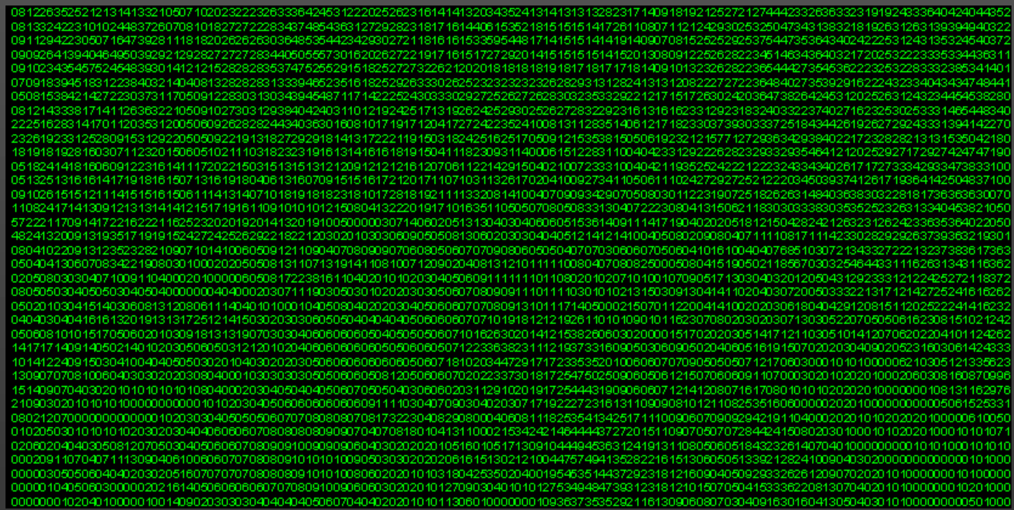
Layered person segmentation [Seguin et al., 2015]

Course outline:

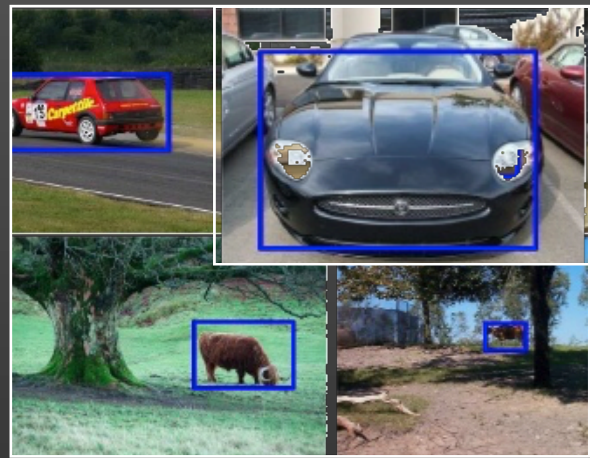
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Projective cameras
8. Range data
9. Segmentation
10. Recognition

Programming assignments + final presentation

How to make sense of "pixel-chaos"?



Object class recognition

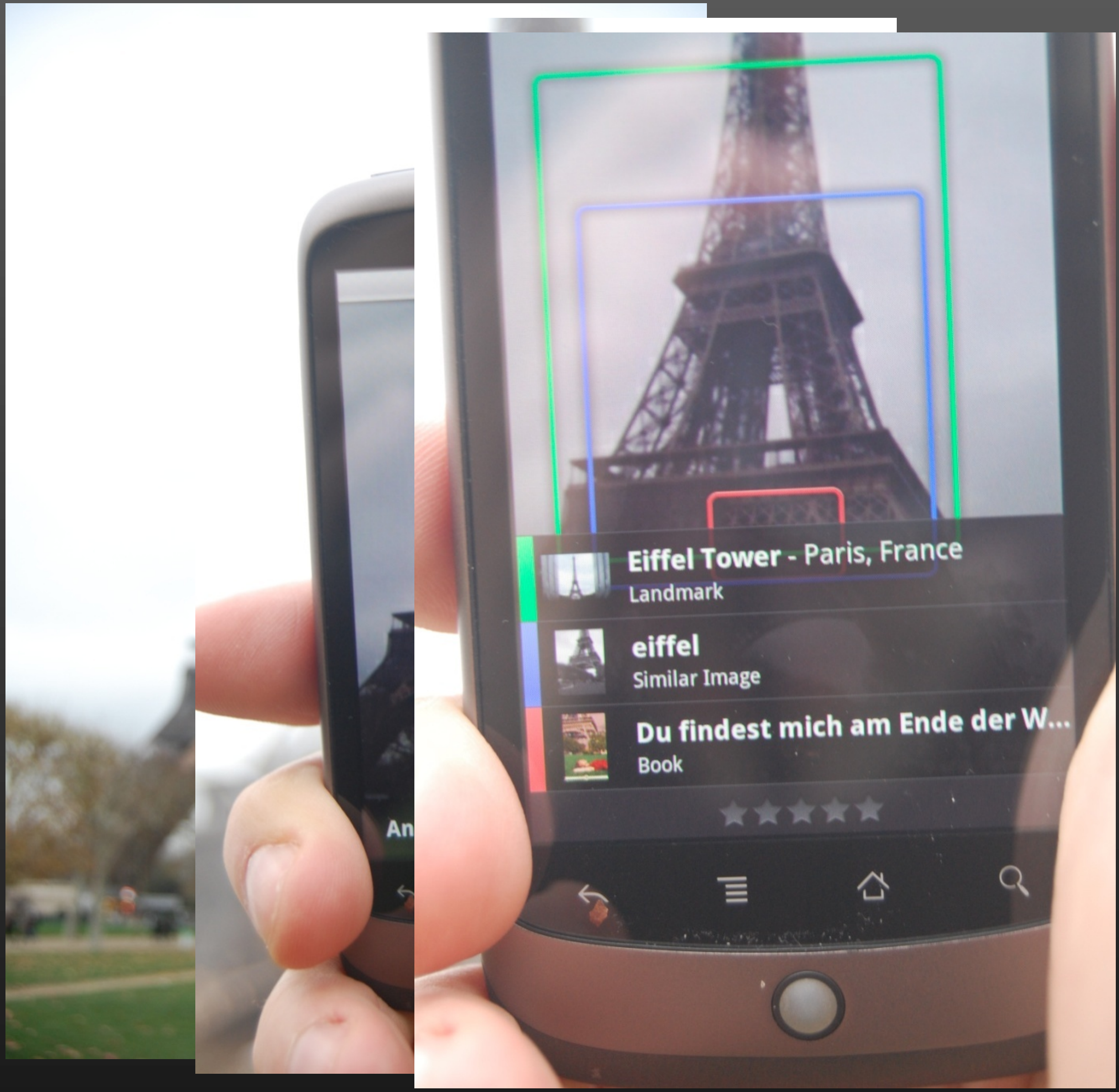


3D Scene reconstruction

Face recognition

Action recognition

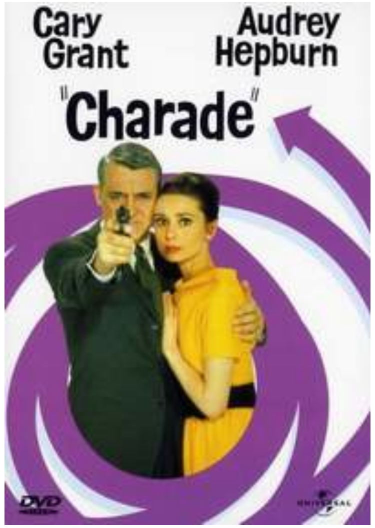
Object instance recognition



Example: Visual search in an entire feature length movie

Visually defined query

"Find this bag"



"Charade" [Donen, 1963]

Demo:
http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html

Instance level recognition : still difficult


See also: this workshop, e.g. Perdoch et al.’15, Fernando et al.’14, Schindler et al.’06, Martin-Buralla’15, Matzen&Snavely’14

Example I.: Localize non-photographic depictions

Inputs: paintings, drawings, historical photographs, reference 3D model

Output: recovered artist/camera viewpoints

[Aubry, Russell, Sivic, 2013]
http://www.di.ens.fr/willow/research/painting_to_3d/

Matching non-photographic depictions

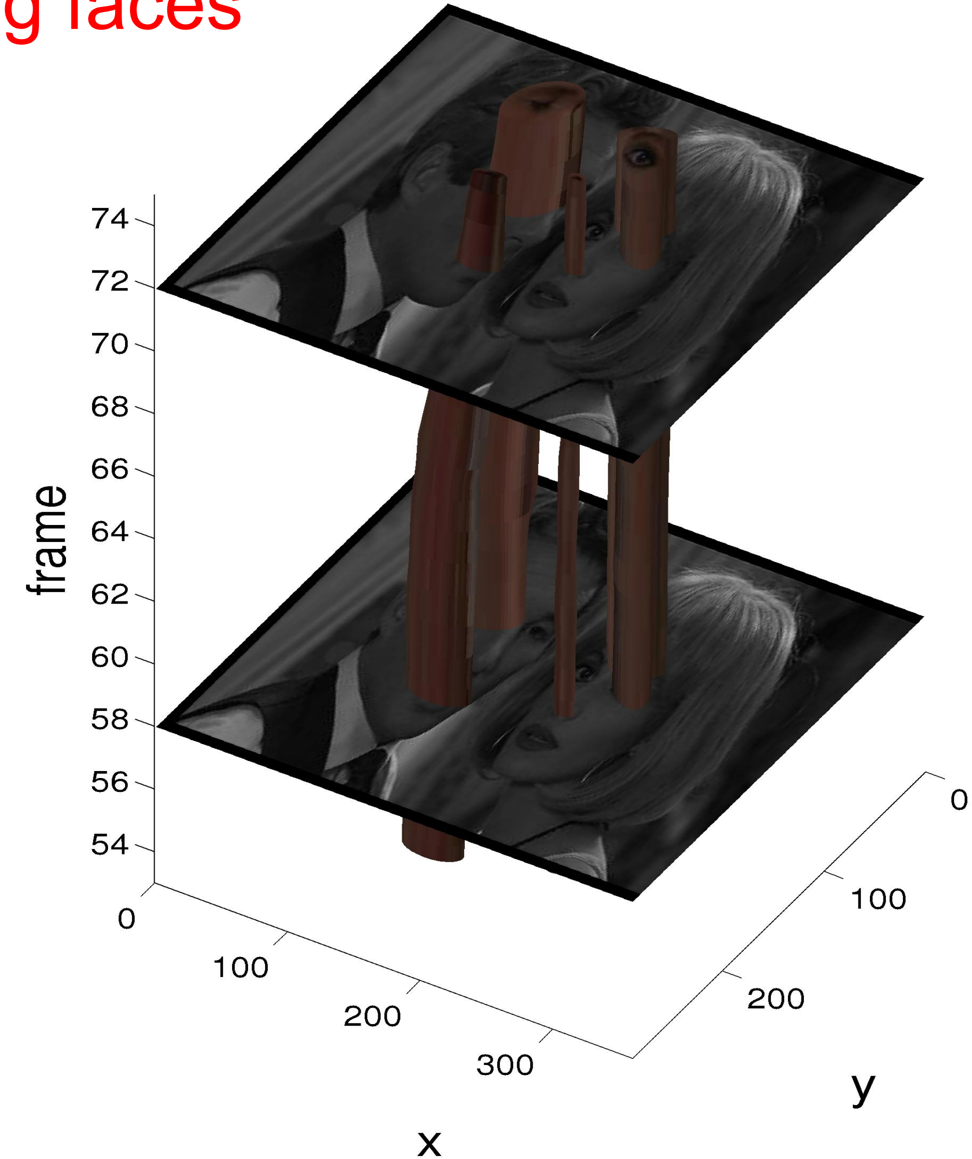
Geo-localization of historical and non-photographic depictions


© 2014 Google
Image Landsat
Image © 2014 DigitalGlobe
Google earth

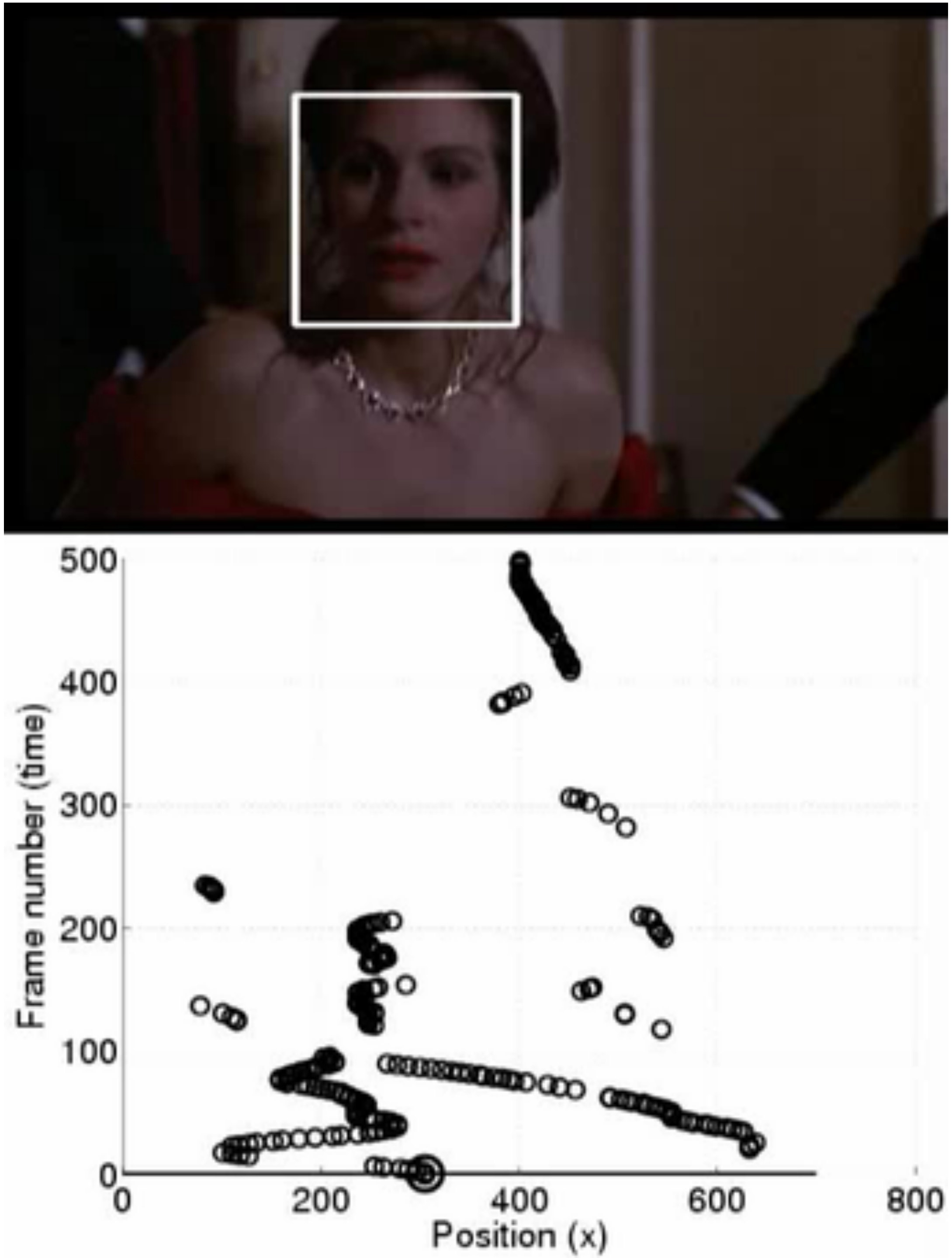
Recognizing people

(Sivic, Everingham, Zisserman, 2005)

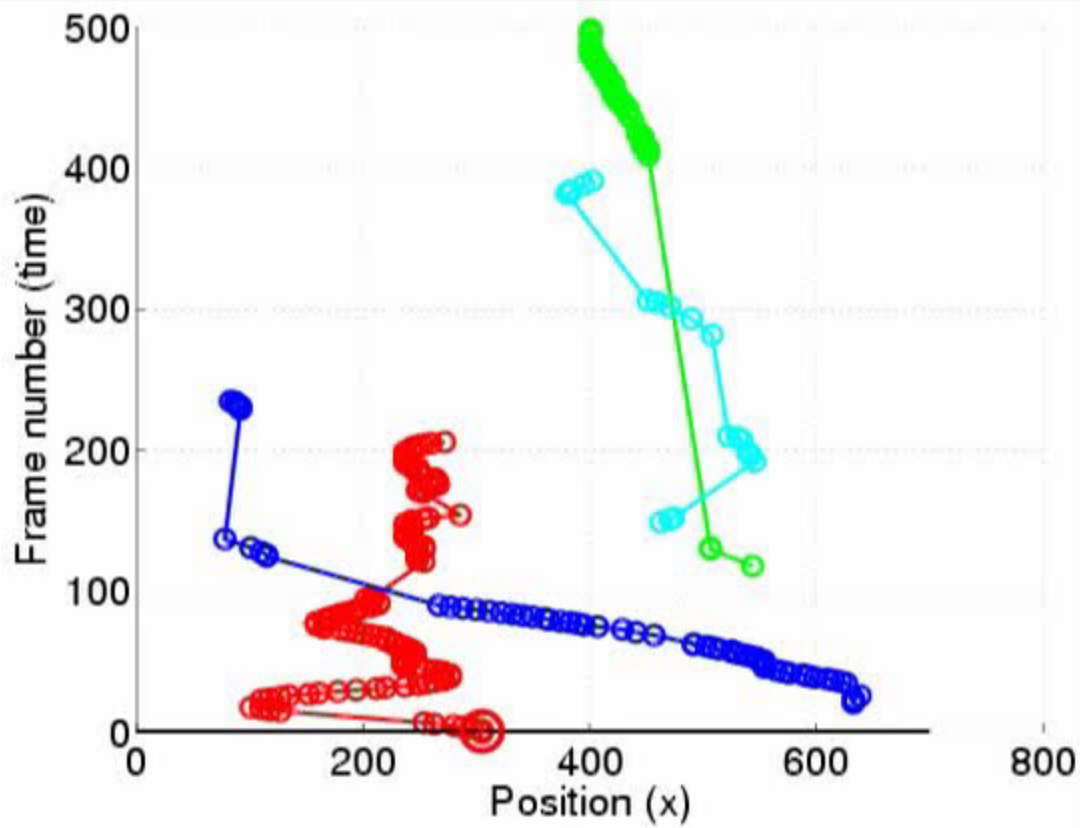
Faces:
Region tubes for tracking faces



[Sivic, Everingham and Zisserman, 2005]



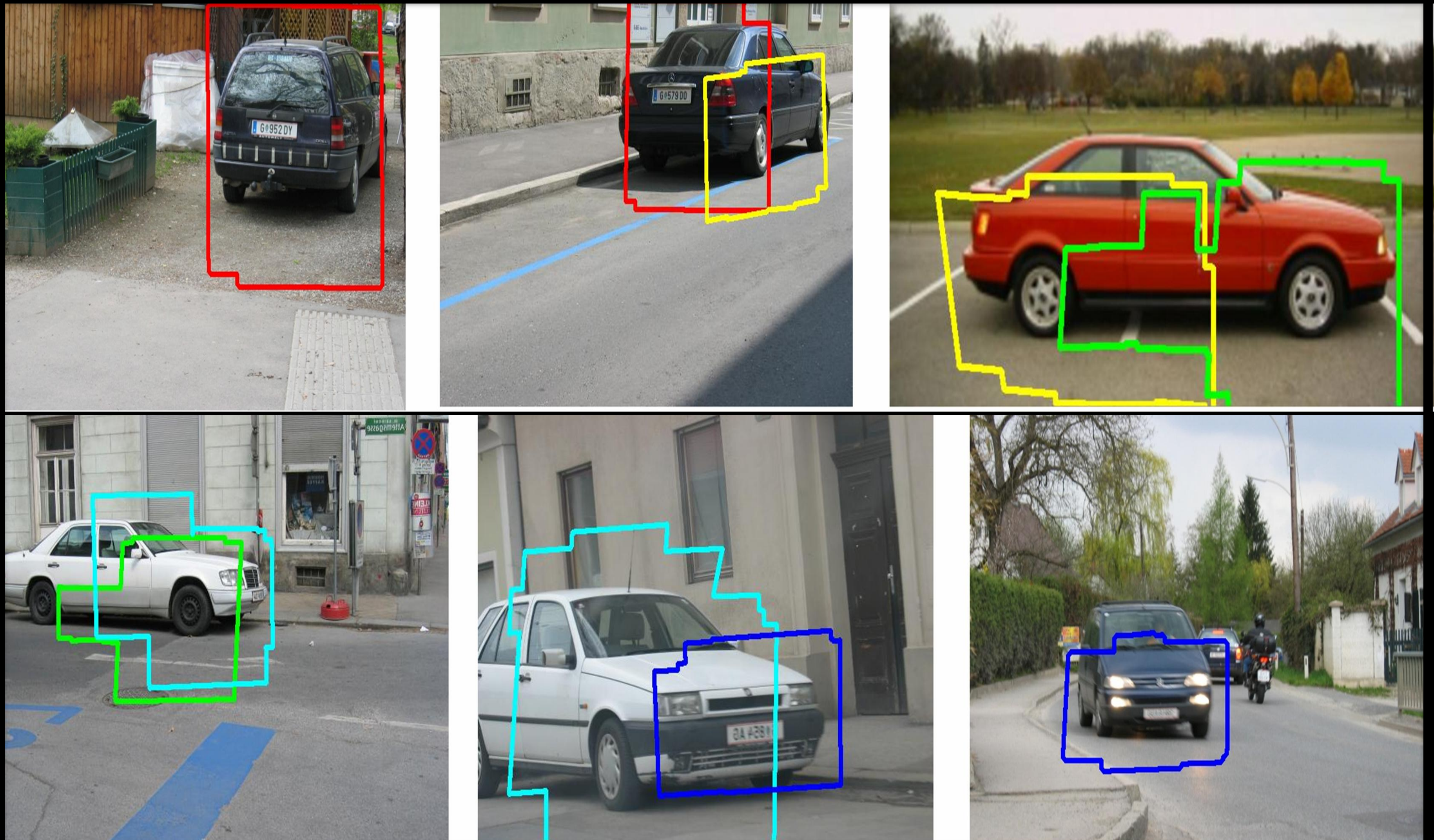
Raw face detections



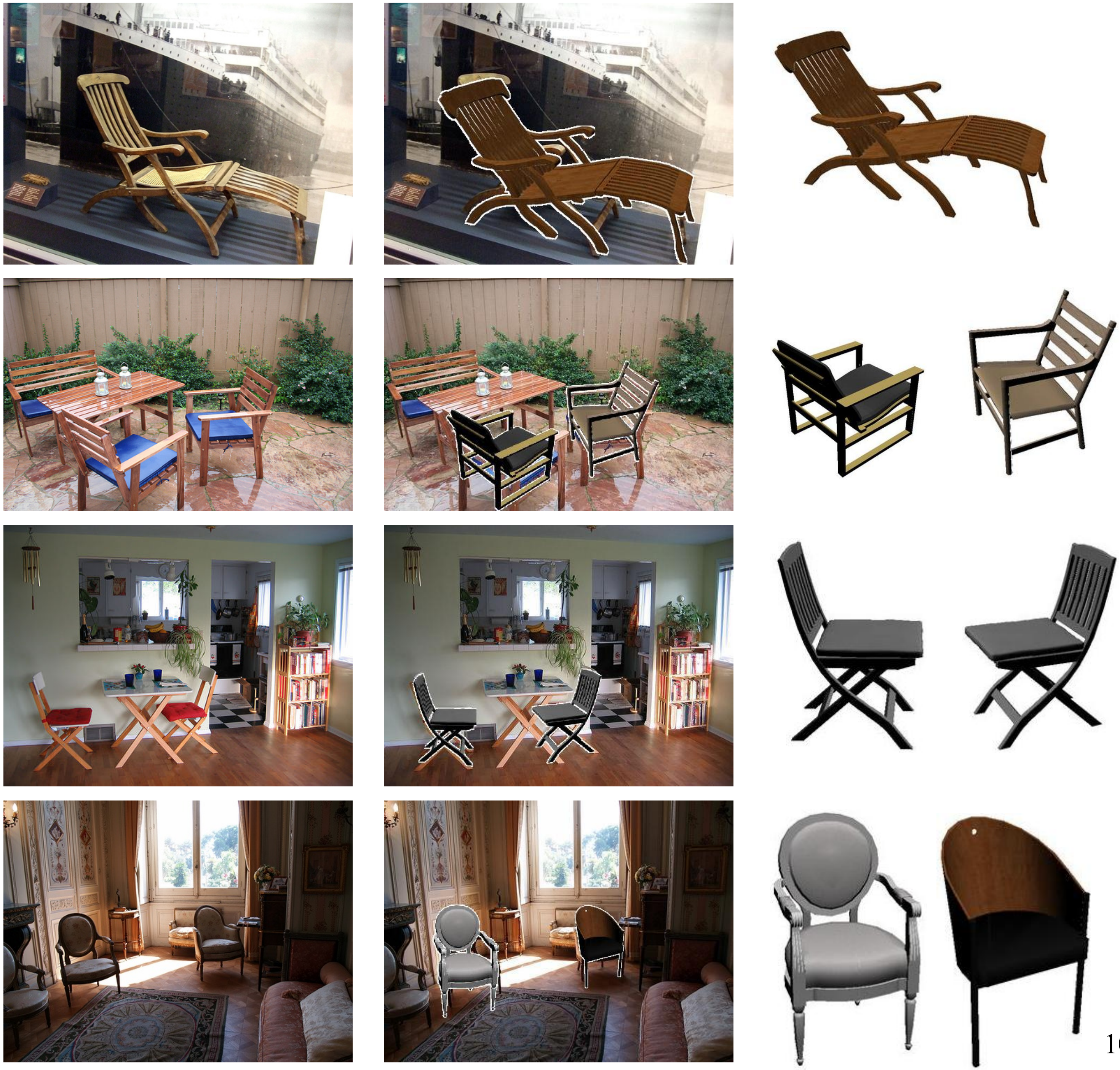
Tracking by detection and recognition

Connected face tracks

Recognition



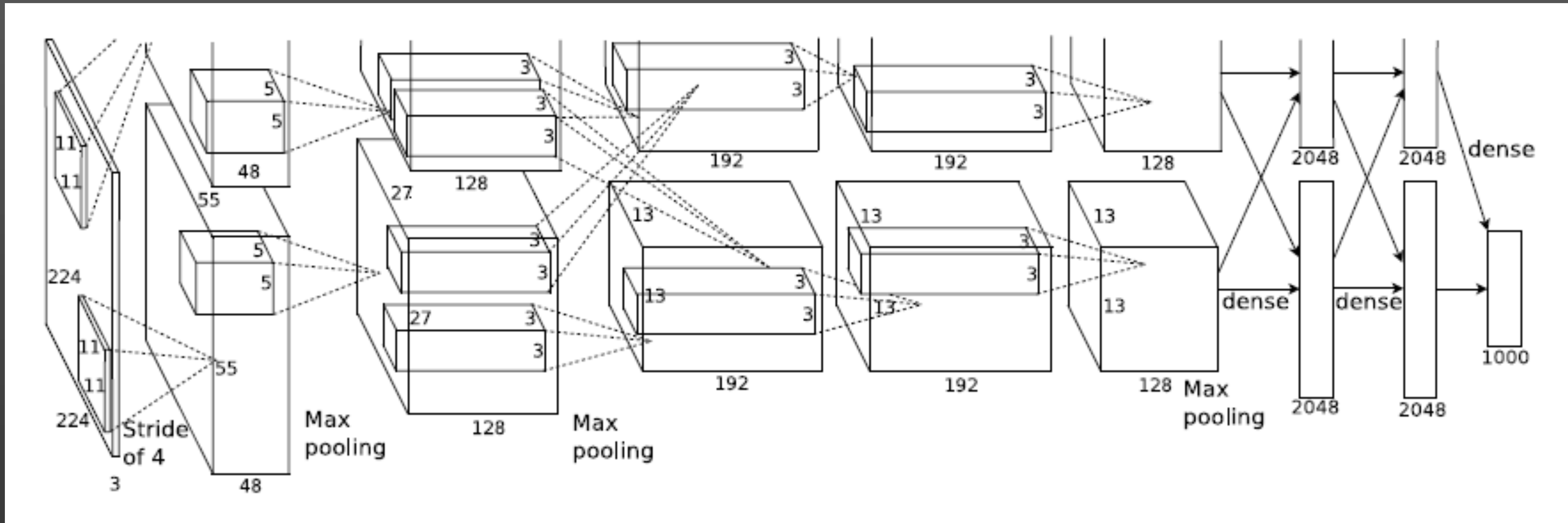
(Kushal et al., 2007)



(Aubry et al., 2014)

Convolutional neural networks

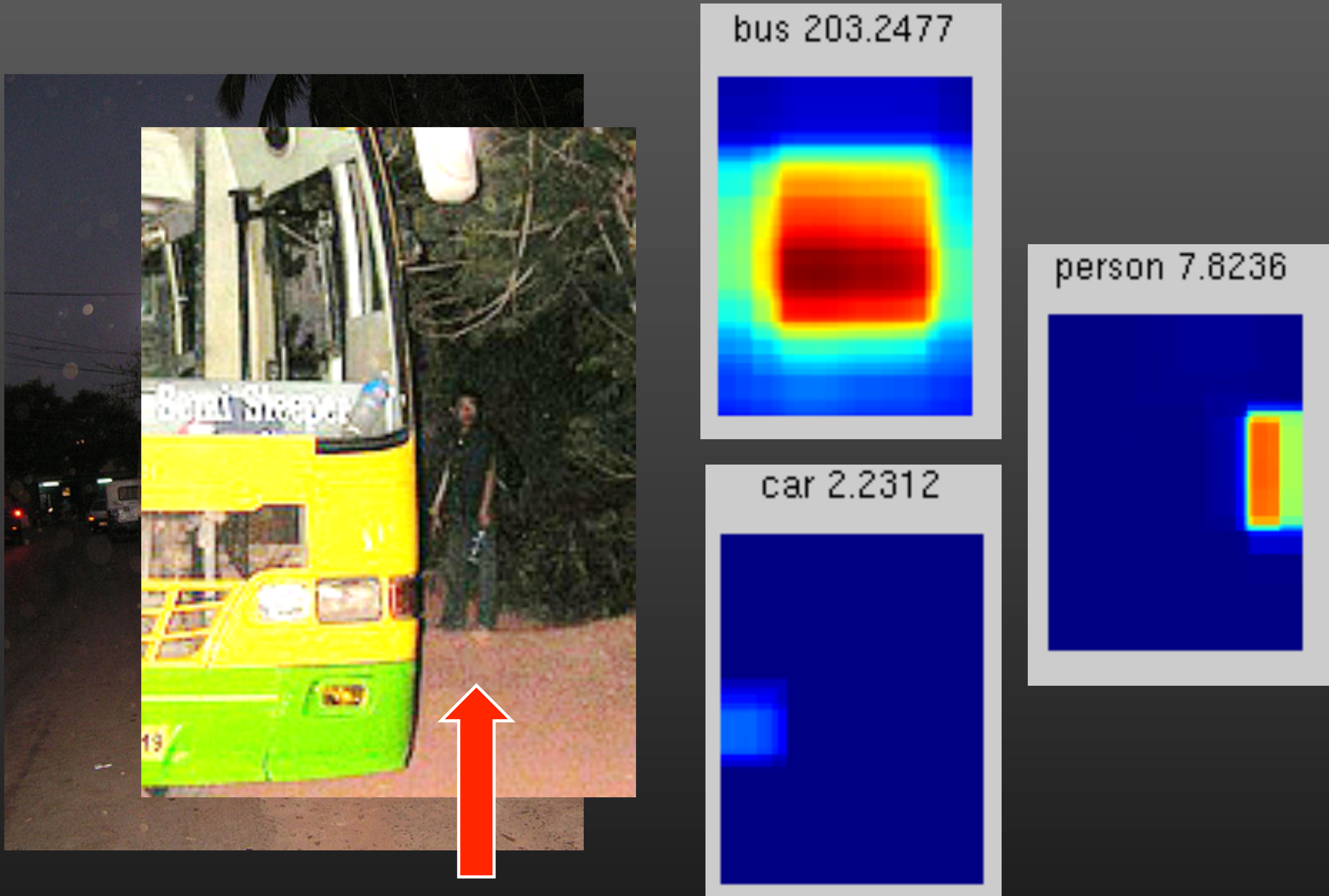
[Krizhevsky et al. NIPS'12]



Convolutional Neural Networks:

- The main principles are known since LeCun'88
- Has 60M parameters and 650K neurons.
- Success is determined by (a) lots of labeled images and (b) fast GPU implementation. Both (a) and (b) have not been available until very recently.

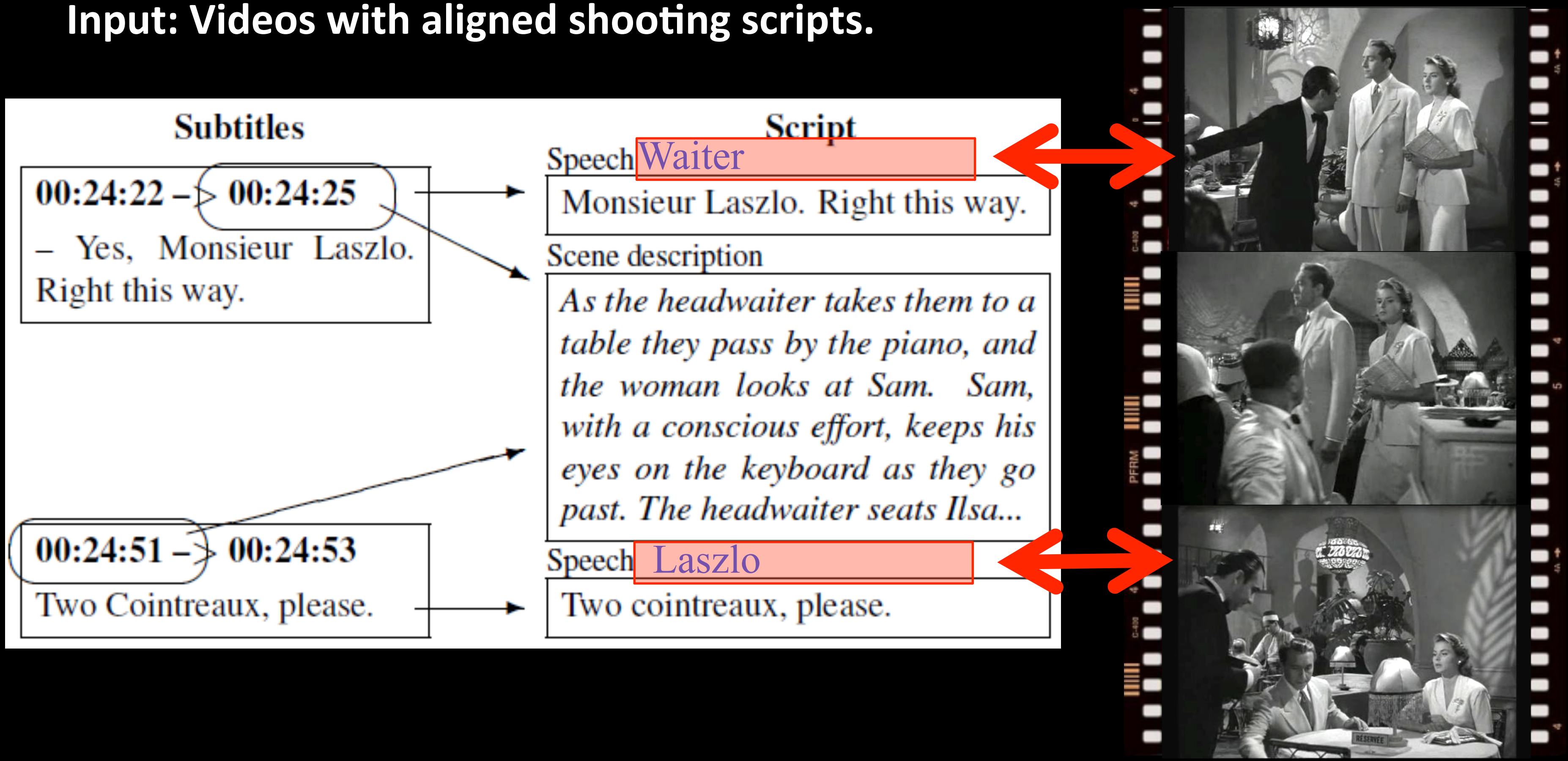
Some results



[Oquab, Bottou, Laptev, Sivic, CVPR 2014]

Automatic learning from video scripts

Input: Videos with aligned shooting scripts.



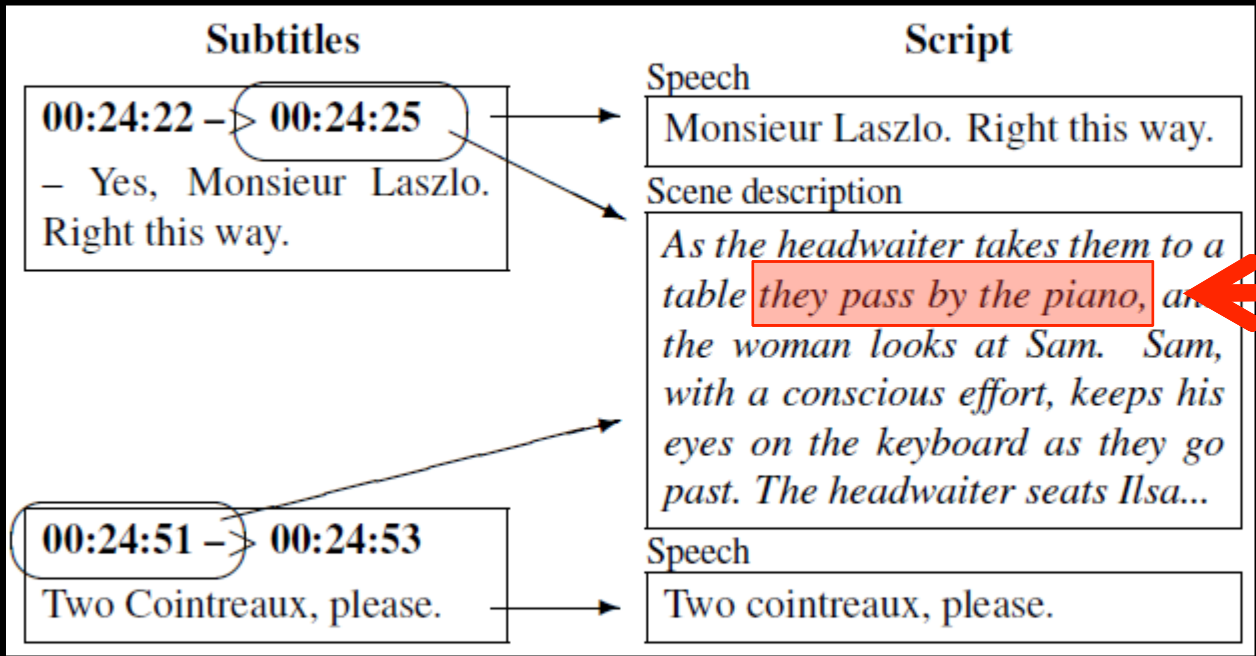
Output: Recognizer for each character in the video

Recognizing people

(Everingham, Sivic, Zisserman, 2009)

Automatic learning from video scripts

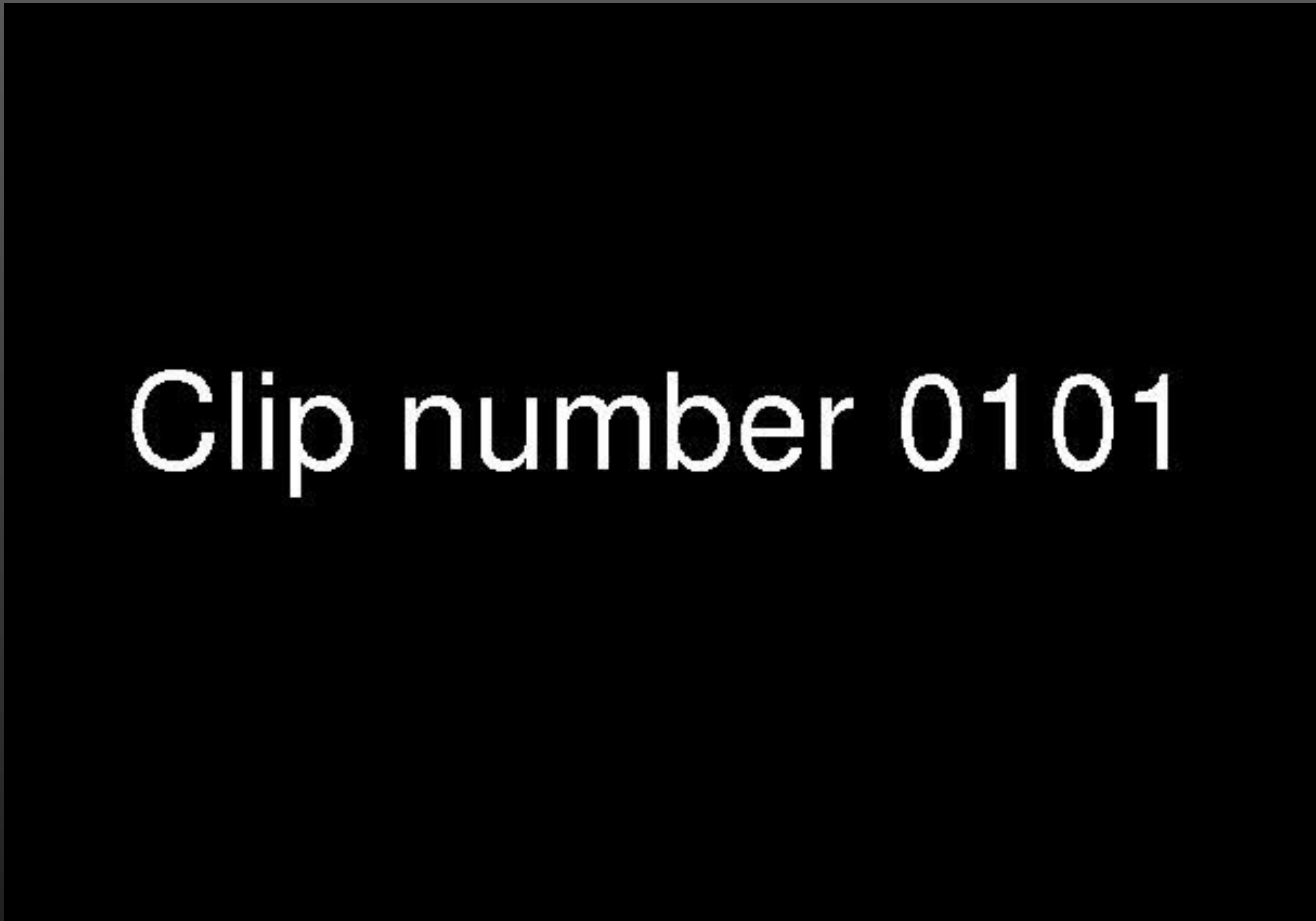
Input: Videos with aligned shooting scripts.



Output: detector of human actions.

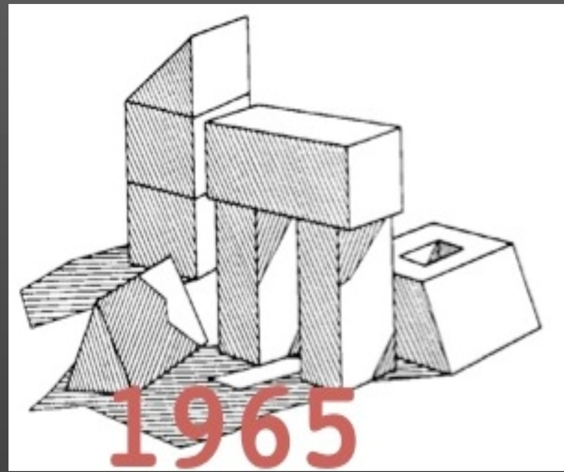
See also [Laptev, Marszałek, Schmid, Rozenfeld 2008]

Weakly-supervised video interpretation

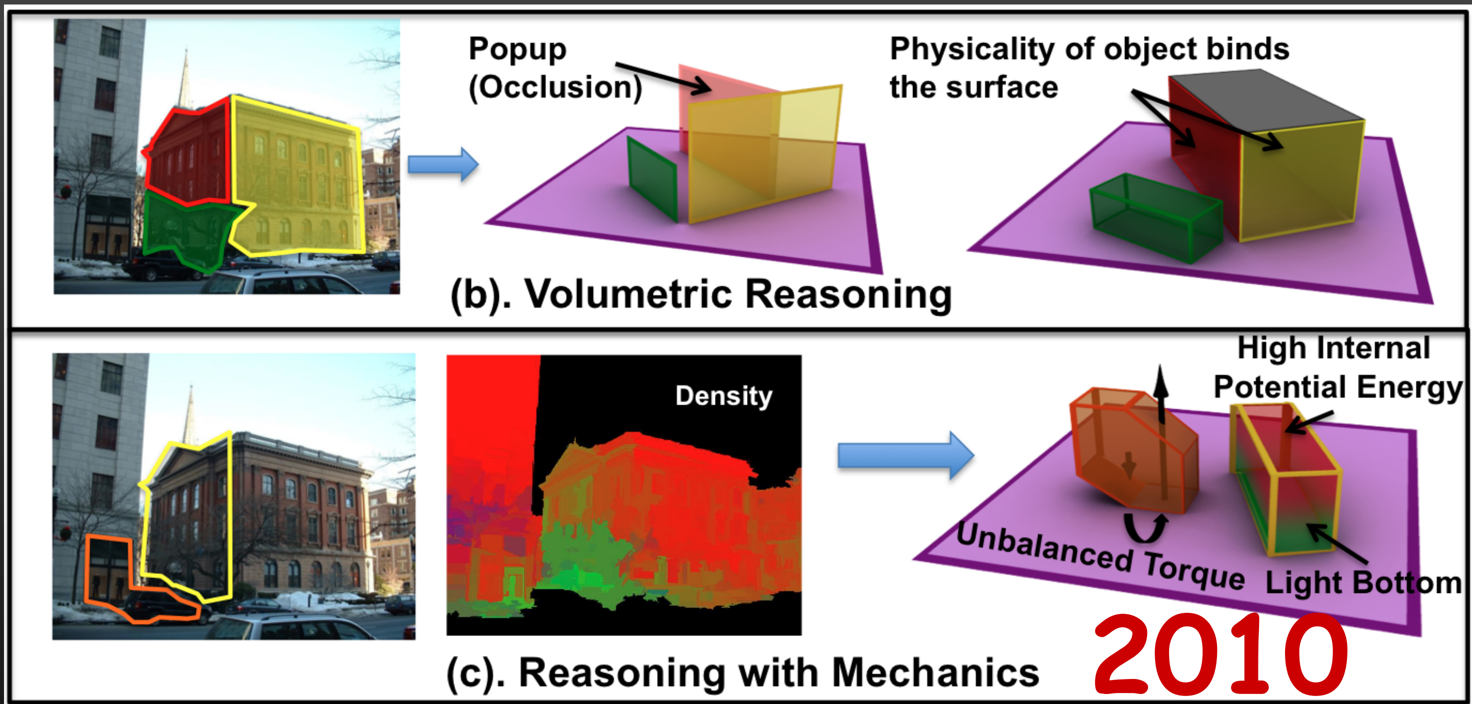


(Bojanowski et al., 2014)

What about scene understanding?



The blocks world revisited



(Gupta, Efros, Hebert, ECCV'10)