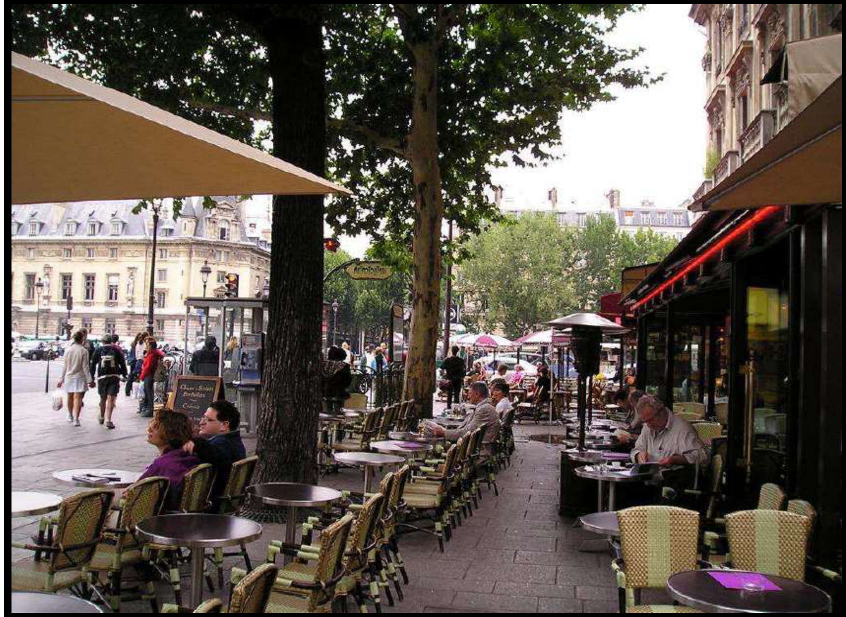
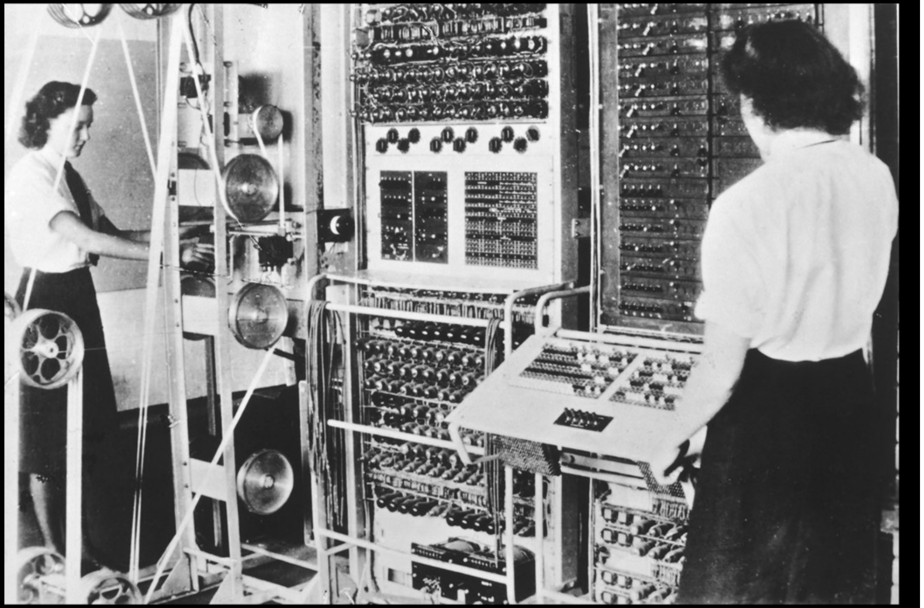
Introduction to computer vision

Instructors: Jean Ponce and Matthew Trager
jean.ponce@inria.fr, matthew.trager@cims.nyu.edu

TAs: Jiachen (Jason) Zhu and Sahar Siddiqui
jiachen.zhu@nyu.edu, ss12414@nyu.edu

Slides will be available after class at:
https://mtrager.github.io/introCV-fall2019/





Description:

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

Computer vision

… extracting information from images and video
(courtesy of I. Laptev)

Vision is hard—this is what the machine "sees"

- The visual cortex is about 50% of the macaque brain
- More human brain is dedicated to vision than anything else

WHY IS VISION DIFFICULT ?

Too much information:

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

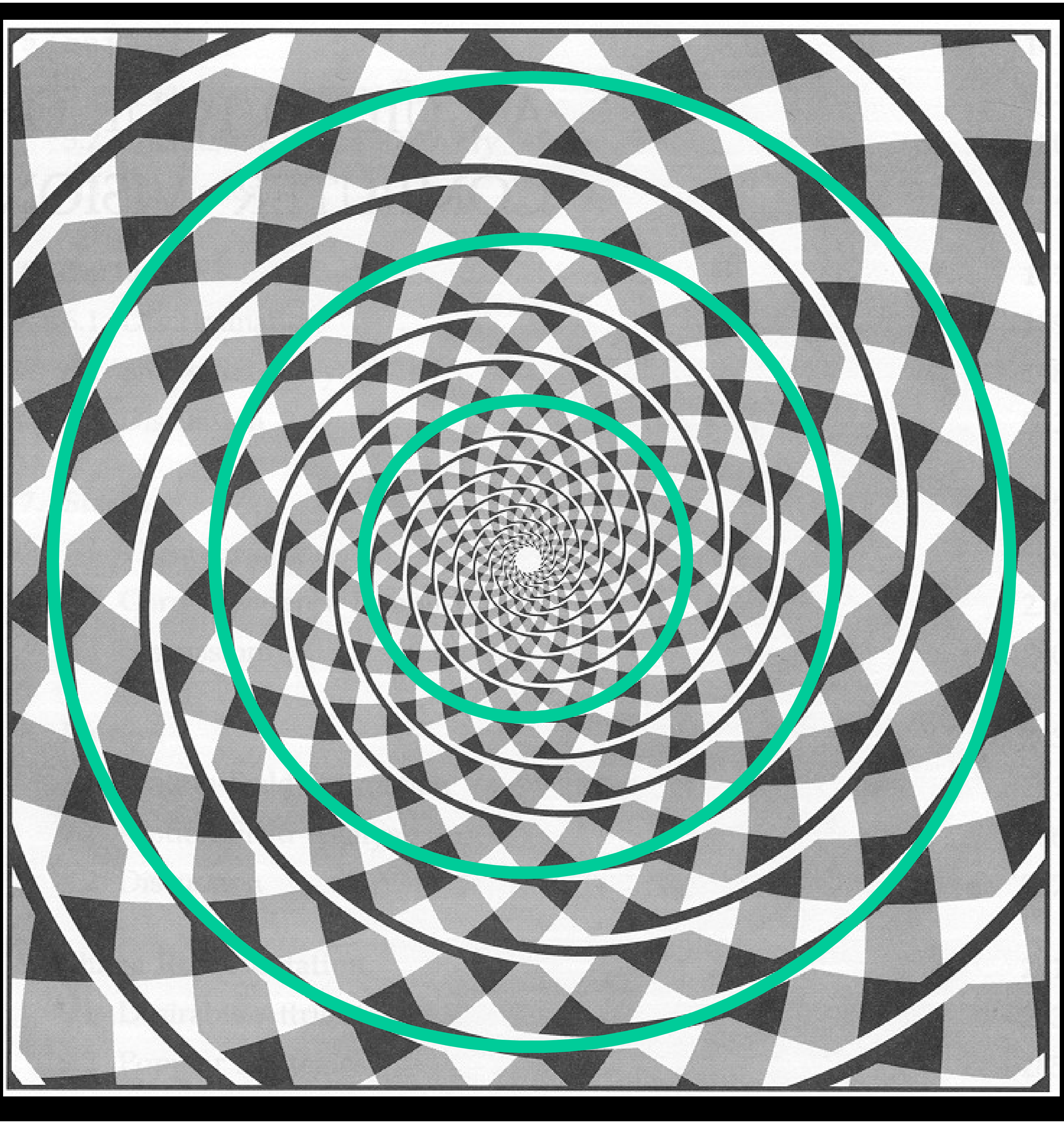
Too little information:

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

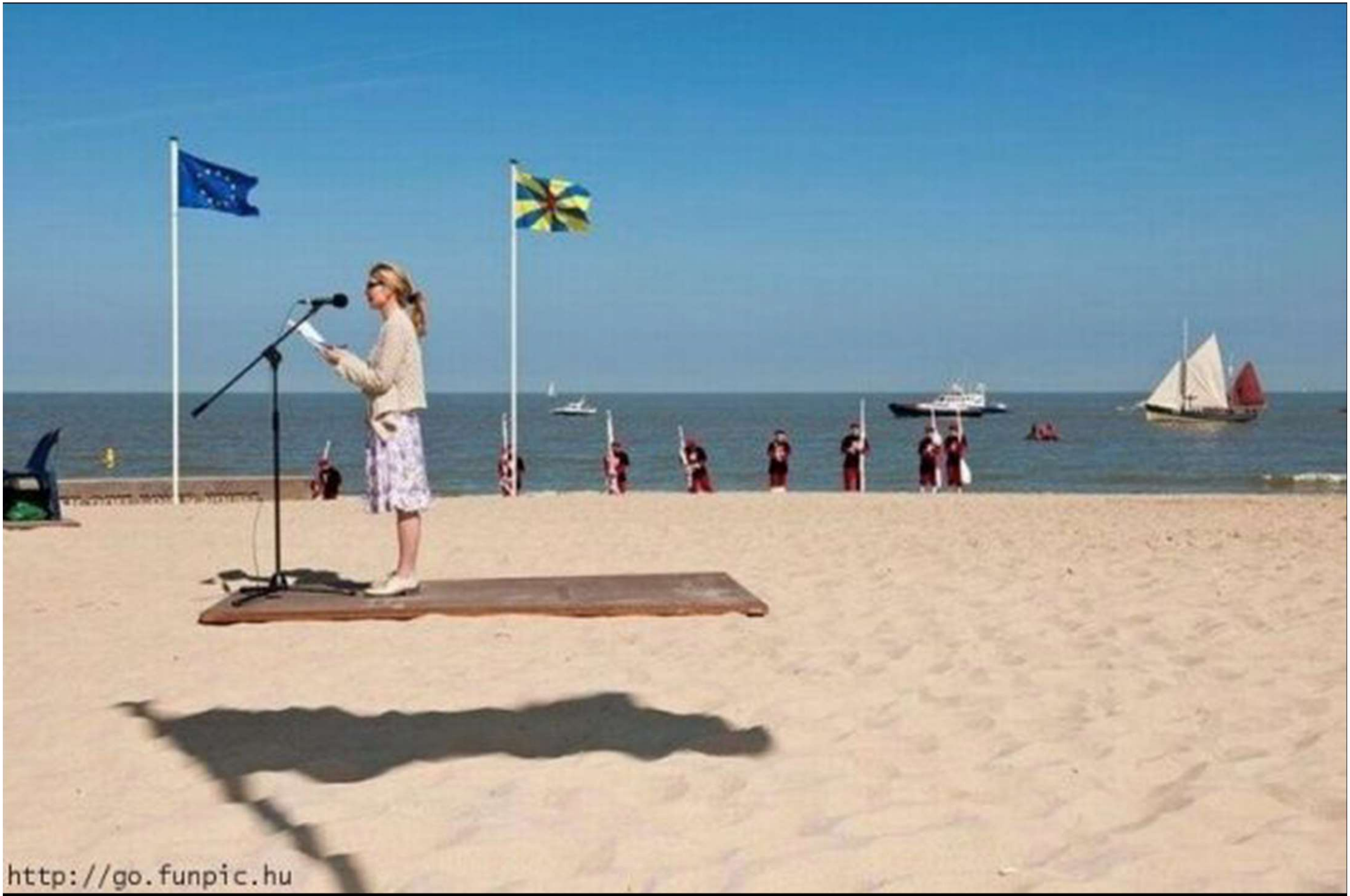
What are appropriate models?

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

What are appropriate algorithms and architectures?



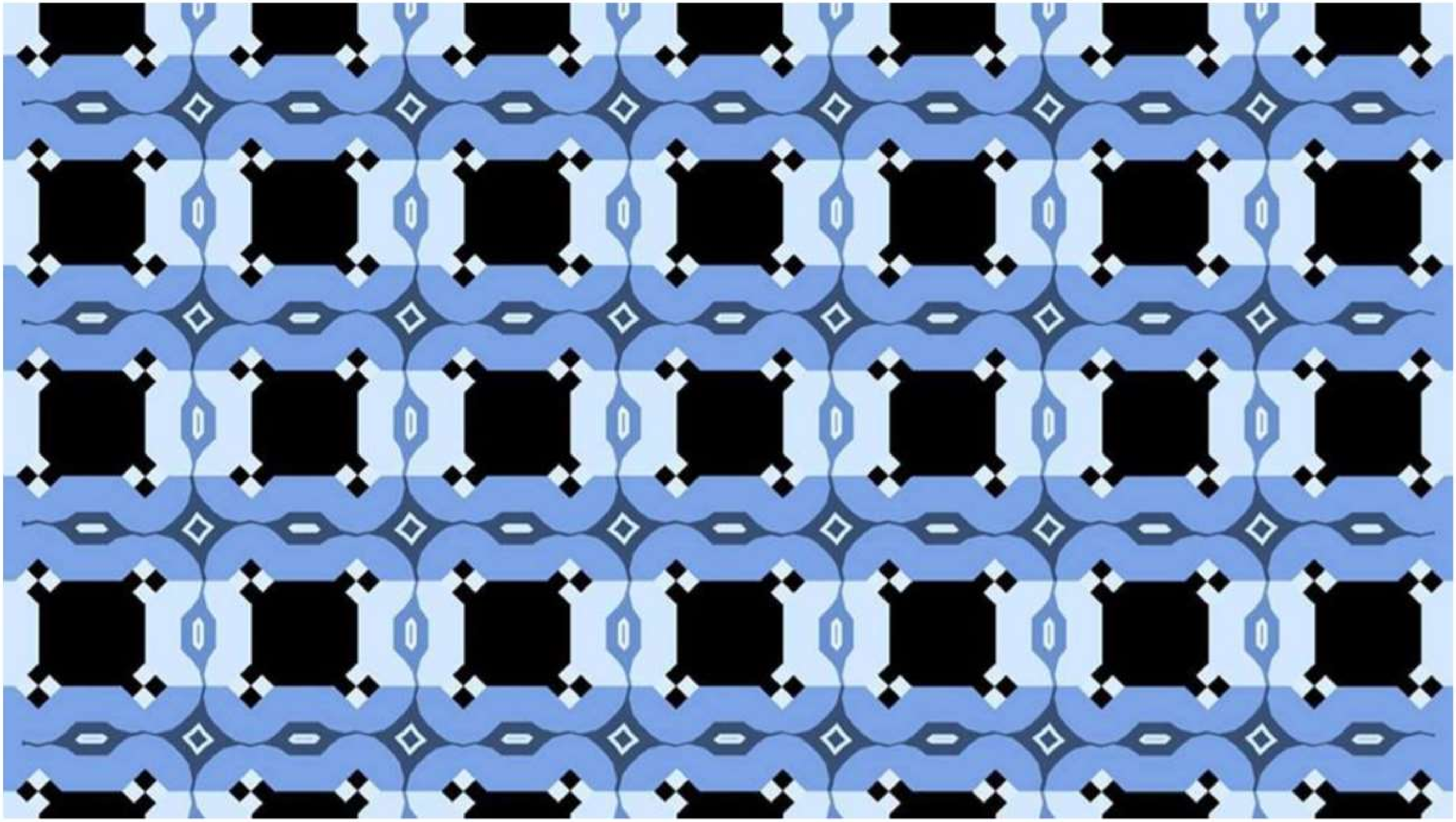
(Nalwa, 1993)



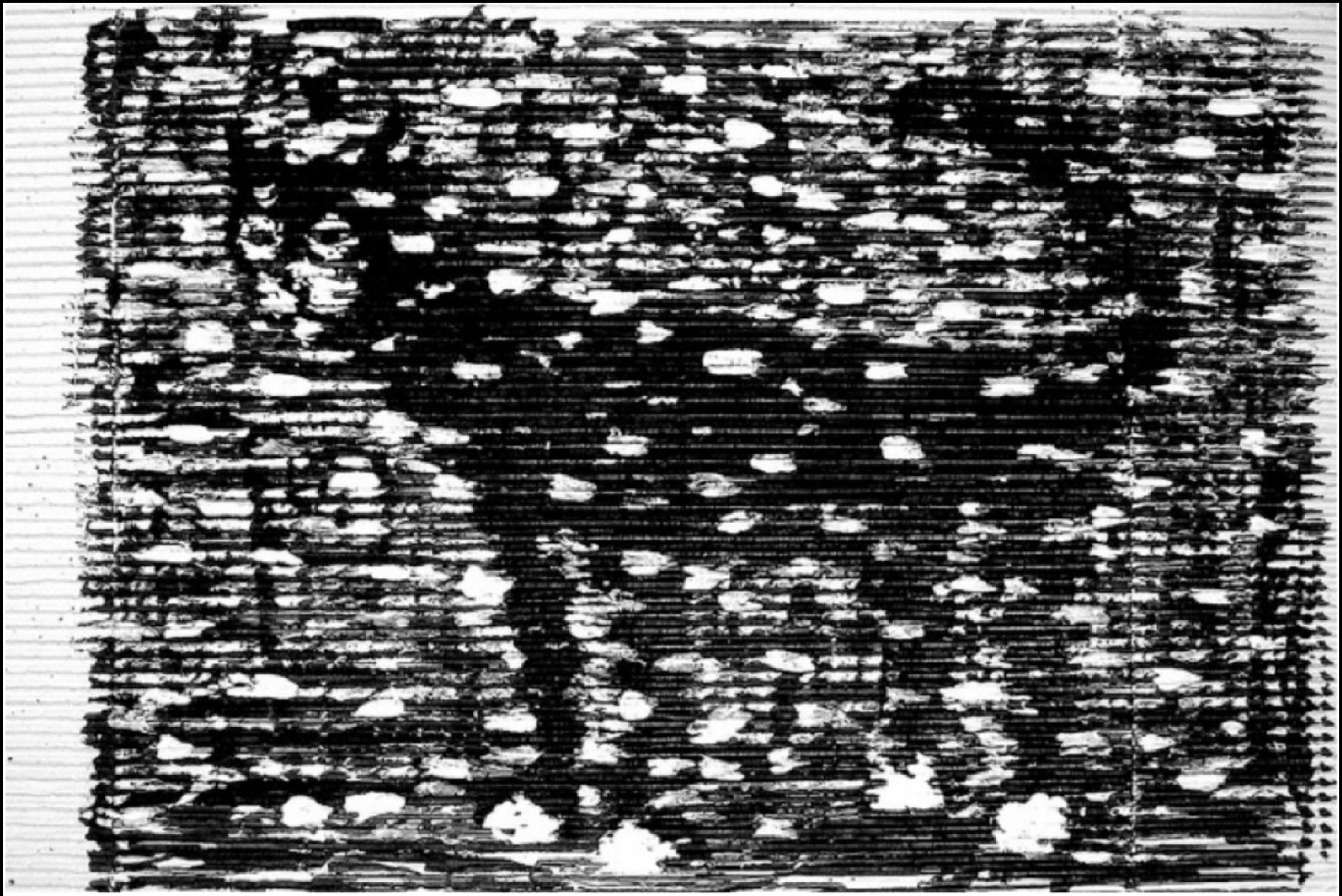
J.J. Koenderink, www.gestaltrevision.be/en/resources/clootcrans-press

COMPUTER VISION IS INTERESTING.

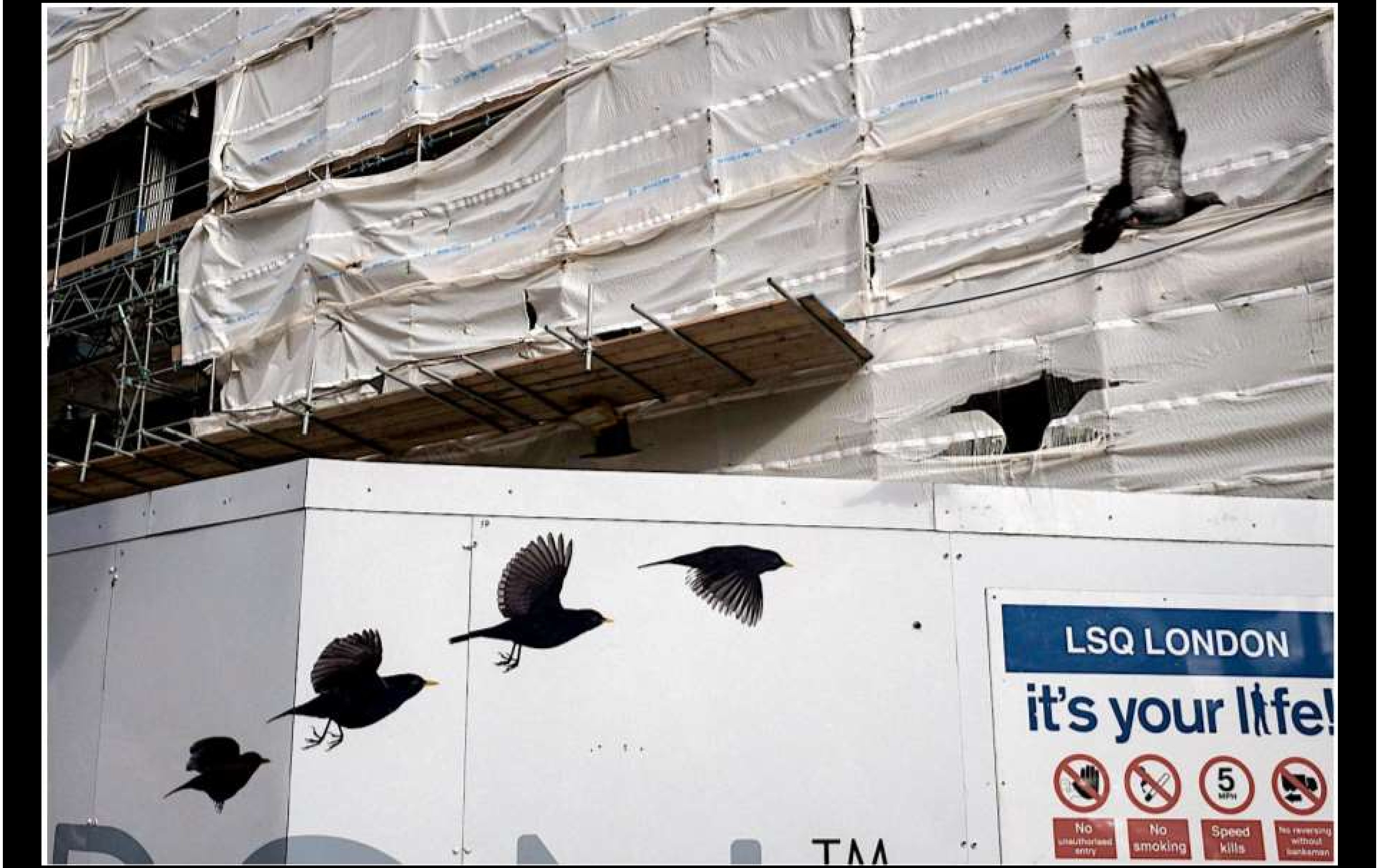
- We know it is possible.
- We know it is difficult.
- We don't (really) know how to do it.



(Victoria Skye)



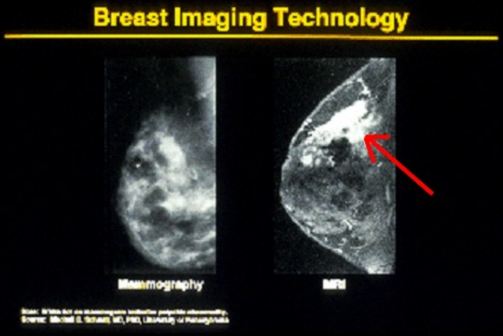
(Franco Mattichio)


LSQ LONDON
it's your life!
No unauthorised entry
No smoking
5 MPH
Speed kills
No reversing without banksman
TM

(Pau Buscato)

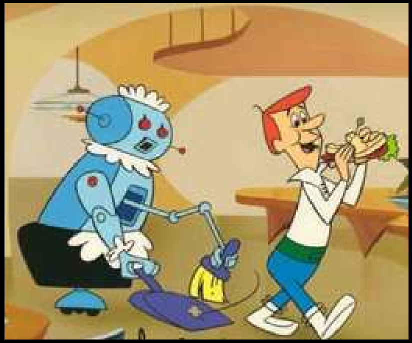
Why computer vision matters

Safety



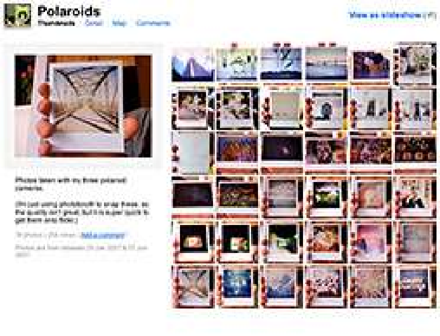
Health

Security



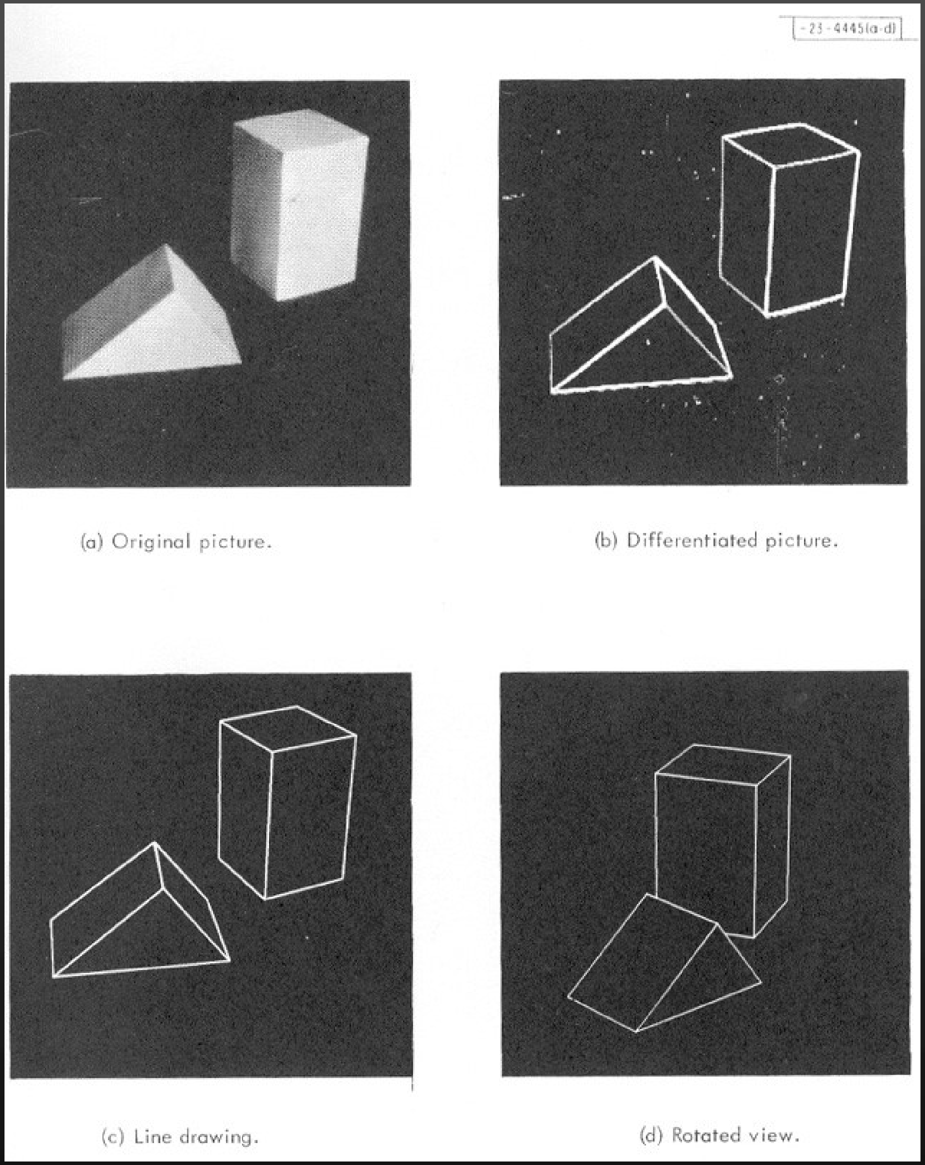
Comfort

Fun



Access

Origins of computer vision



(a) Original picture.

(b) Differentiated picture.

(c) Line drawing.

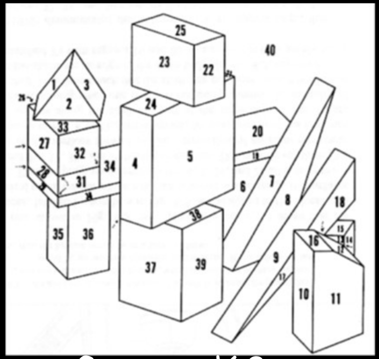
(d) Rotated view.

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

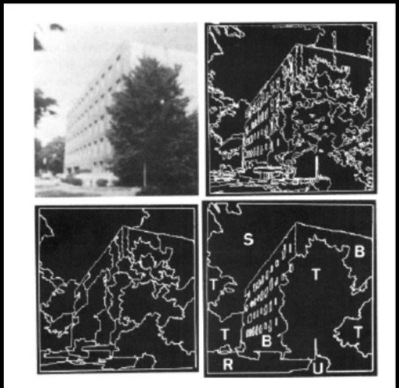
photo credit: Joe Mundy

After Roberts: a ridiculously brief history of computer vision

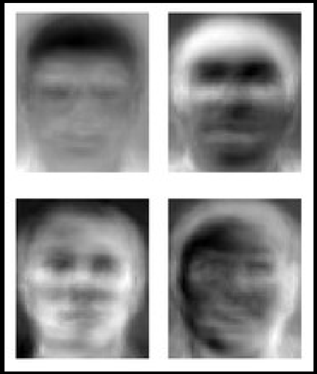
- 1966: Minsky assigns computer vision as an undergrad summer project (??)
- 1960's: interpretation of extremely simple images & synthetic worlds
- 1970's: some progress on interpreting selected images
- 1980's: ANNs come and go; shift toward geometry and increased mathematical rigor
- 1990's: face recognition; 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:

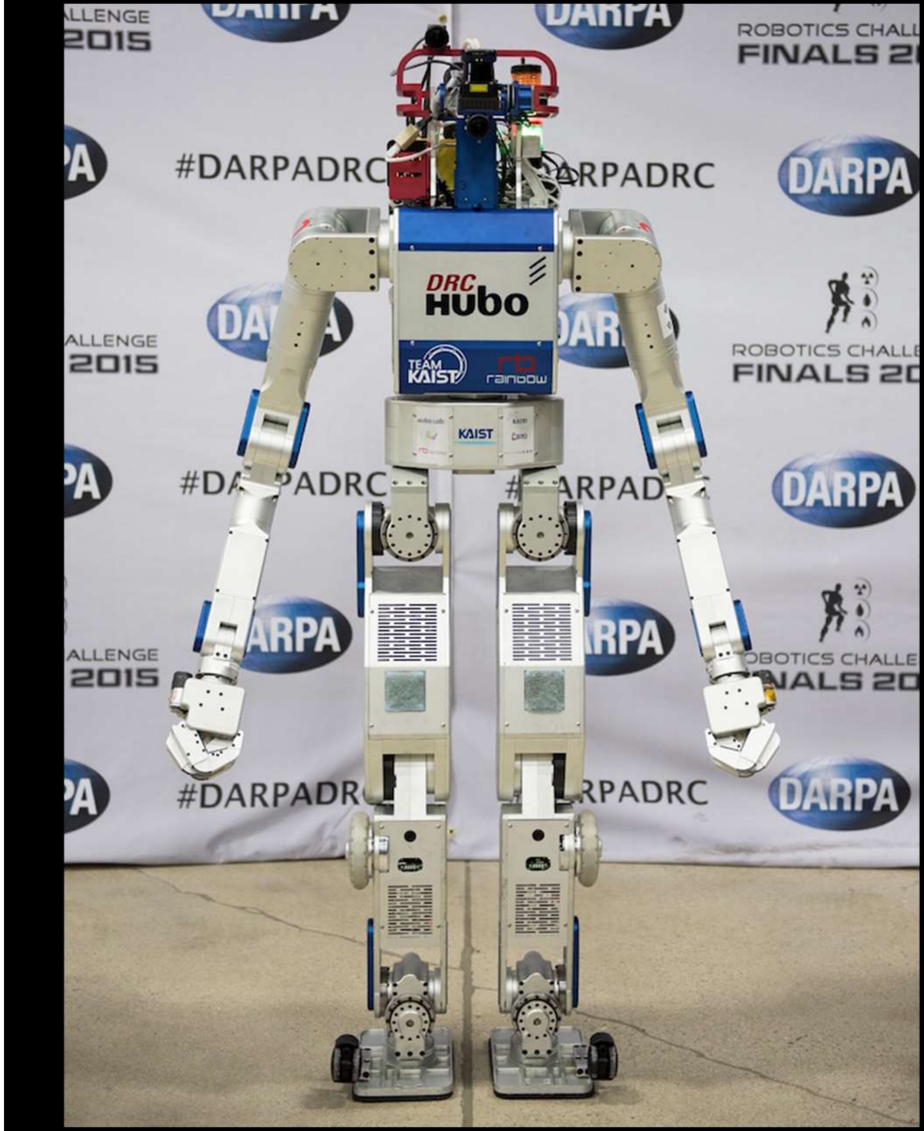
- Manufacturing: inspection, bin picking;
- Defense: ATR, photogrammetry, surveillance;
- Robotics: navigation, visual servoing.

Recently:

- Computer graphics, medical imaging, HCI;
- 3D vision and recognition;
- The Web, Internet, social networks;
- Robotics again;
- And zillions of other industries.

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

NREC
NATIONAL ROBOTICS ENGINEERING CENTER

CMU's Chimp

How vision is used now

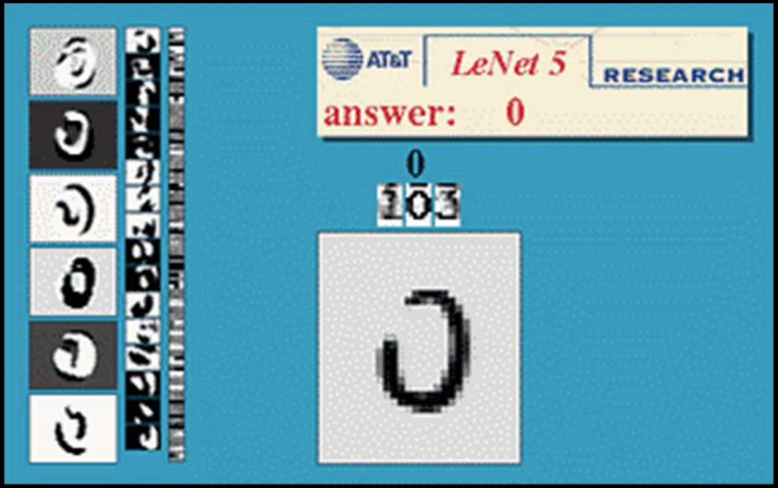
- Examples of recent real world applications

Some of the following slides by Steve Seitz

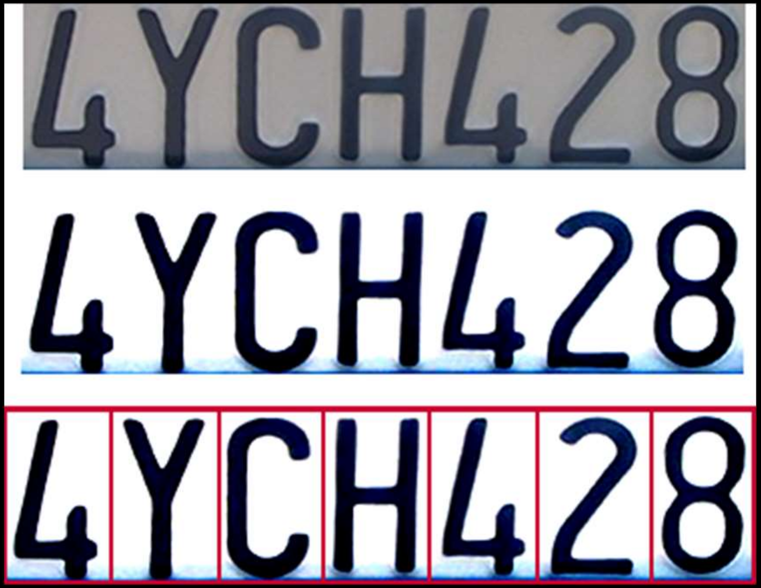
Optical character recognition (OCR)

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/

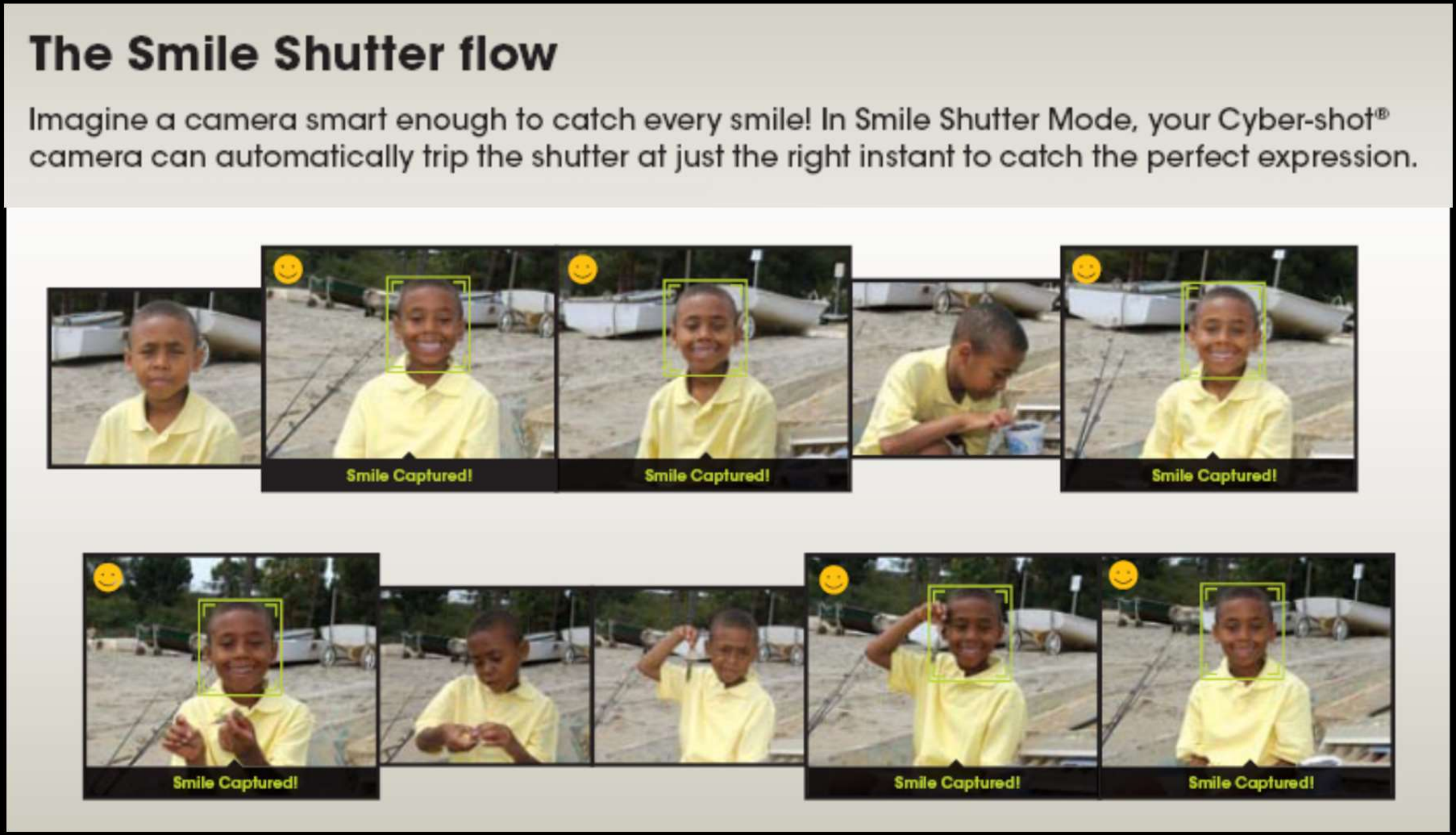


License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

Face detection

- All digital cameras detect faces

Smile detection



The Smile Shutter flow

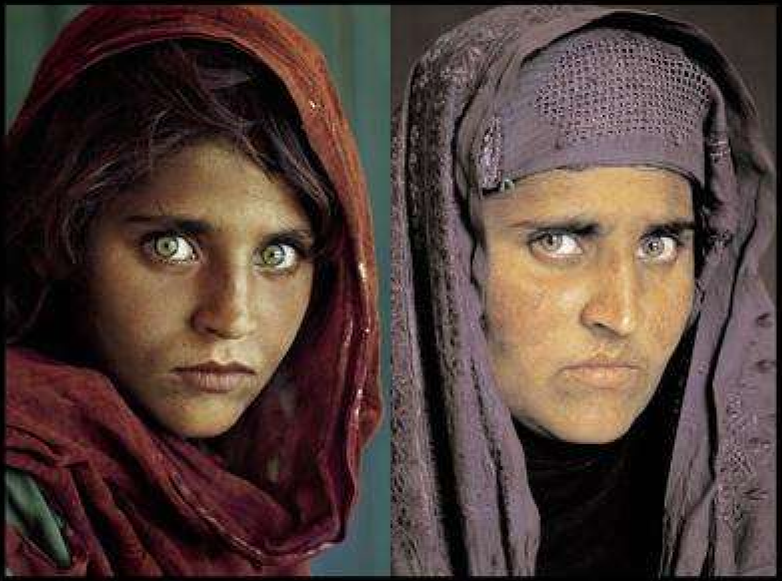
Imagine a camera smart enough to catch every smile! In Smile Shutter Mode, your Cyber-shot® camera can automatically trip the shutter at just the right instant to catch the perfect expression.

Sony Cyber-shot® T70 Digital Still Camera

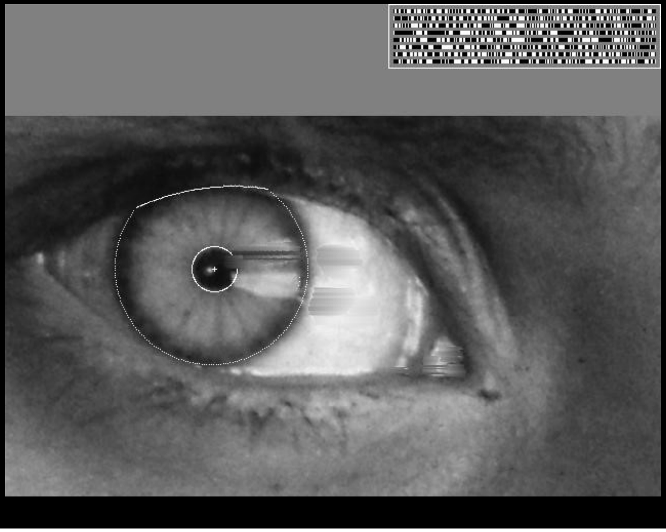
Structure from motion from busloads of images

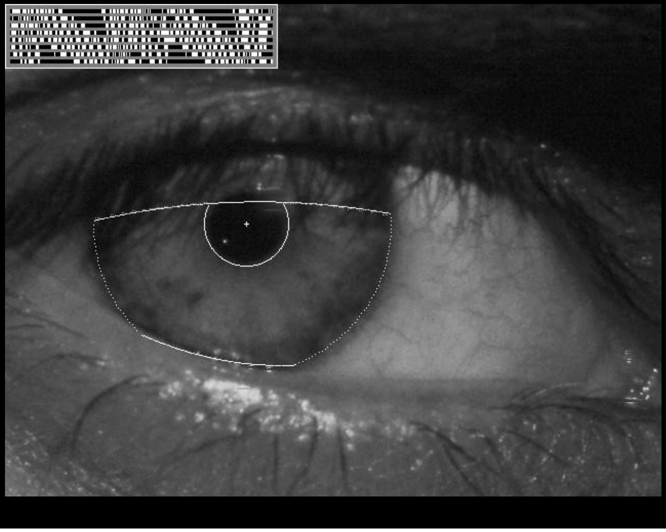
(Agarwal et al. 2009)

Vision-based biometrics



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



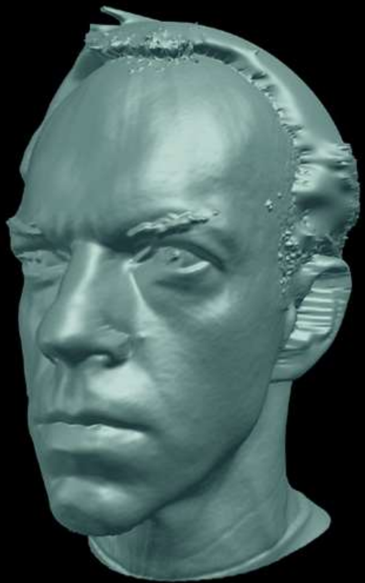


Object recognition (in mobile phones)

Point & Find, Nokia

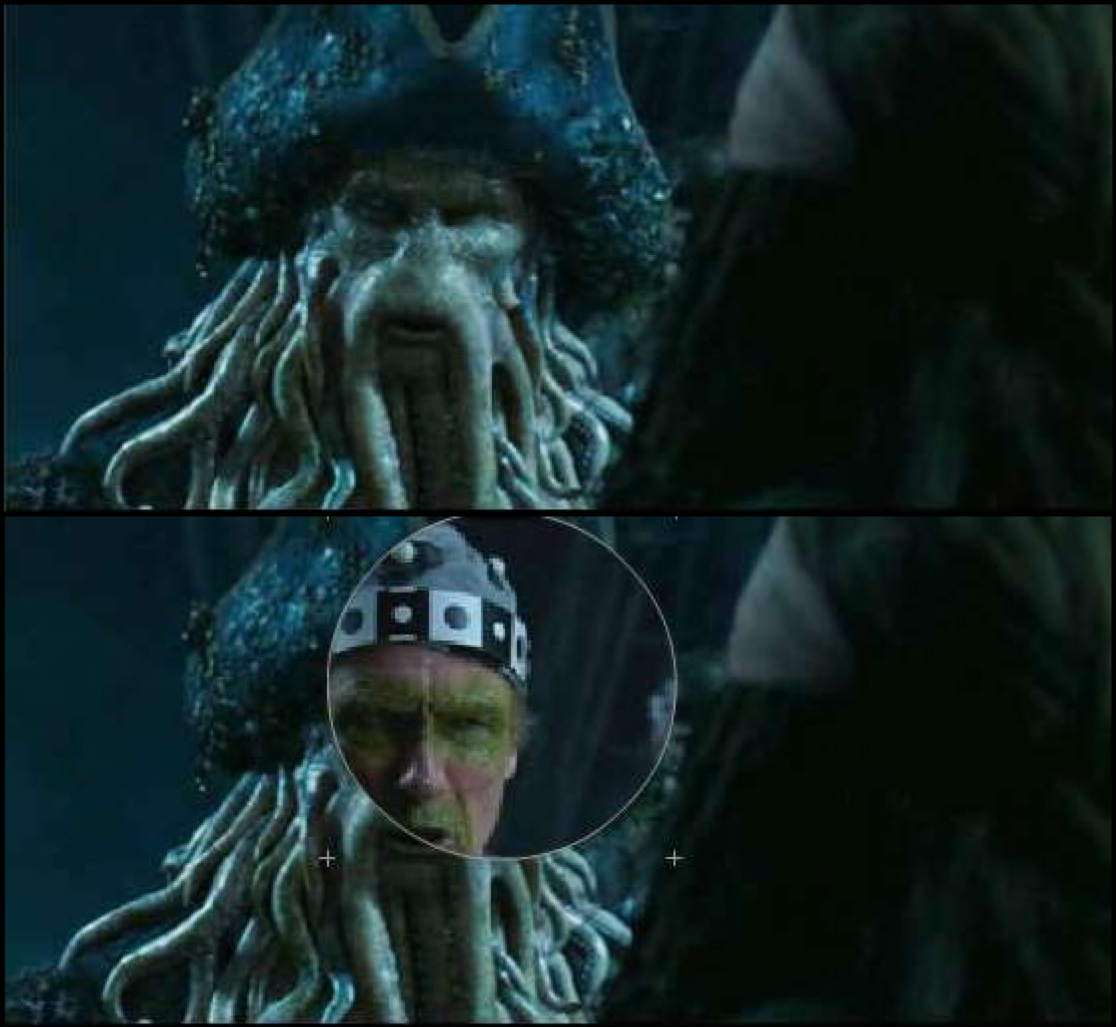
Google Goggles

Special effects: shape capture



The Matrix movies, ESC Entertainment, XYZRGB, NRC

Special effects: motion capture

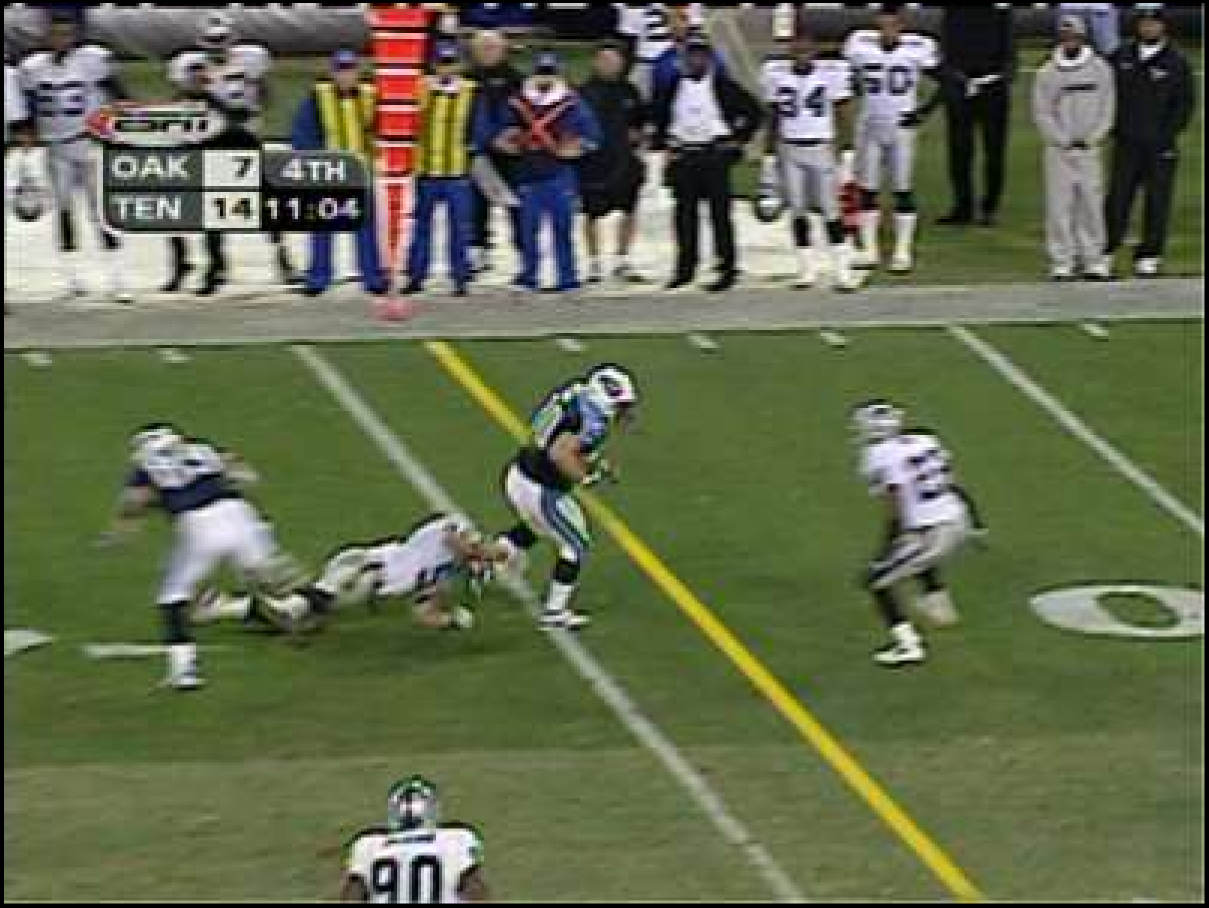


Pirates of the Carribean, Industrial Light and Magic

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

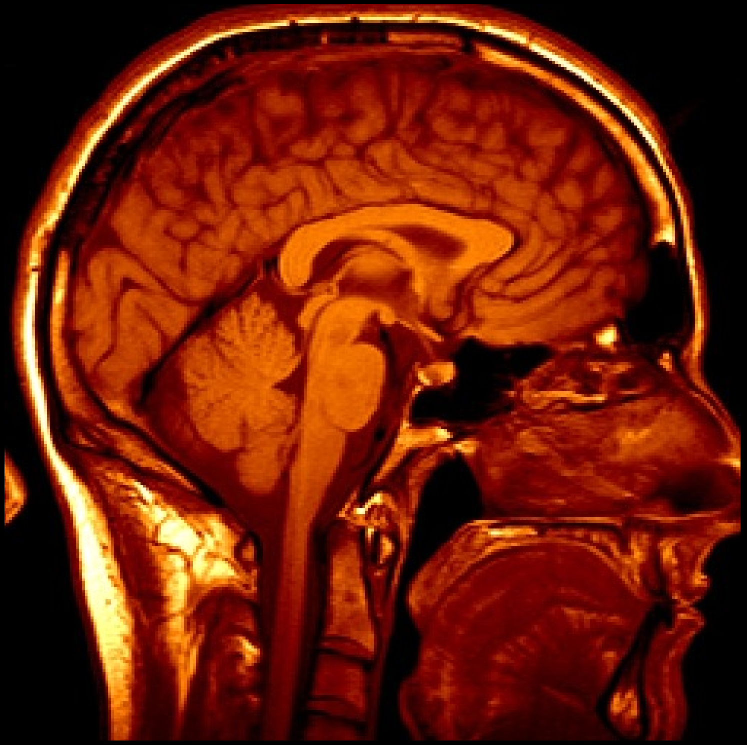
Sports



Sportvision first down line
Nice explanation on www.howstuffworks.com

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

Medical imaging



3D imaging
MRI, CT

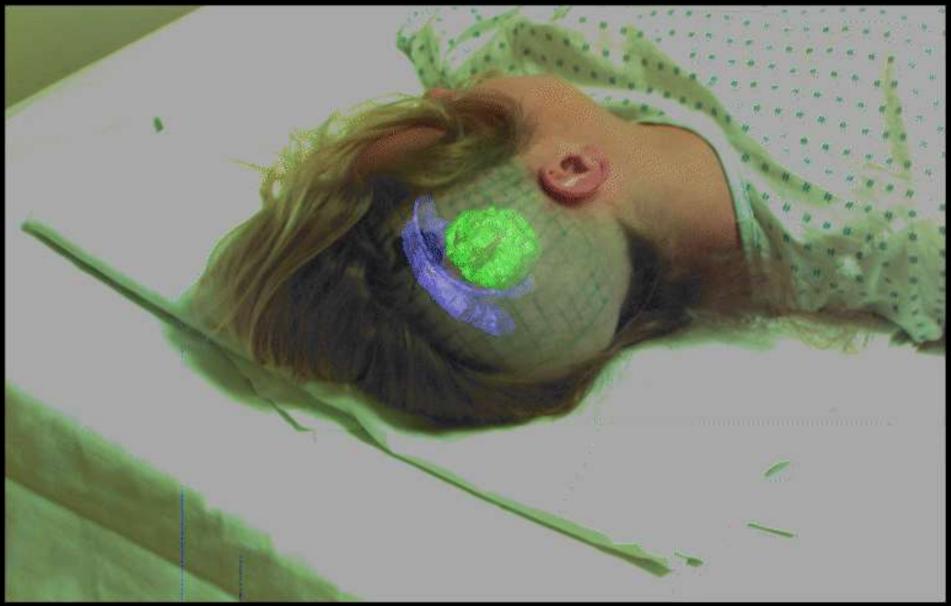
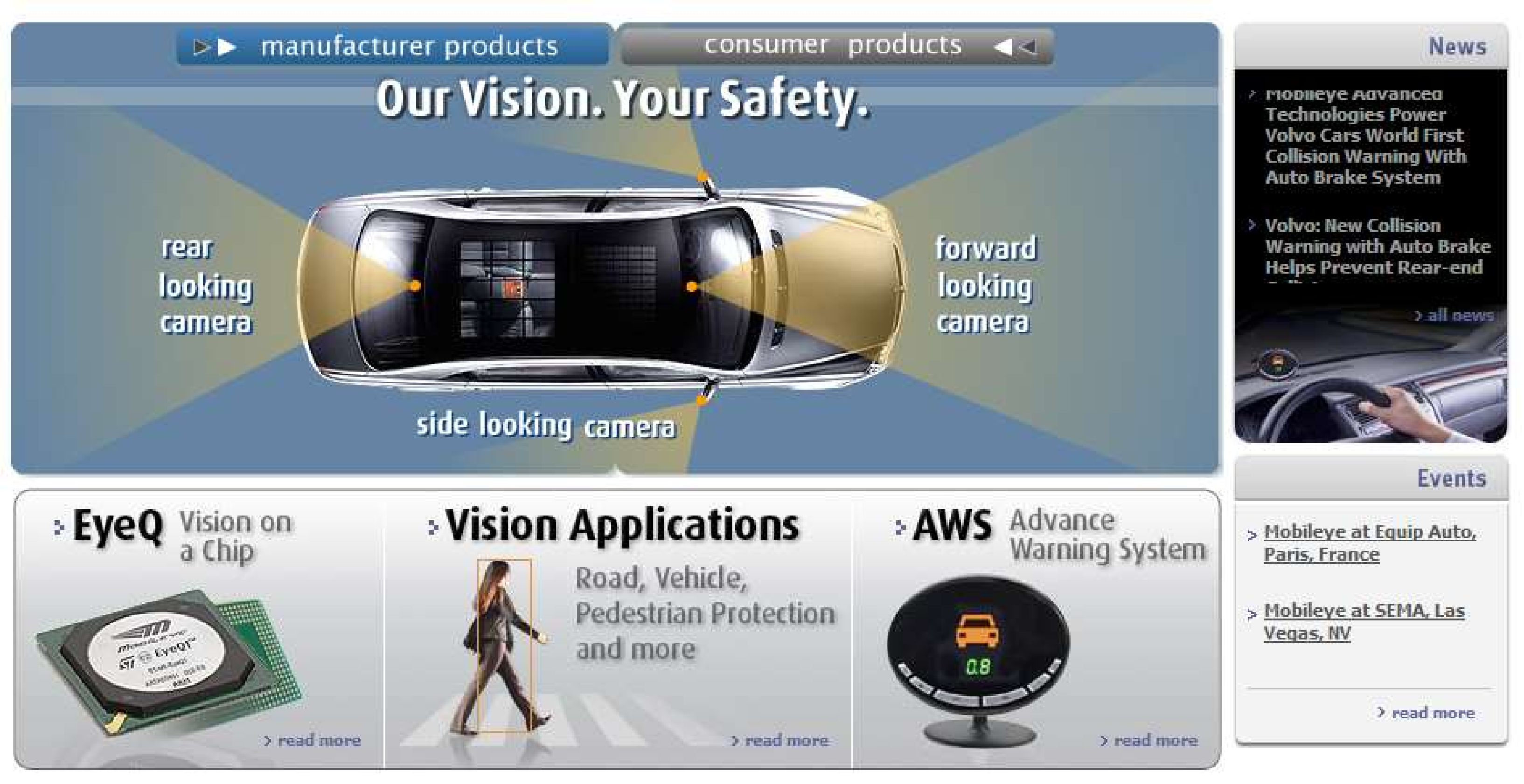


Image guided surgery
Grimson et al., MIT

Smart cars

Slide content courtesy of Amnon Shashua



News

- Mobileye Advanced Technologies Power Volvo Cars World First Collision Warning With Auto Brake System
- Volvo: New Collision Warning with Auto Brake Helps Prevent Rear-end

Events

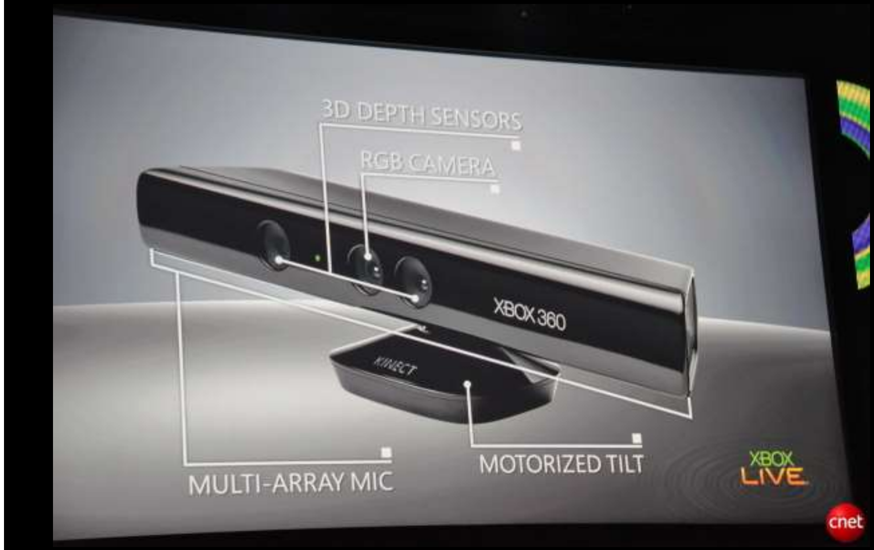
- Mobileye at Equip Auto, Paris, France
- Mobileye at SEMA, Las Vegas, NV

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

The Waymo autonomous car

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

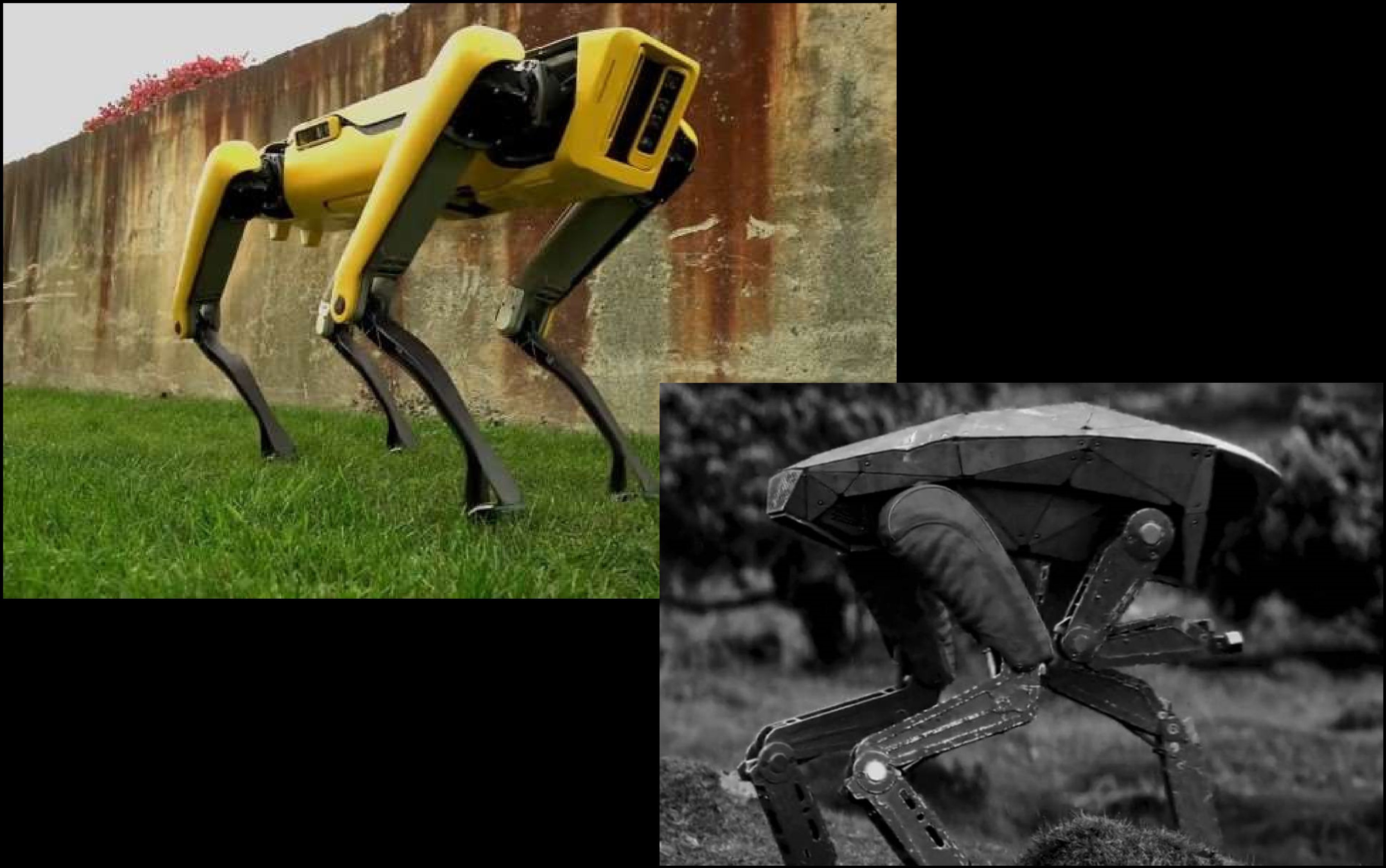


Robots

Vision-guided robots position nut runners on wheels

The Atlas robot from Boston Dynamics (1m80, 150kg)

The SpotMini robot from Boston Dynamics



..and MetalHead from Black Mirror

Automated trucks roaming the Australian desert

© Caterpillar and http://www.nrec.ri.cmu.edu

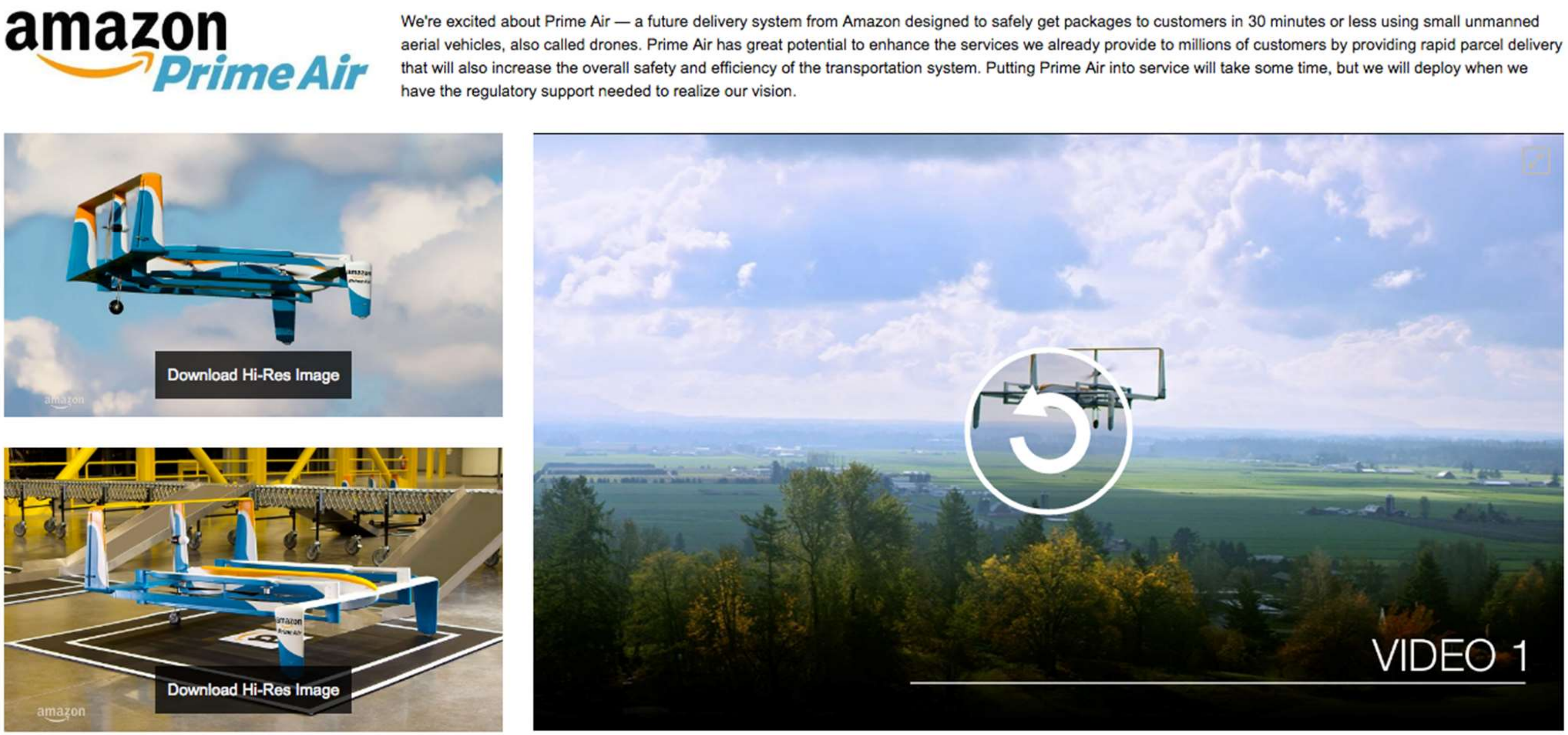
Vision in space

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.

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

Augmented Reality and Virtual Reality

Magic Leap, Oculus, Hololens, etc.

State of the art today?

With enough training data, computer vision (sometimes) 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 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, “Geometry”, 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. Range data
8. Segmentation
9. Recognition

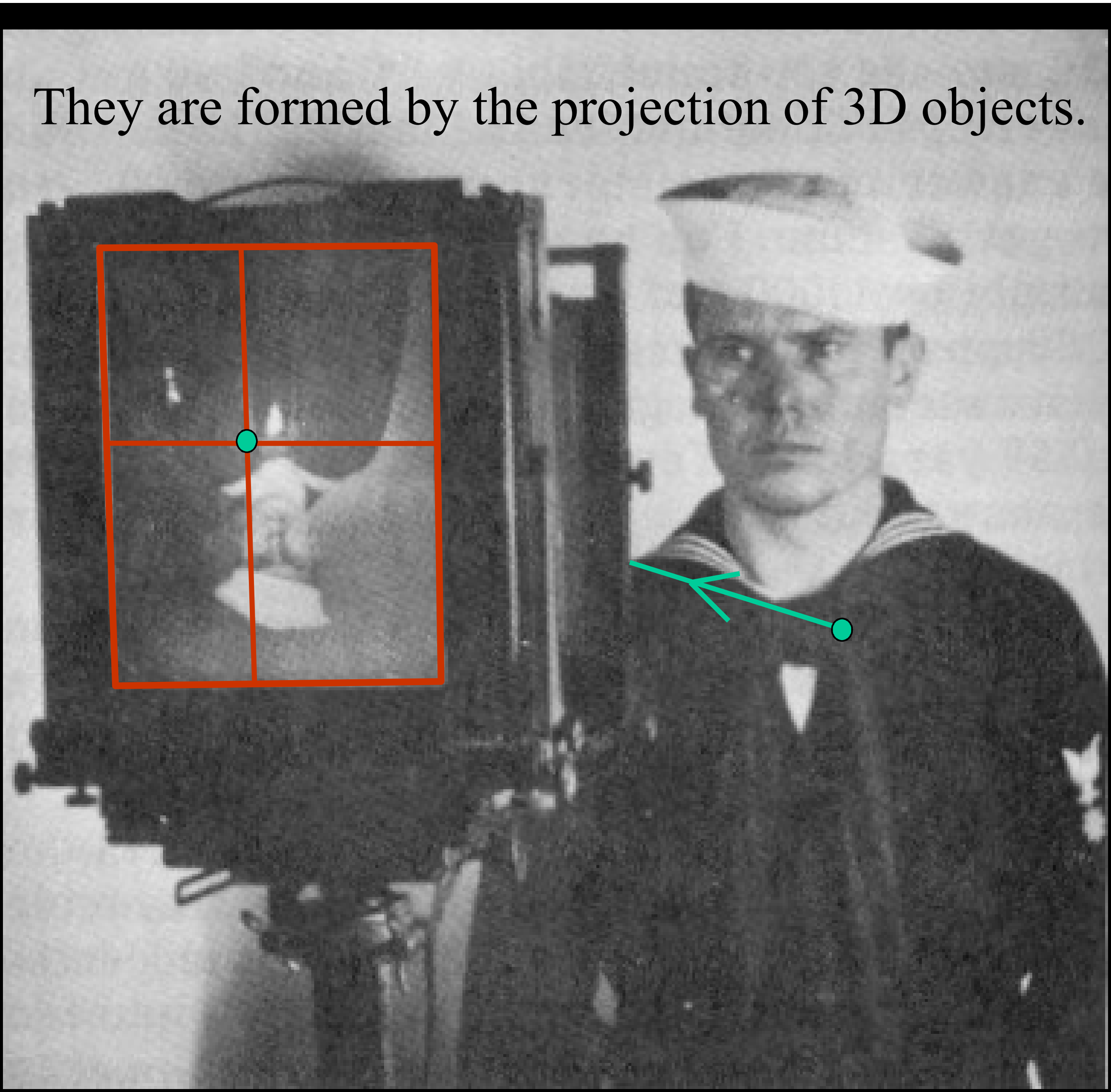
Programming assignments + final presentation

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. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

They are formed by the projection of 3D objects.



Images are two-dimensional patterns of brightness values.

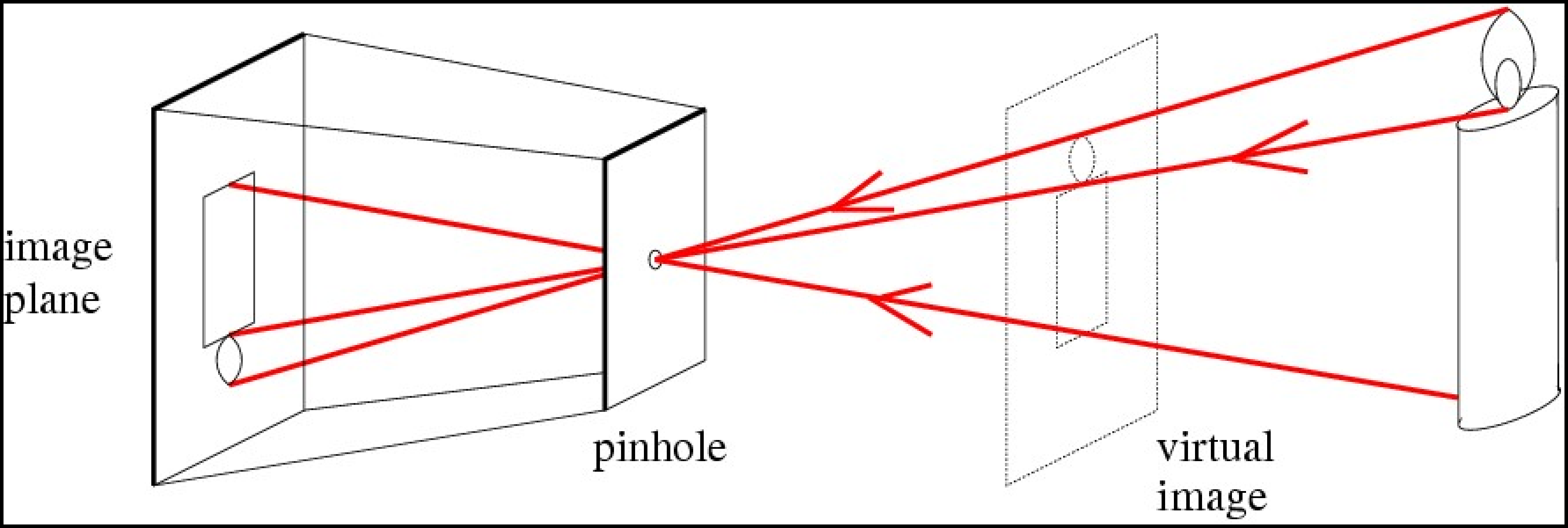
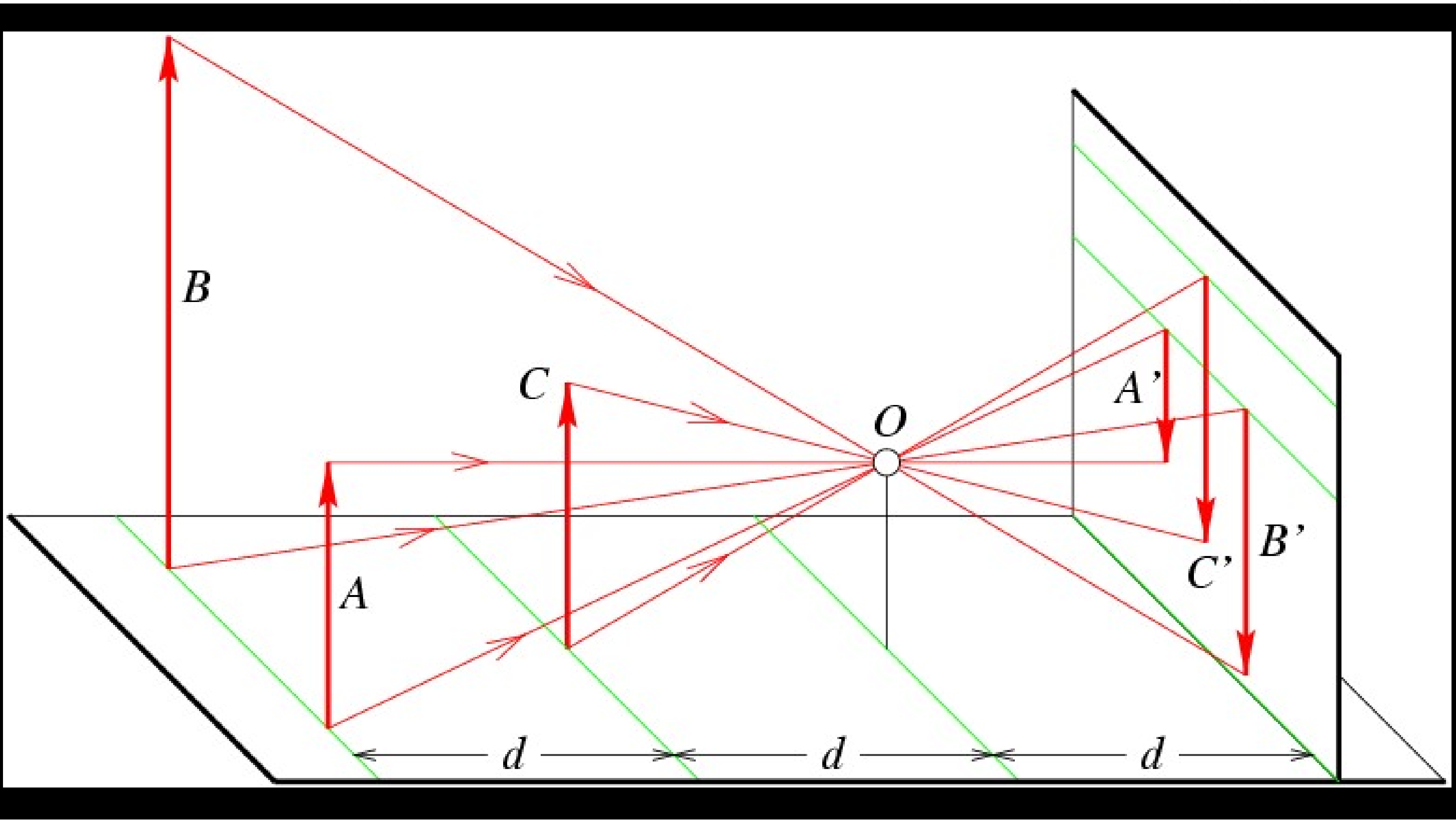


image plane
pinhole
virtual image

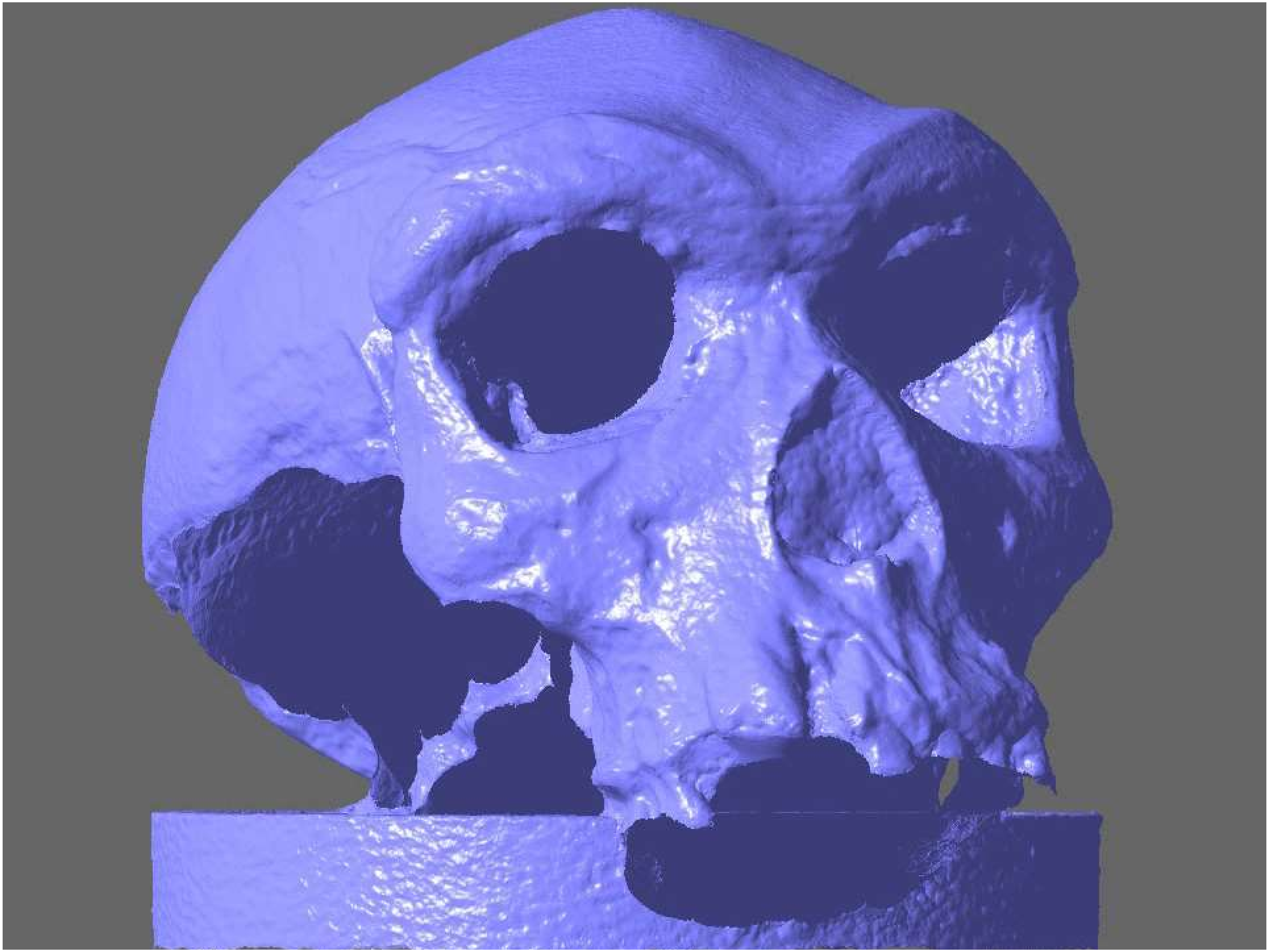

B
C
O
A’
A
C’
B’
d
d
d

High-fidelity multi-view stereopsis

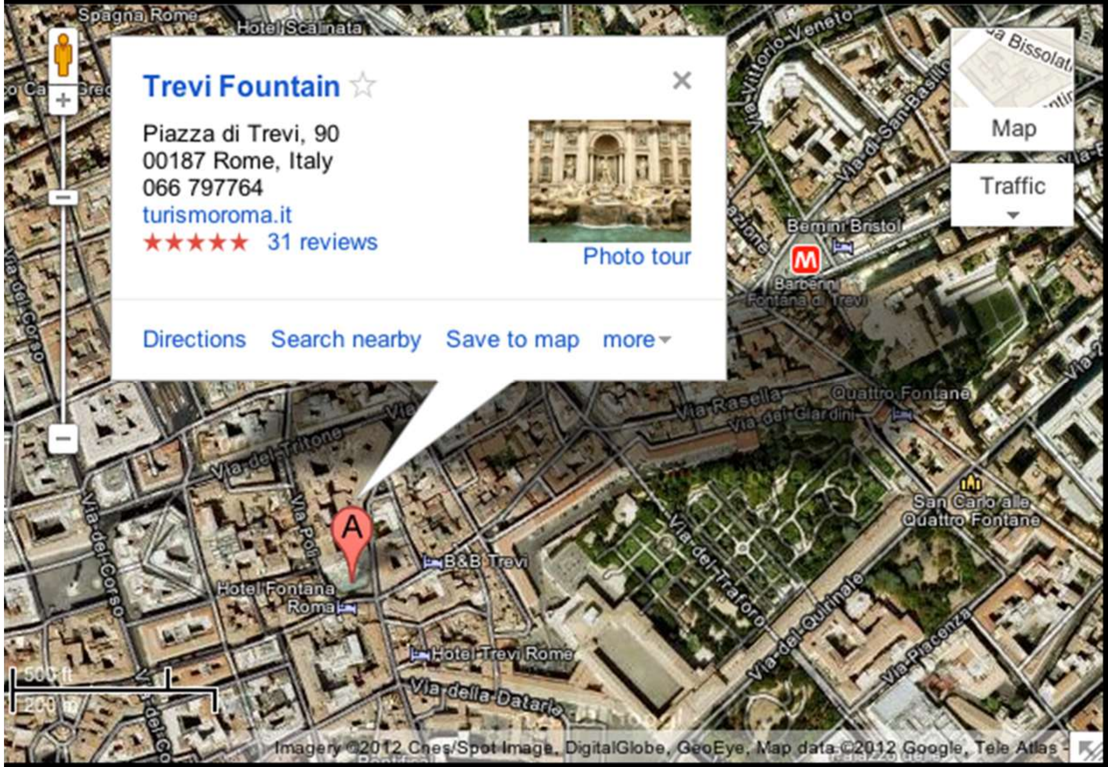
(Furukawa and Ponce, CVPR'07,PAMI'10)

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)


Trevi Fountain ☆

Piazza di Trevi, 90
00187 Rome, Italy
066 797764
turismoroma.it
★★★★★ 31 reviews

Photo tour

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

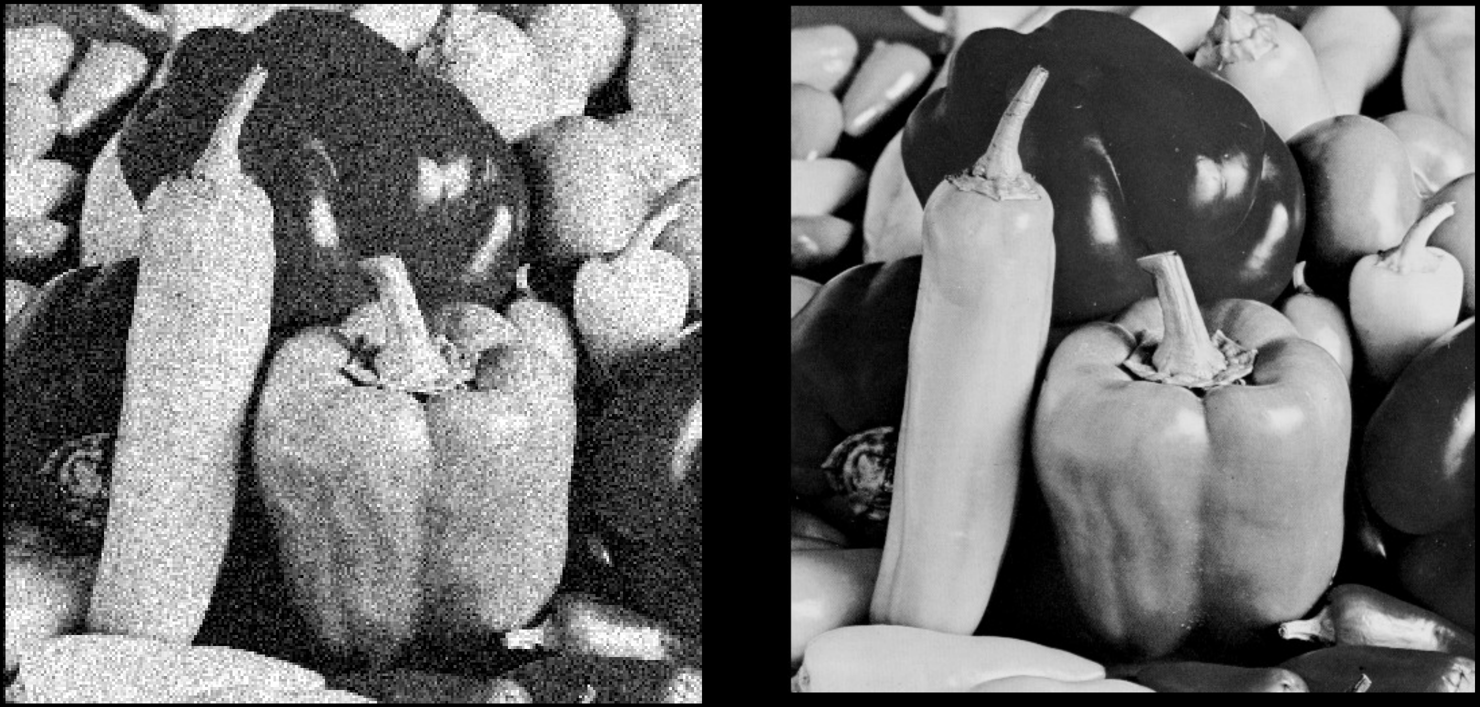
- Google Maps Photo Tour
- Lucasfilm
- Weta Digital

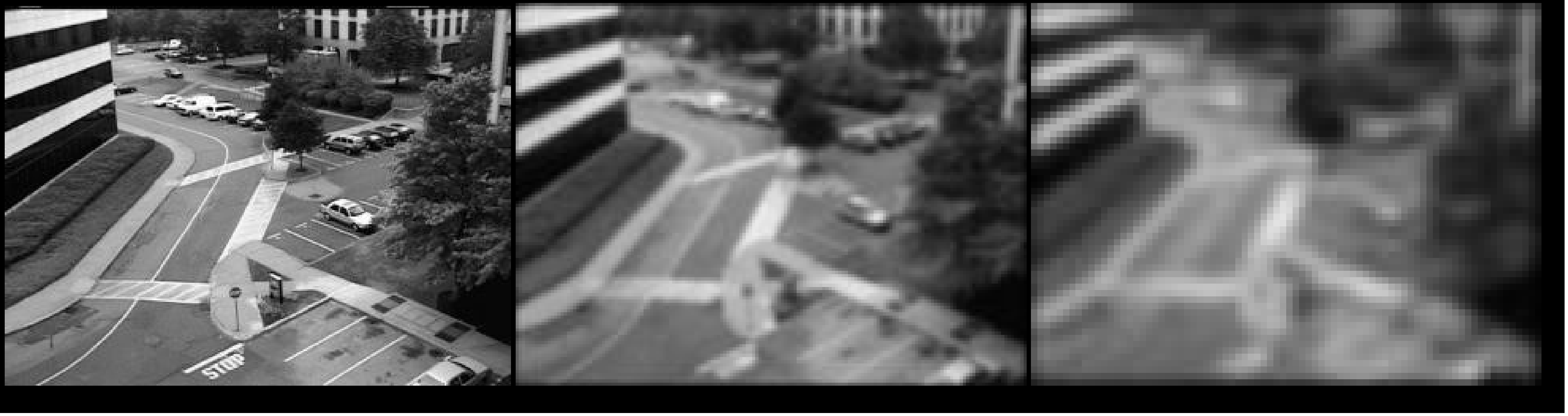
Course outline:

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. Range data
8. Segmentation
9. Recognition

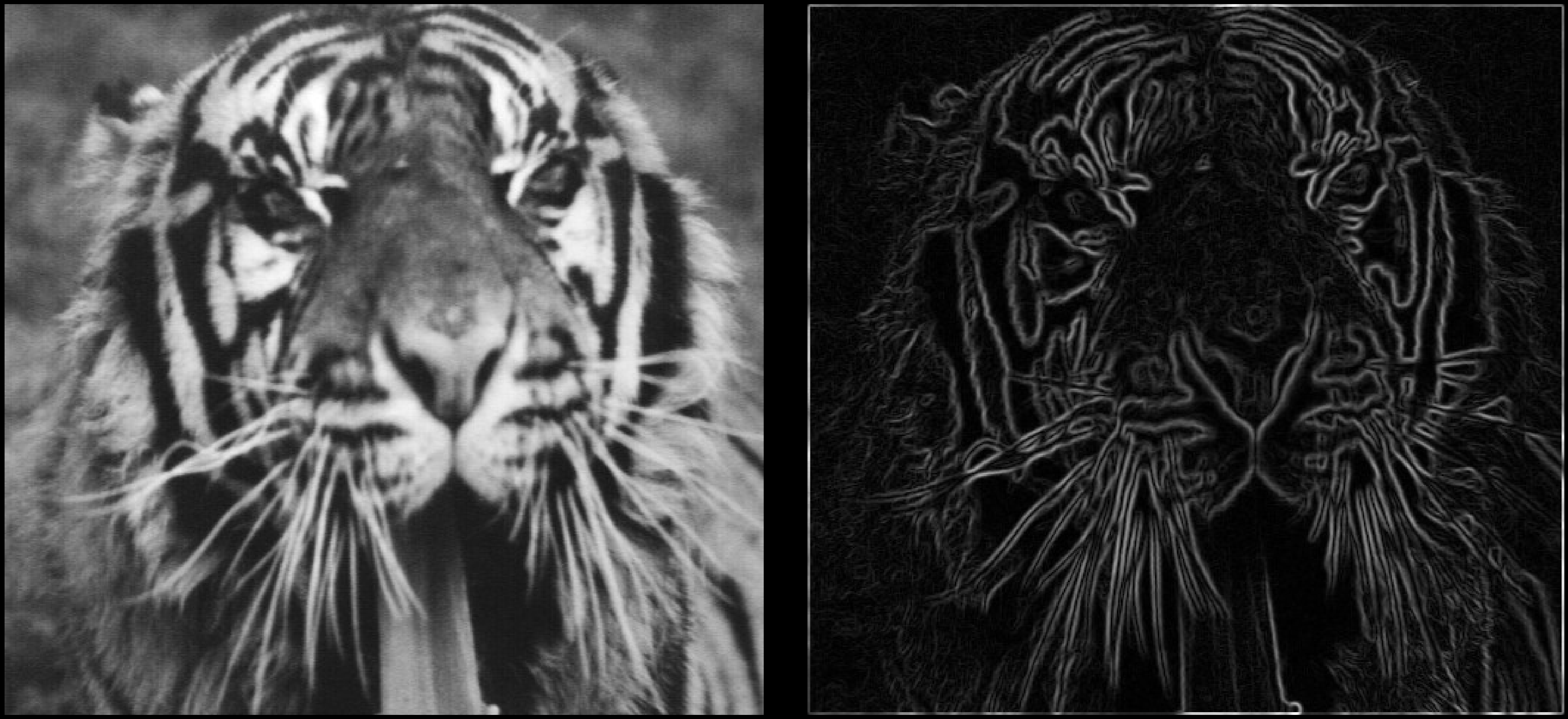
Programming assignments + final presentation

Filtering

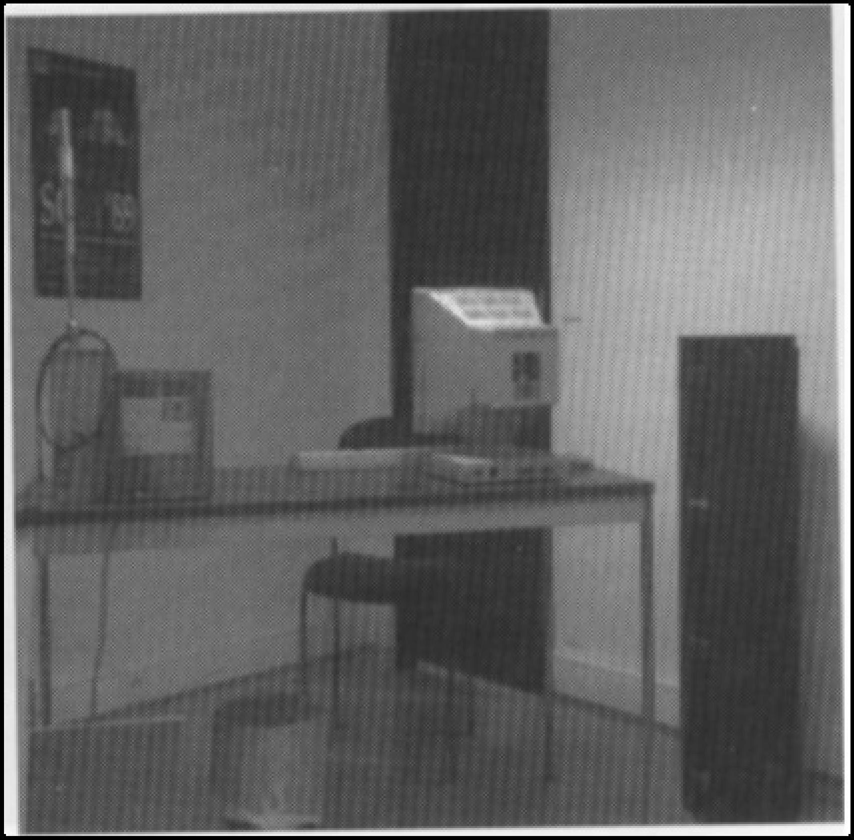


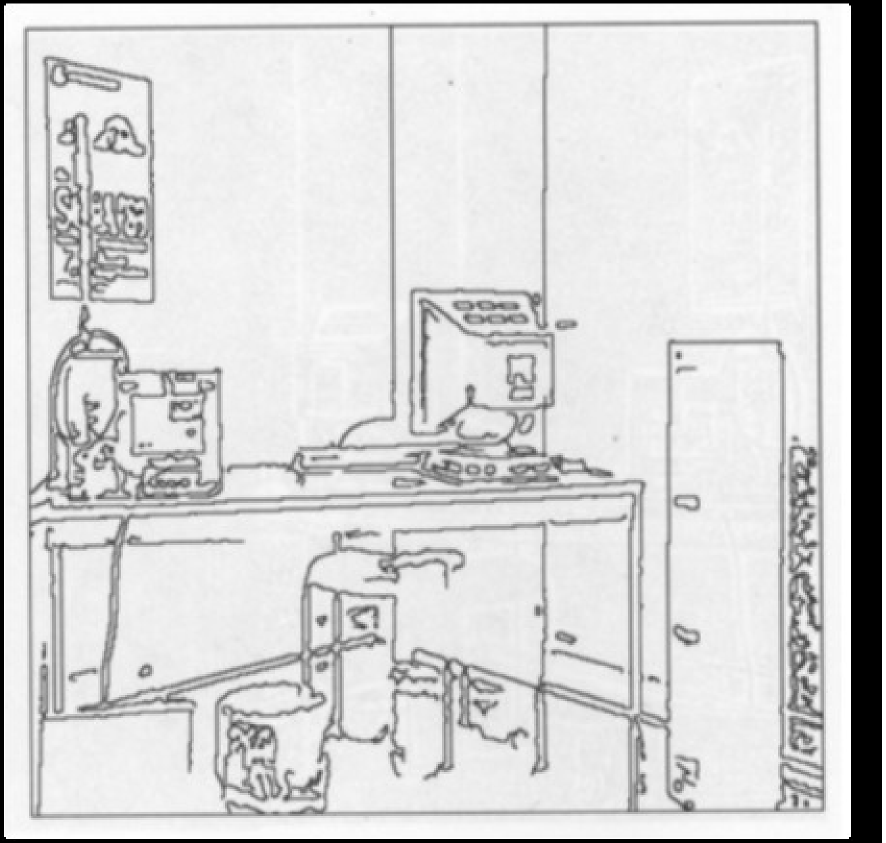


Edge Detection

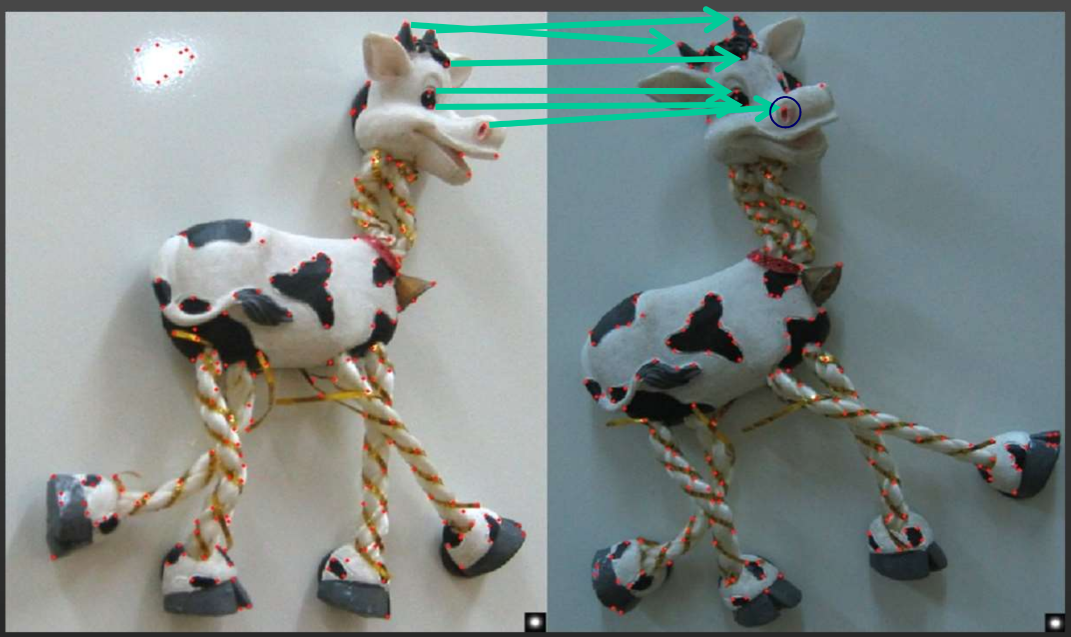


Edge Detection

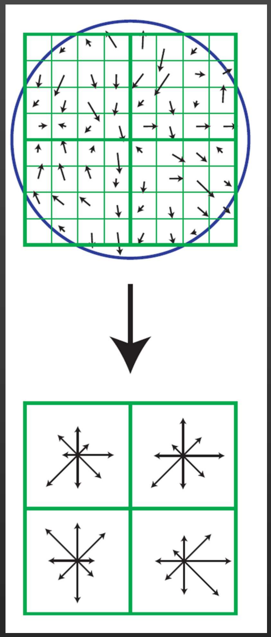




Interest points and local appearance models



(Image courtesy of C. Schmid)



(Lowe 2004)

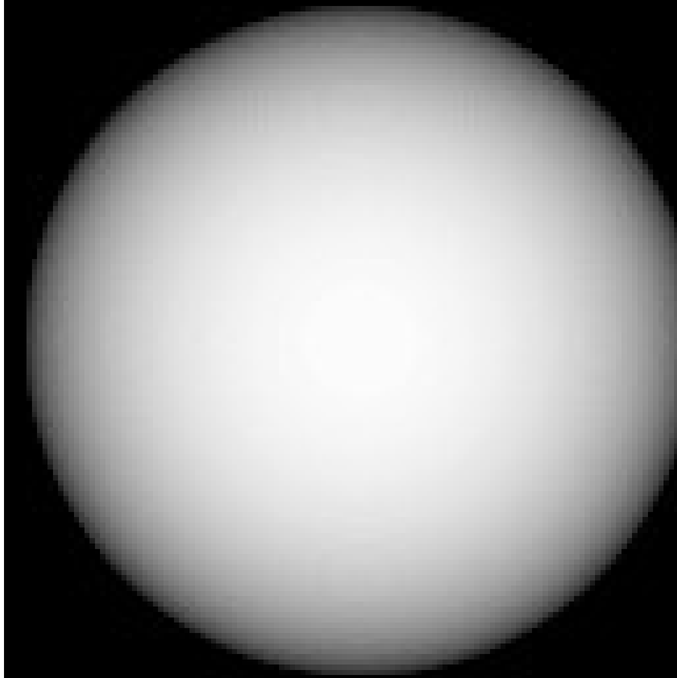
- Find features (interest points)
- Match them using local invariant descriptors (jets, SIFT)

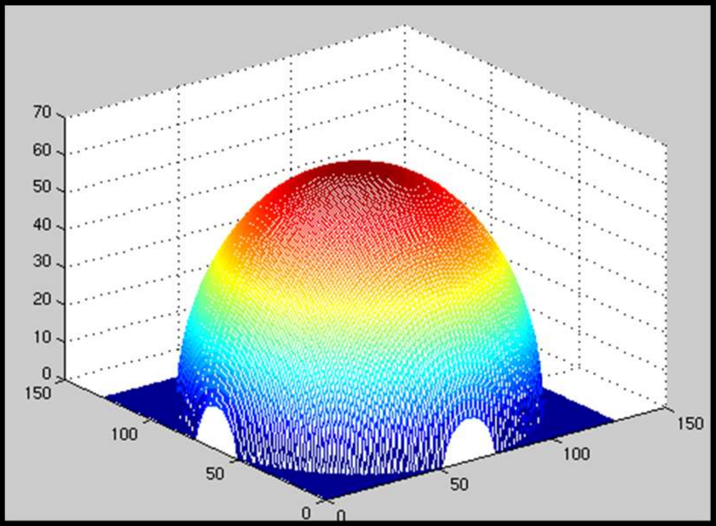
Course outline:

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. Range data
8. Segmentation
9. Recognition

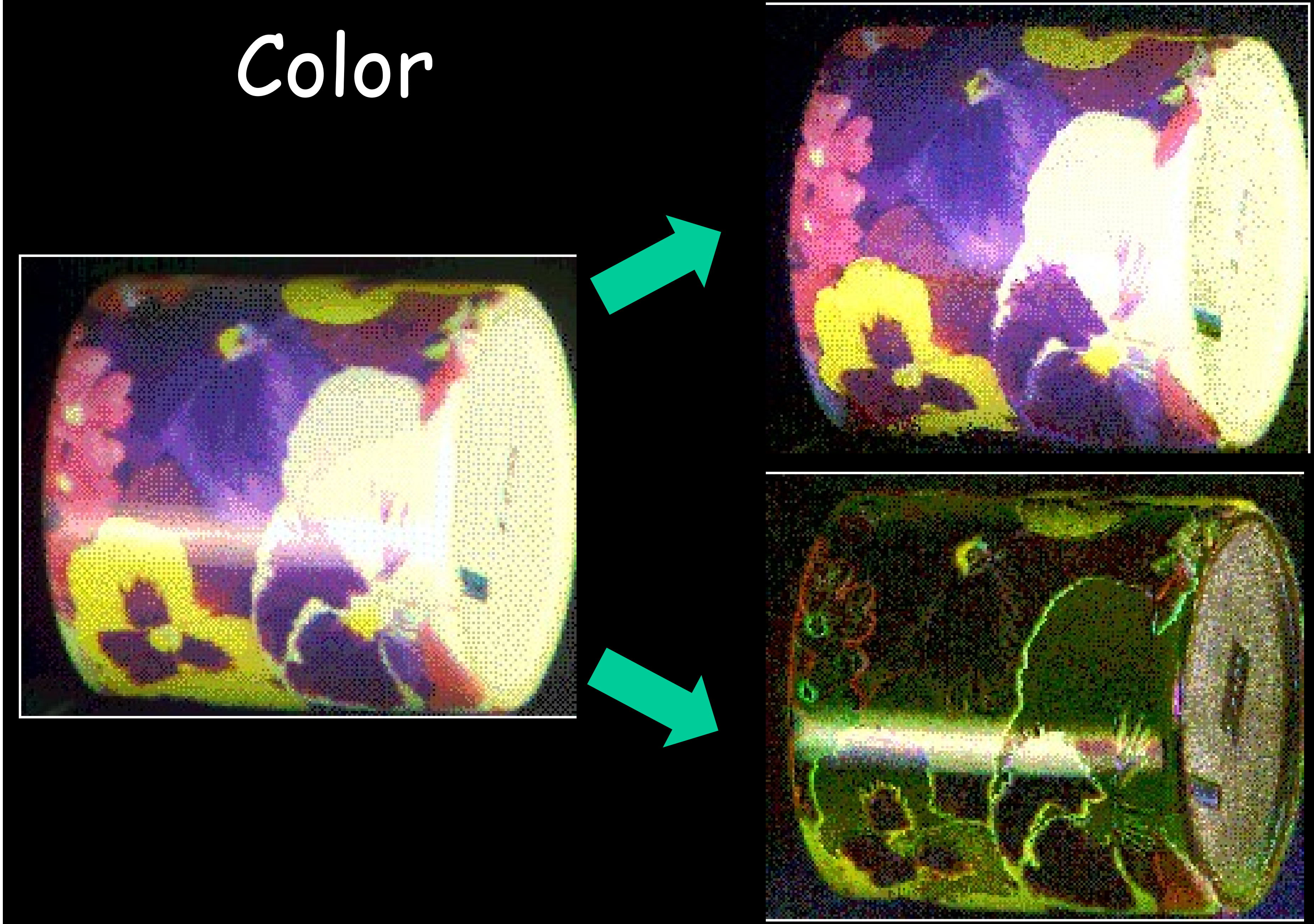
Programming assignments + final presentation

Radiometry/Shading





Color

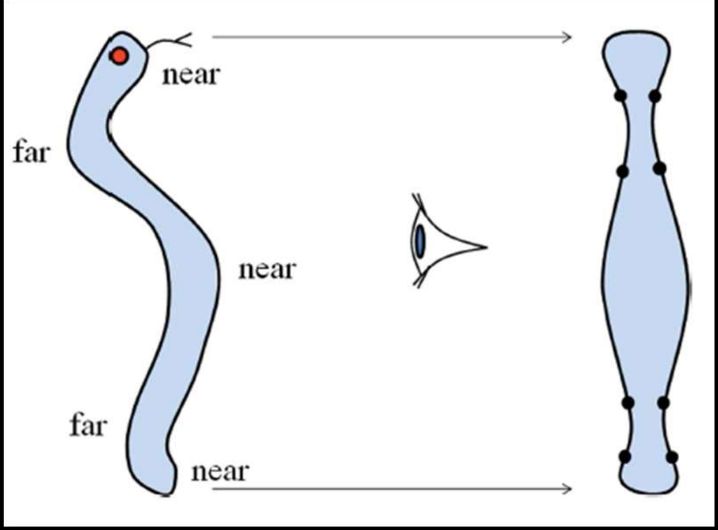


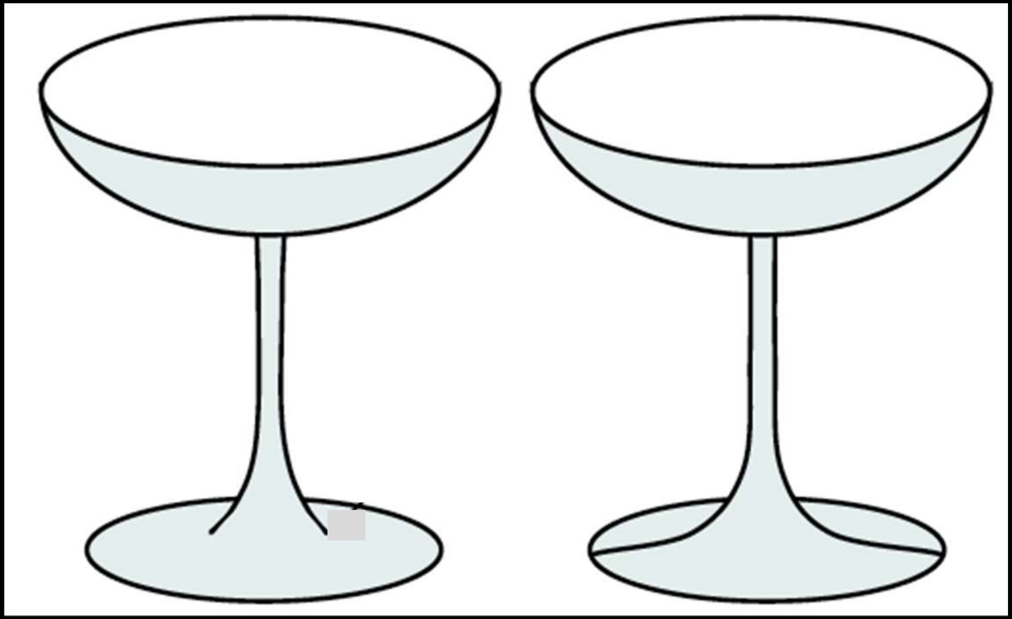
Course outline:

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. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

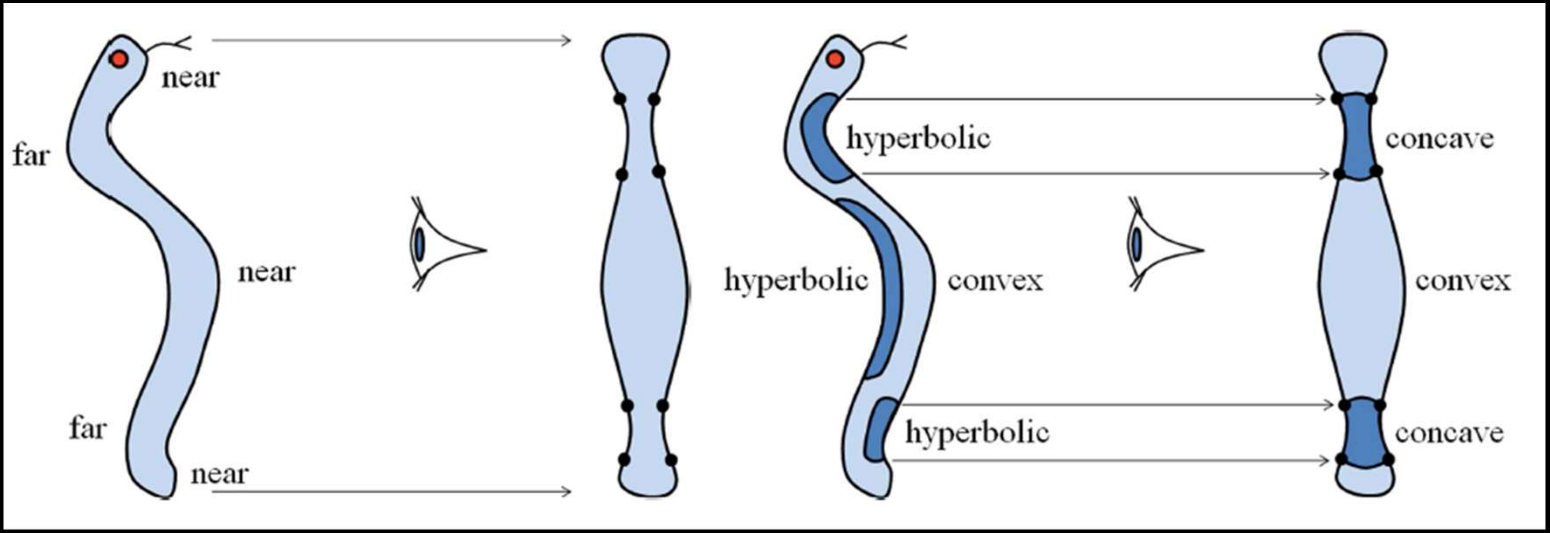
One-view (differential) geometry

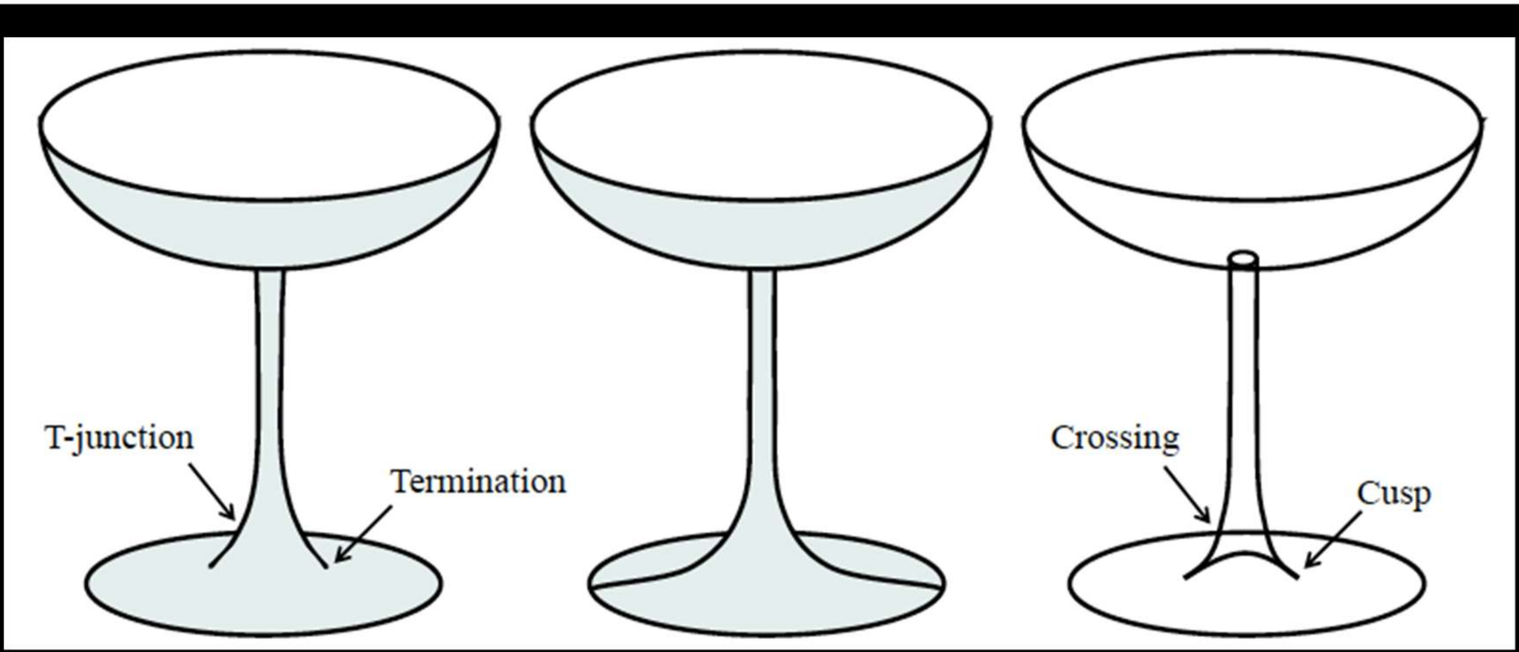




(Marr & Nishihara, 1978; Koenderink, 1984)

One-view (differential) geometry





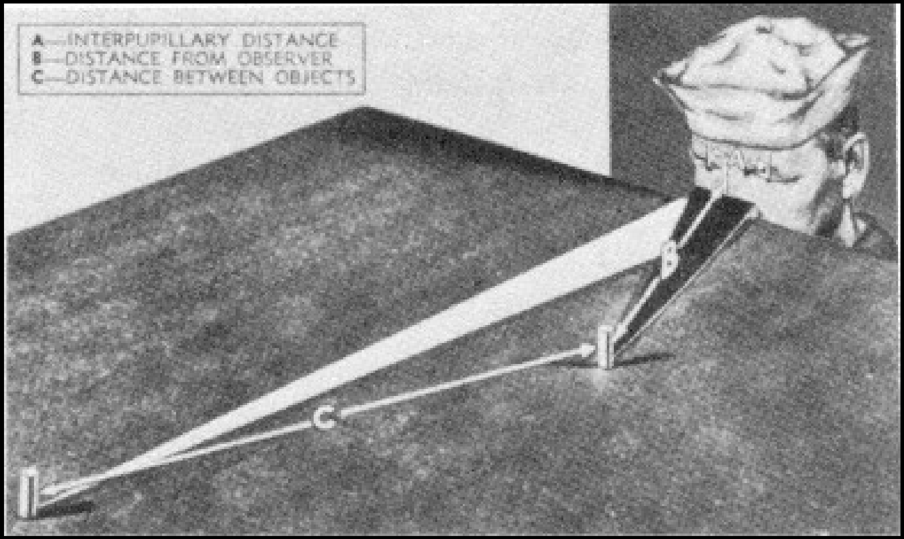
(Marr & Nishihara, 1978; Koenderink, 1984)

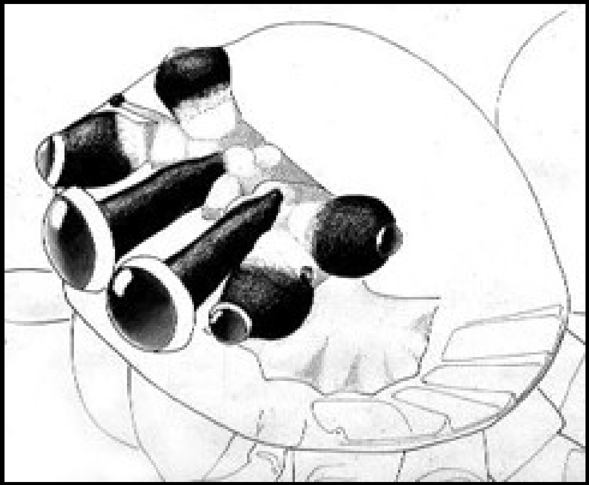
Course outline:

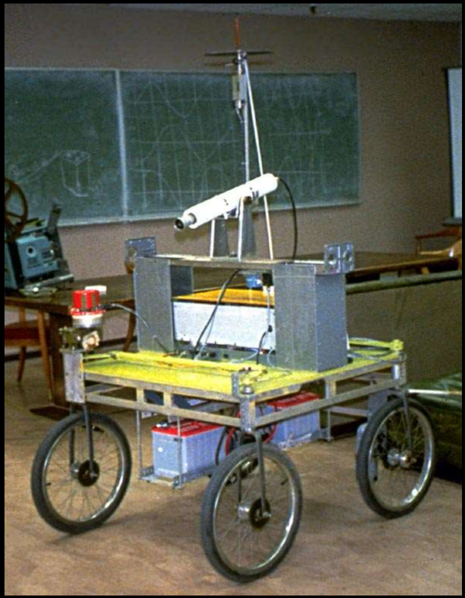
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

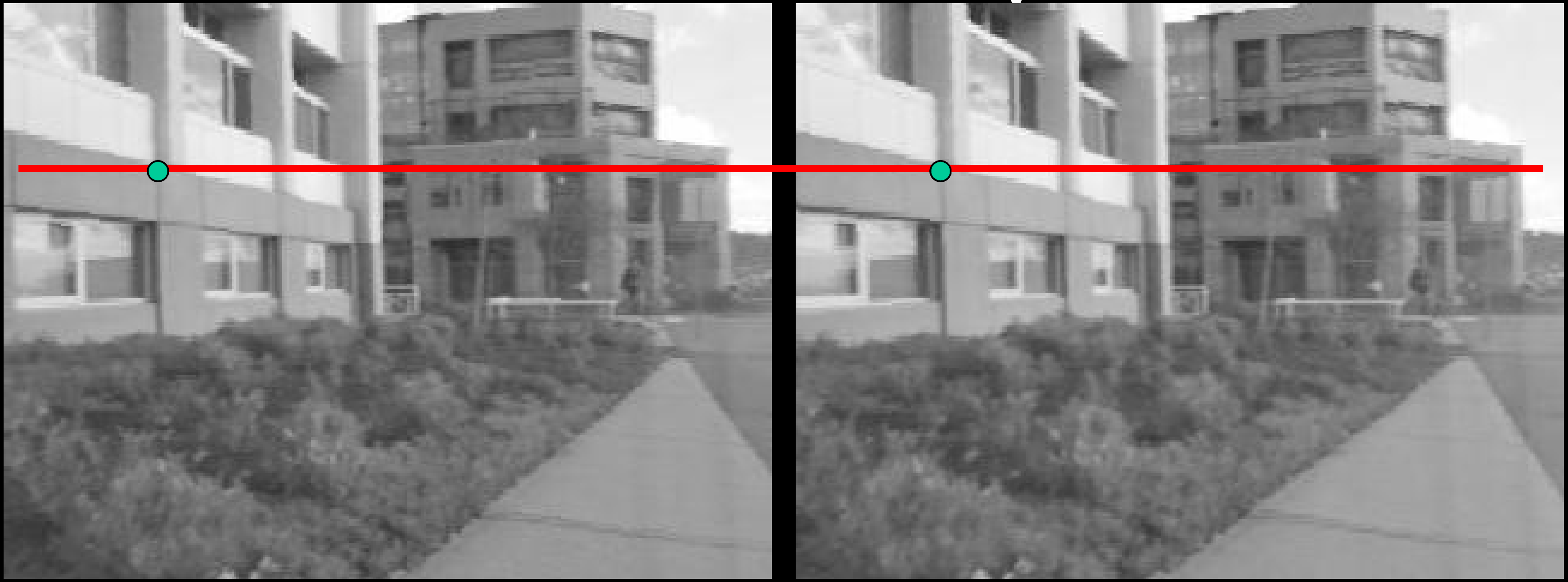
How do we perceive depth?

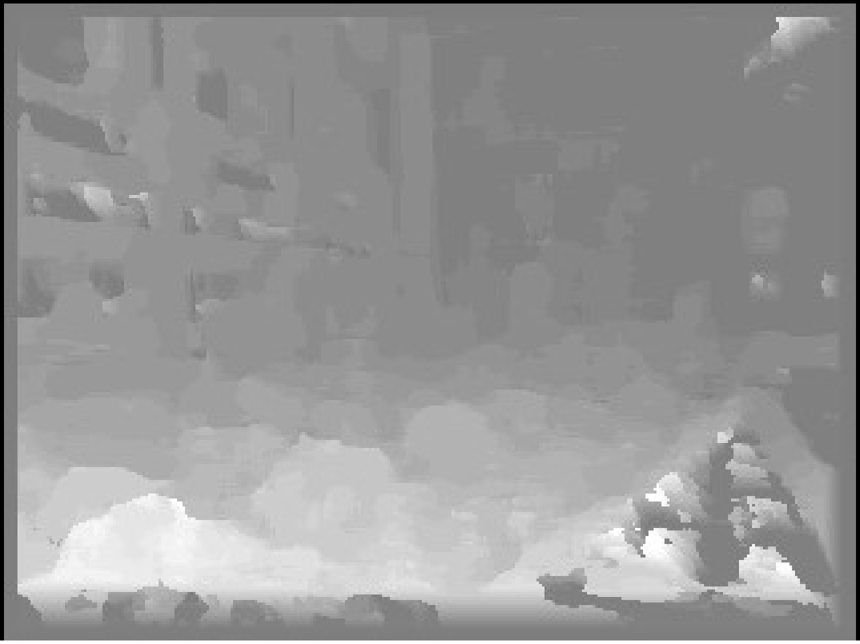






Two-View Geometry: Stereo

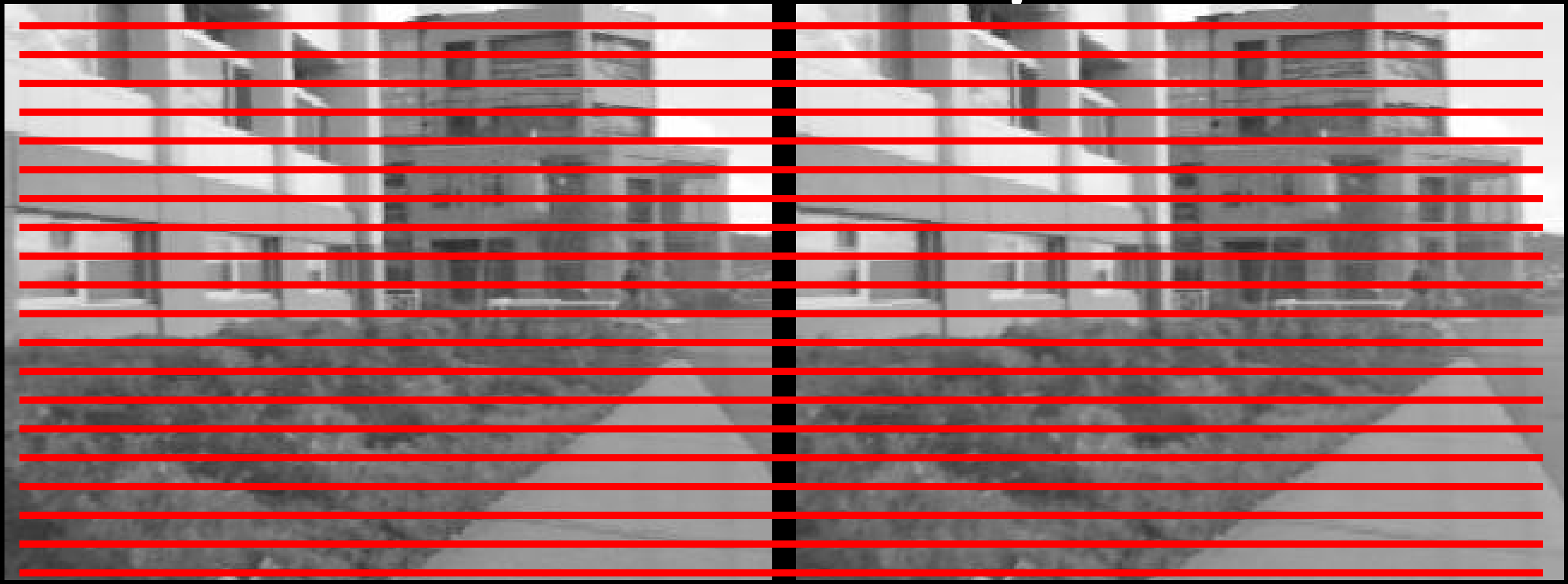




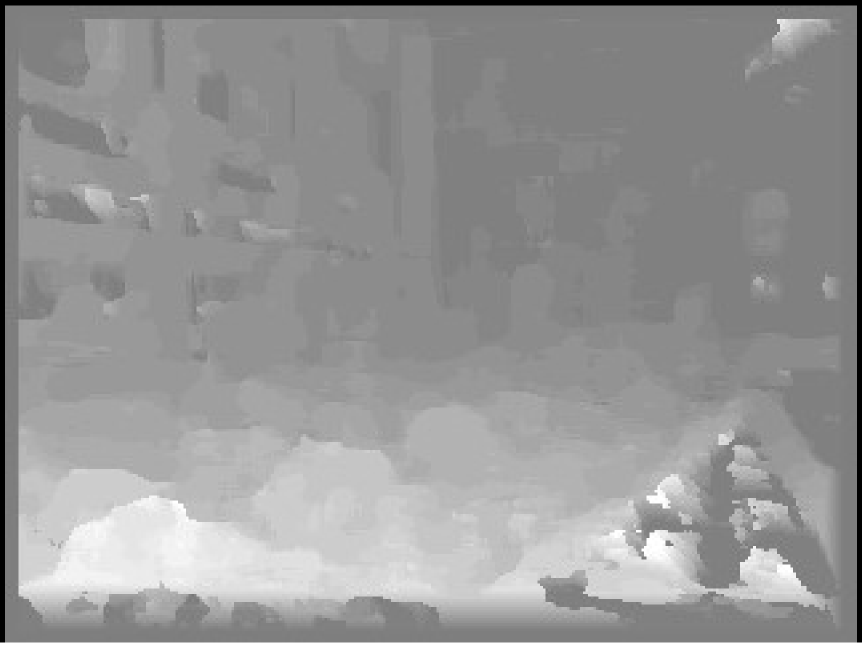
Method:

- Find correspondences
- Along epipolar lines

Two-View Geometry: Stereo



Epipolar lines for rectified cameras

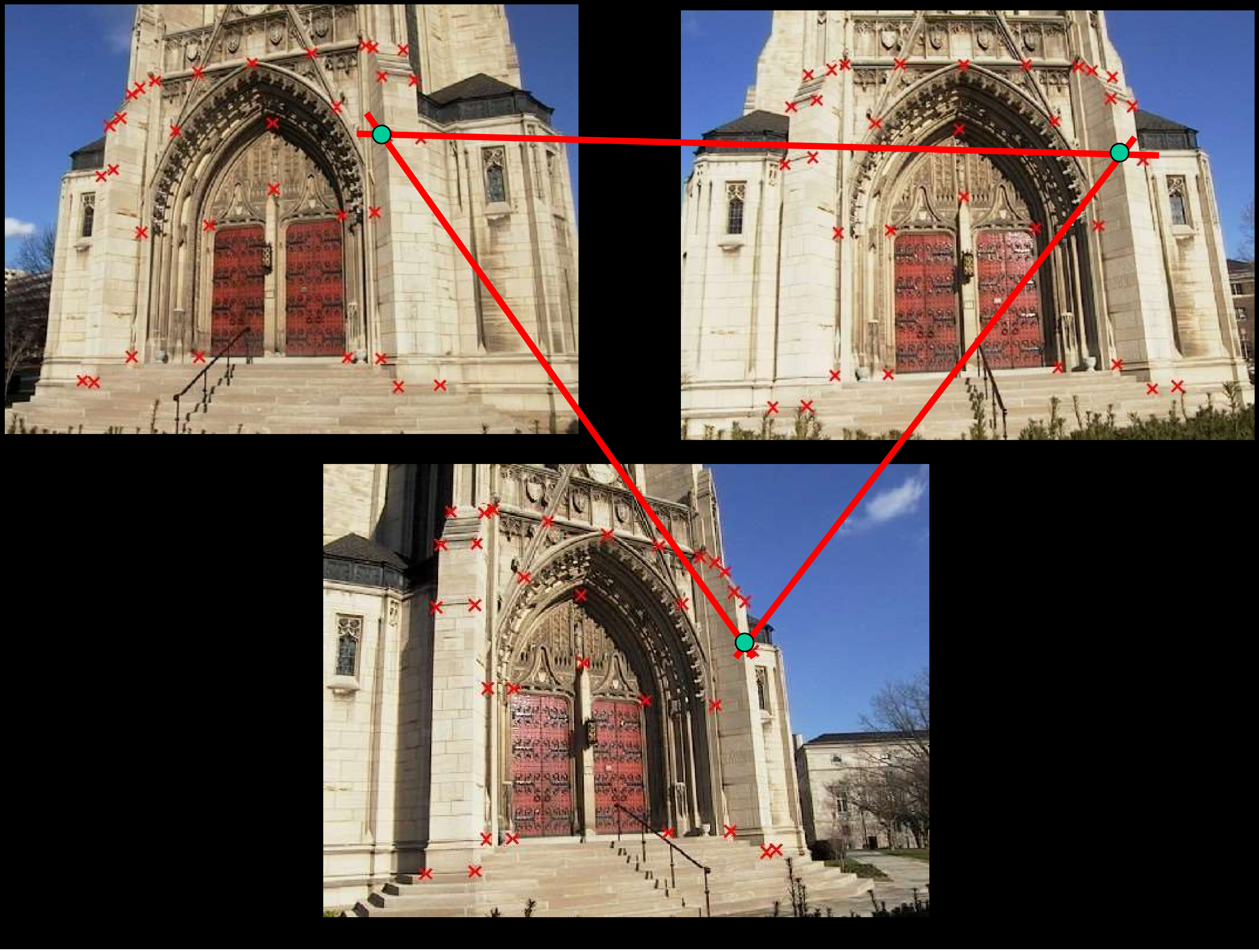


Course outline:

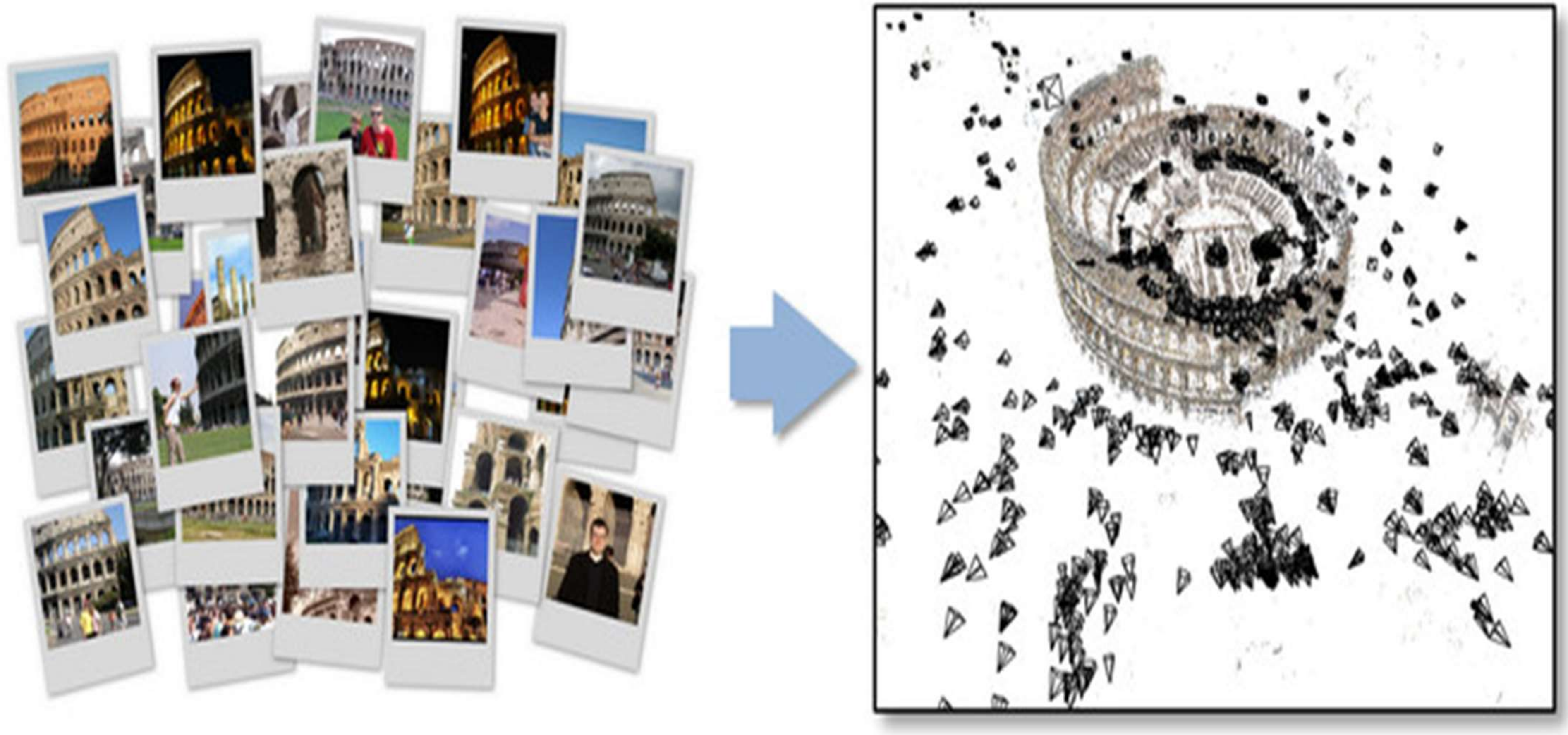
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

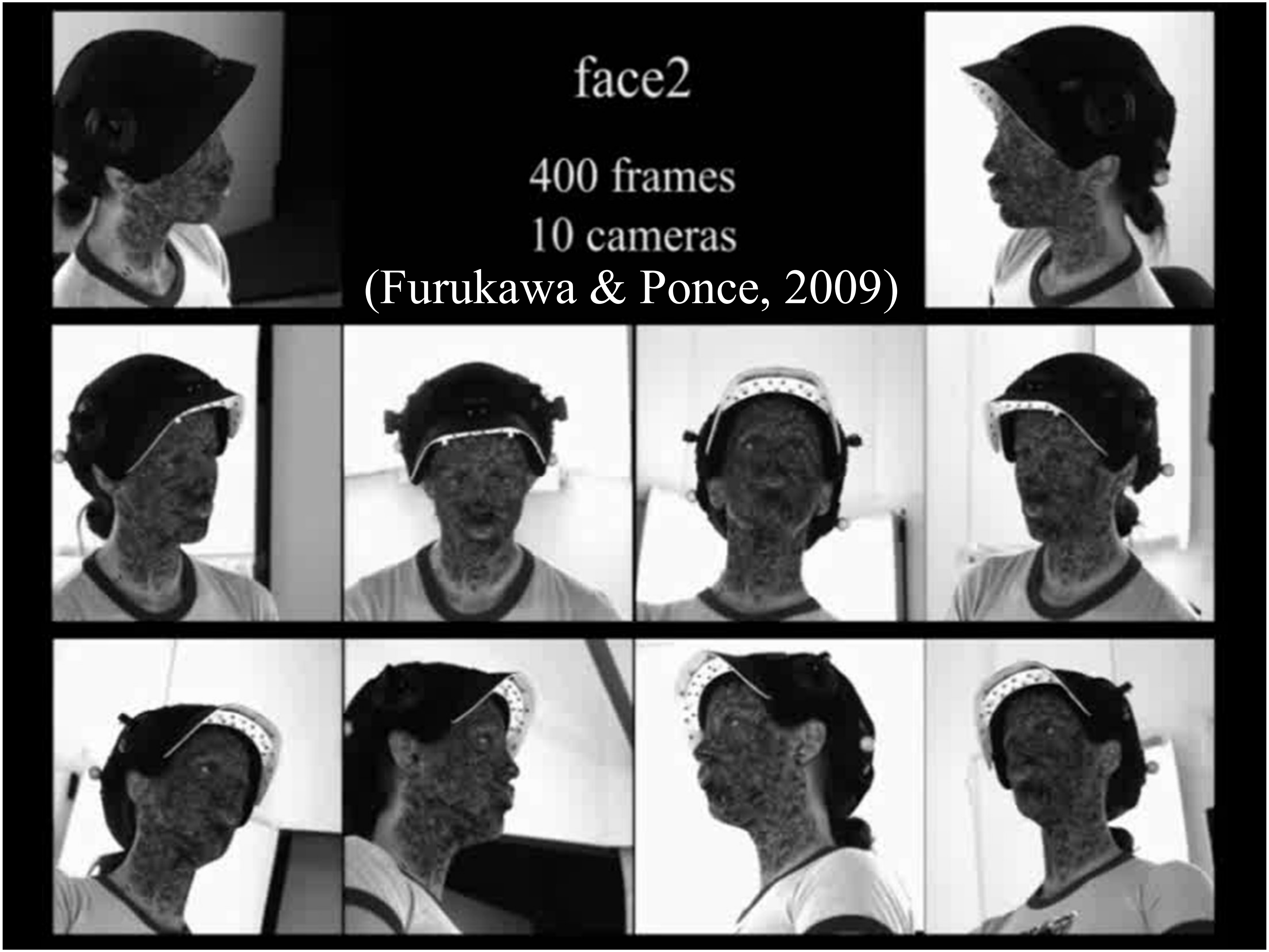
Multi-Camera Geometry



Phototourism



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



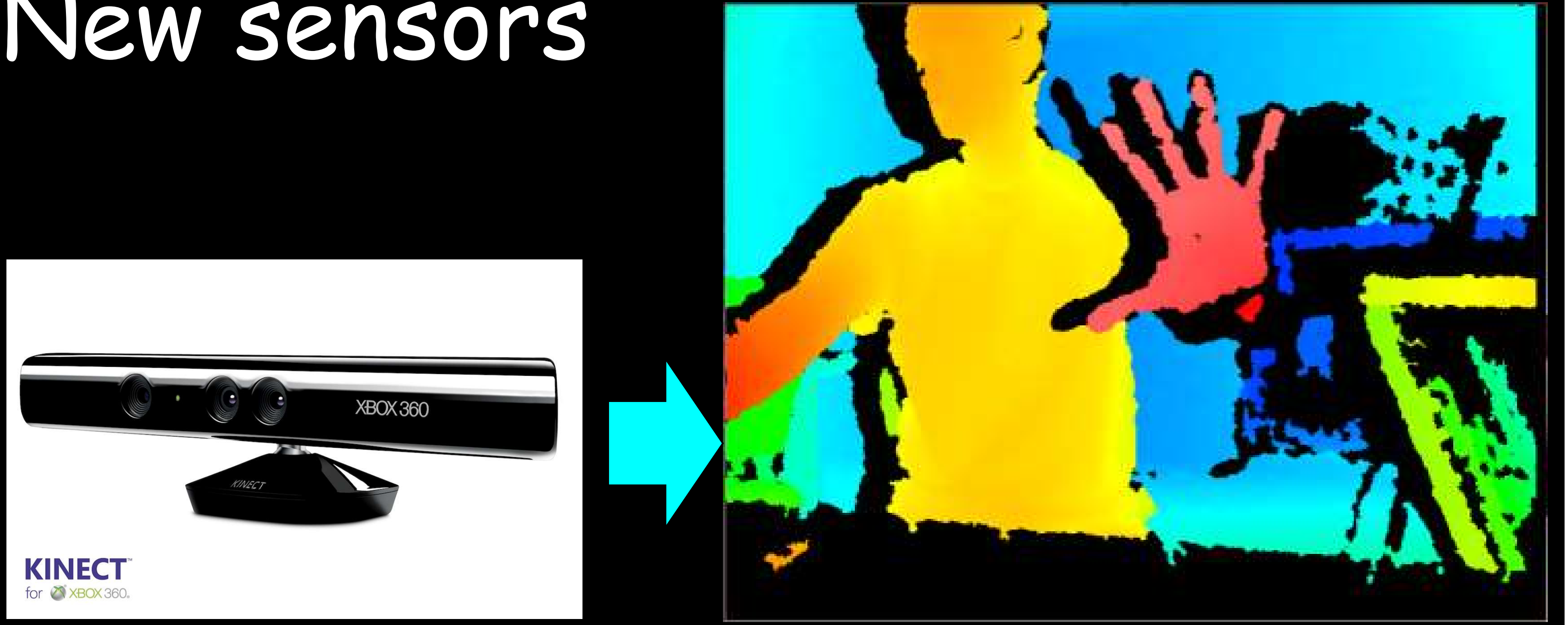
(Furukawa & Ponce, 2009)

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. Range data
8. Segmentation
9. Recognition

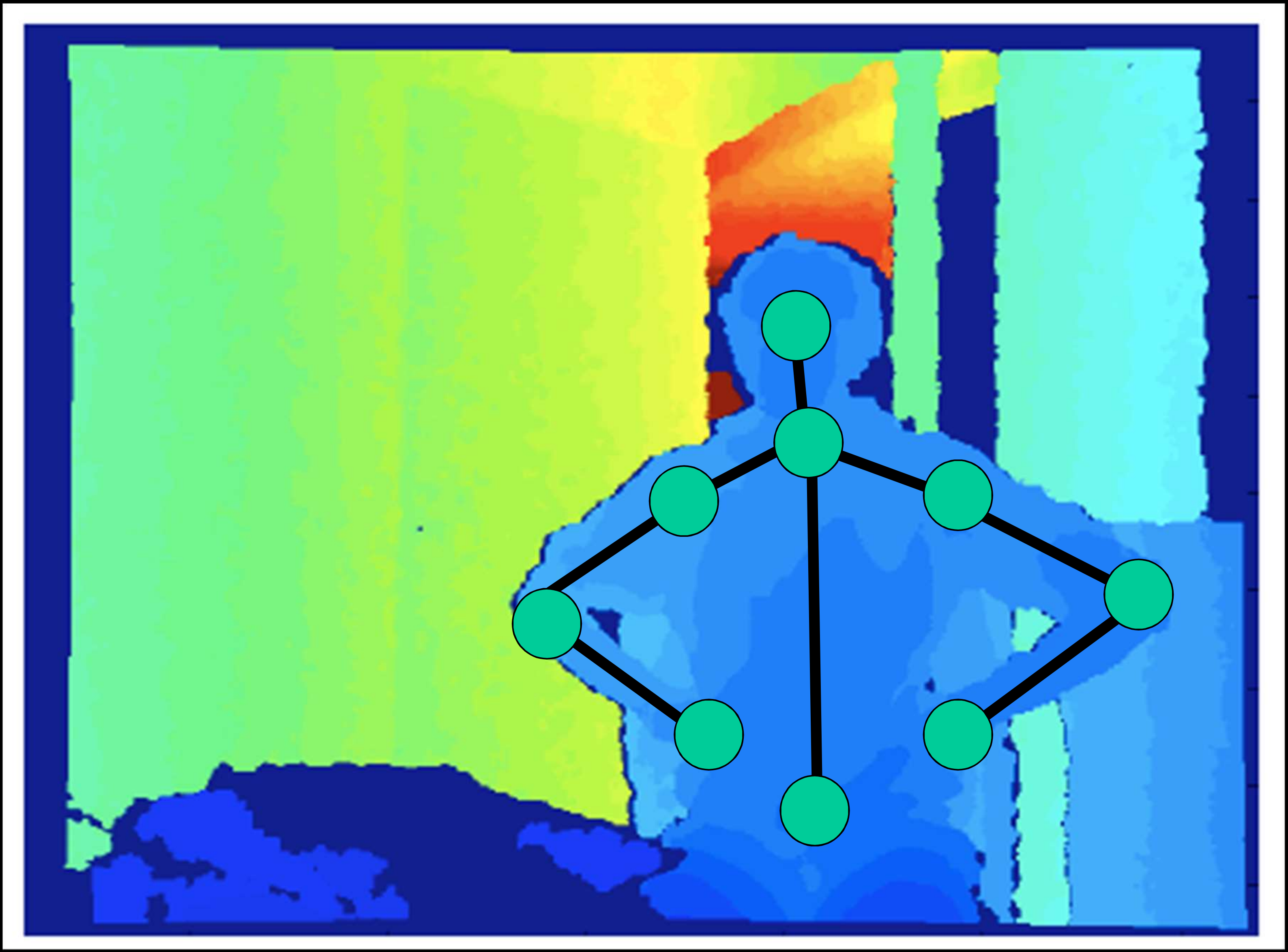
Programming assignments + final presentation

New sensors



Problem: find the 3D skeleton of people

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



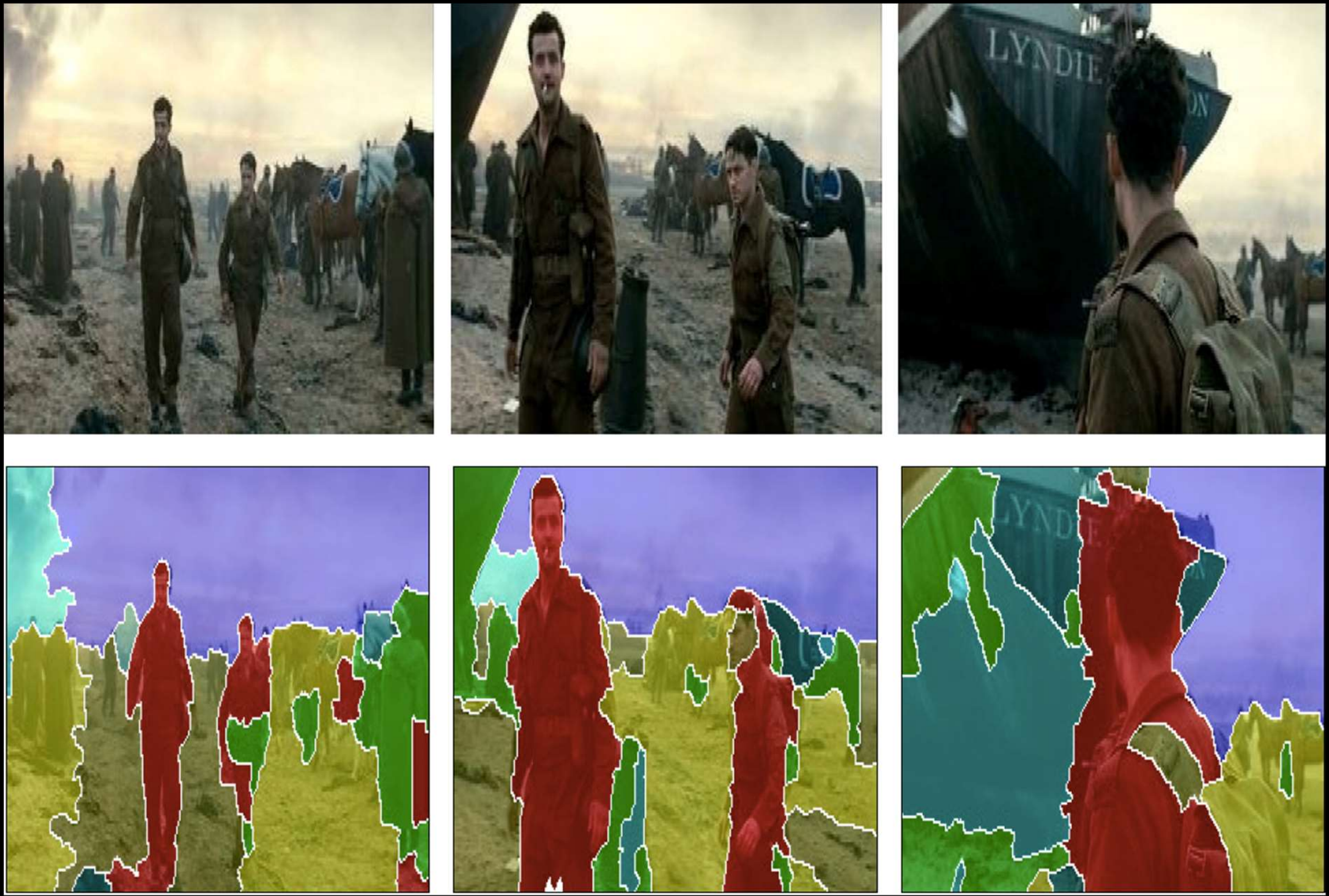
(Shotton et al., 2011)

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. Range data
8. Segmentation
9. Recognition

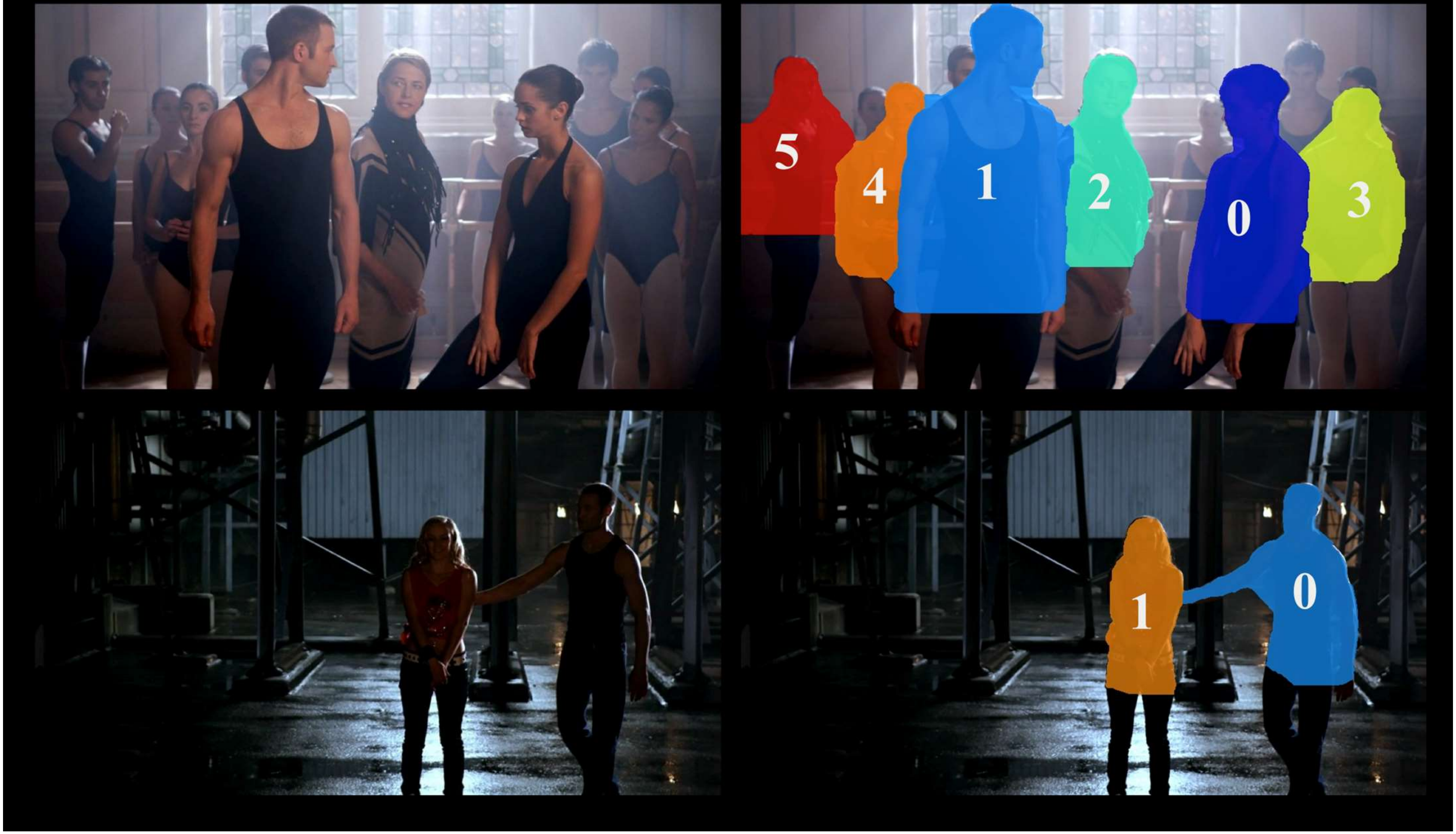
Programming assignments + final presentation

Segmentation



(Joulin, Bach, Ponce, CVPR'12)

Layered person segmentation [Seguin et al., 2015]

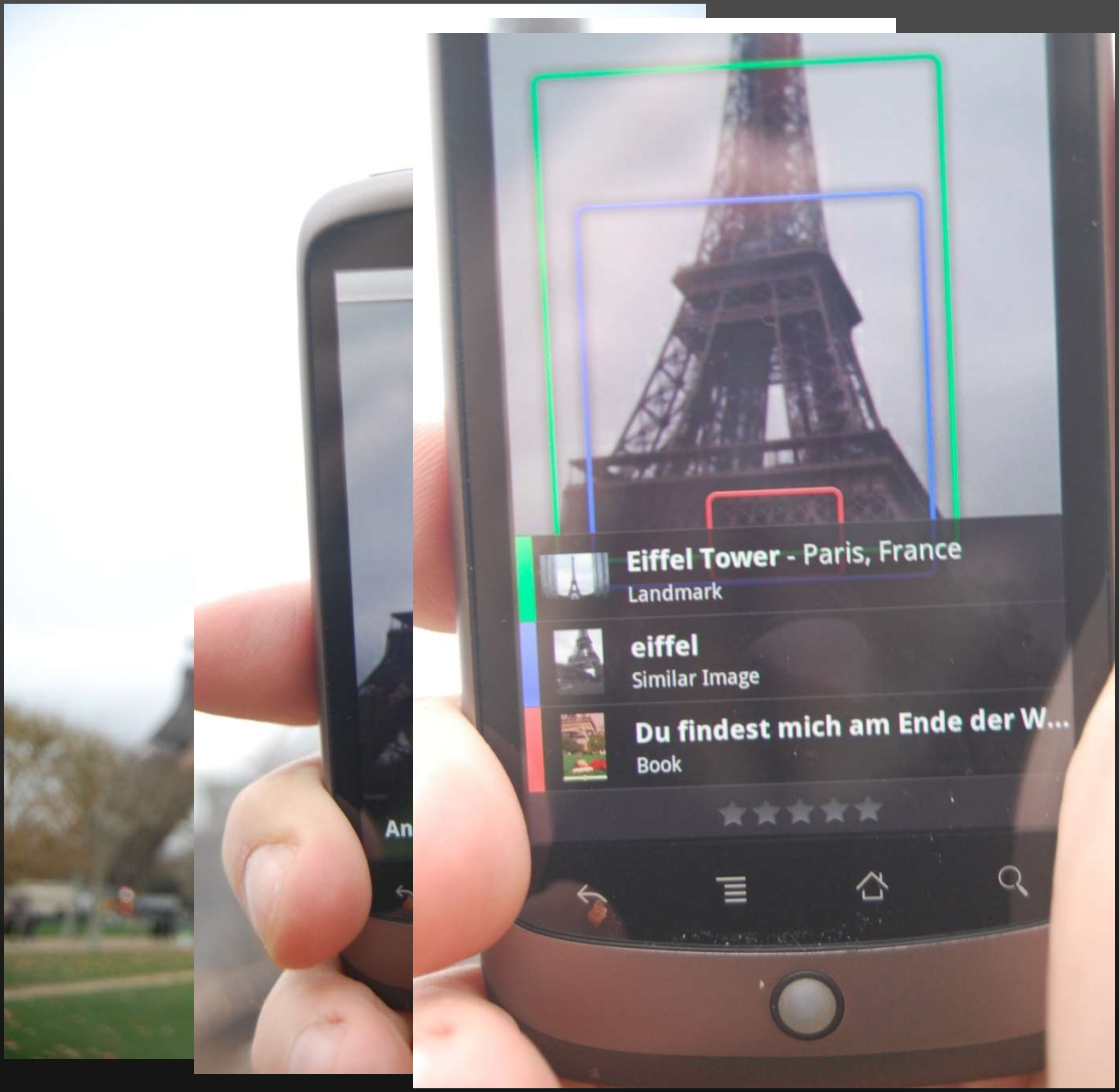


Course outline:

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. Range data
8. Segmentation
9. Recognition

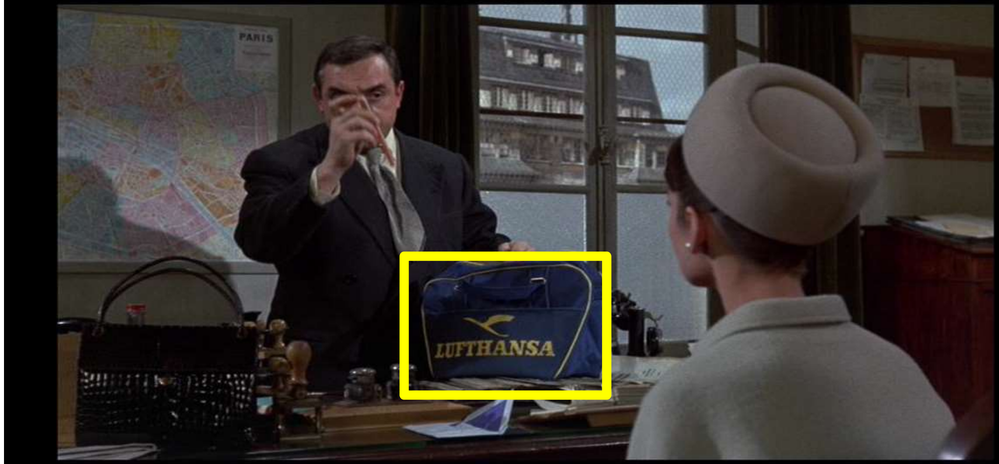
Programming assignments + final presentation

Object *instance* recognition

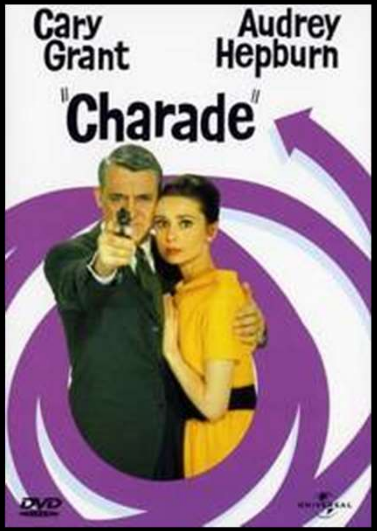

Eiffel Tower - Paris, France
Landmark
eiffel
Similar Image
Du findest mich am Ende der W...
Book
An

Example: Visual search in an entire feature length movie

Visually defined query



"Find this bag"

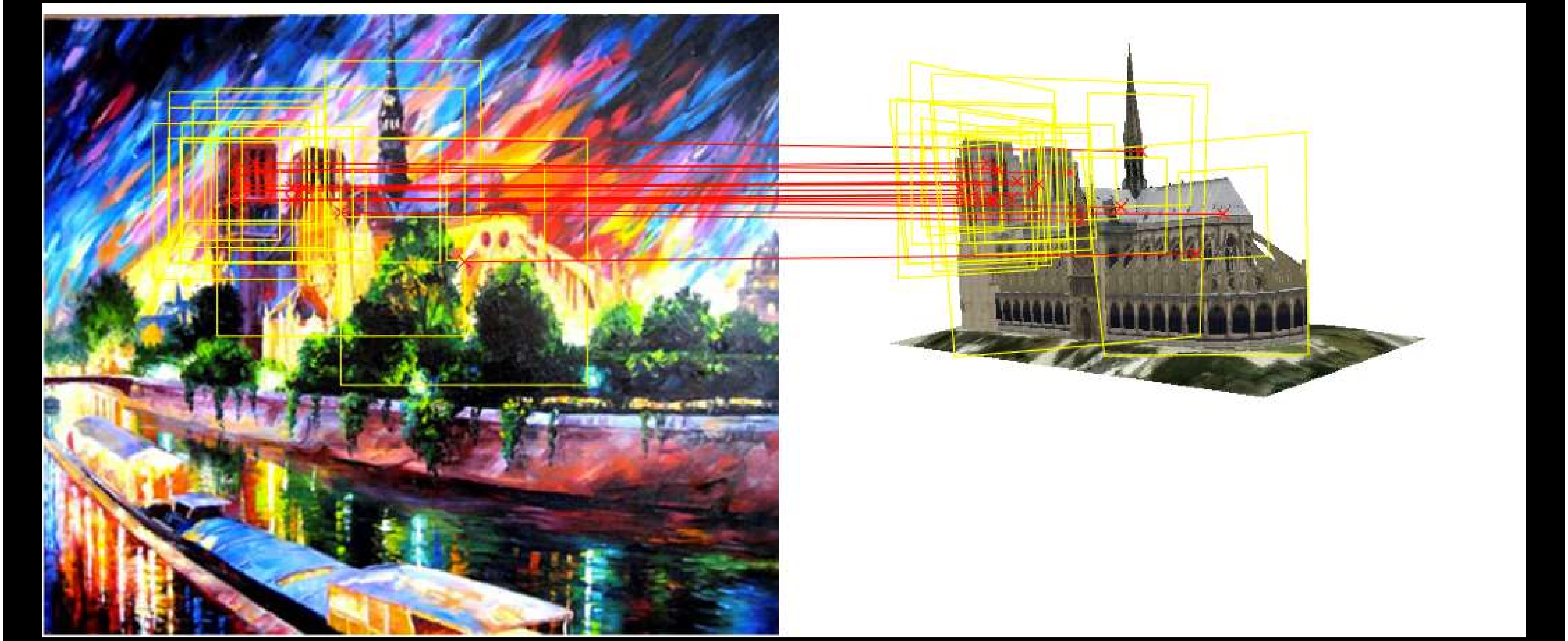


"Charade" [Donen, 1963]

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

Instance level recognition: still difficult

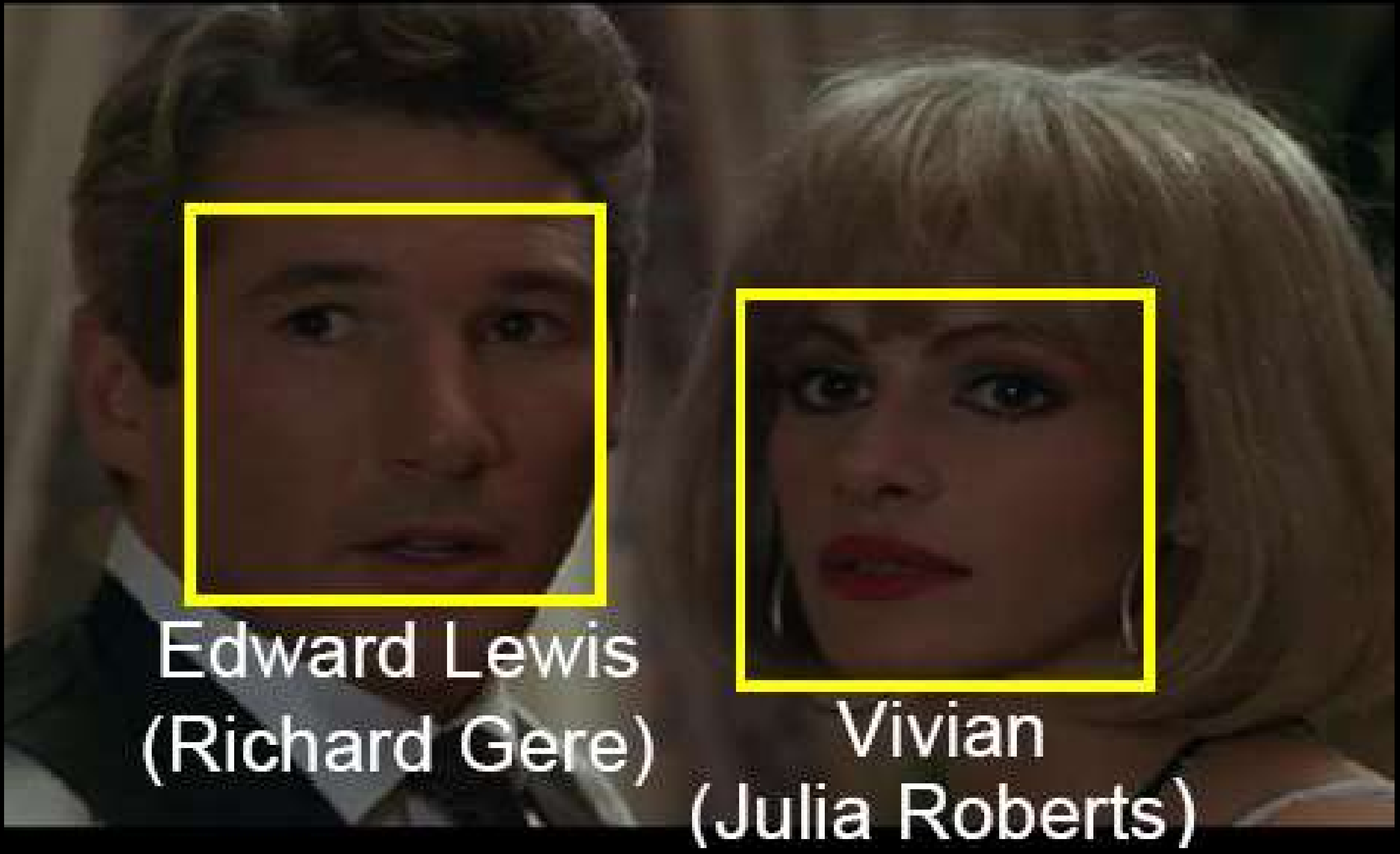
Example: Matching non-photographic depictions



Geo-localization of historical and non-photographic depictions

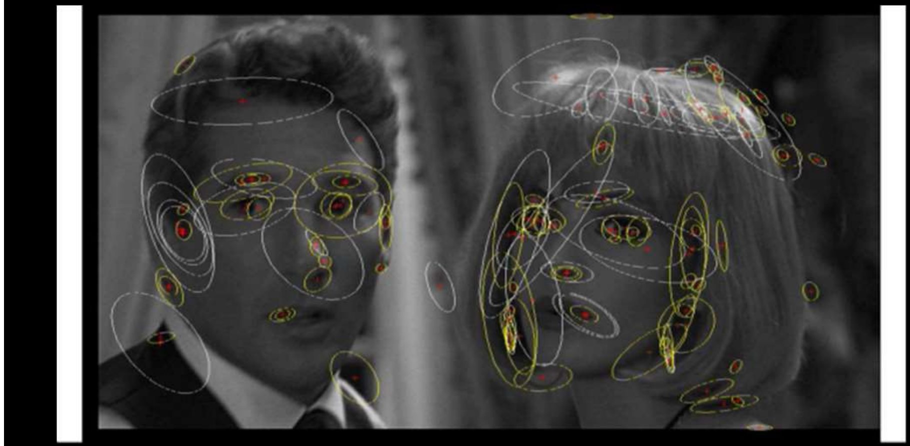
Image Landsat
Google earth

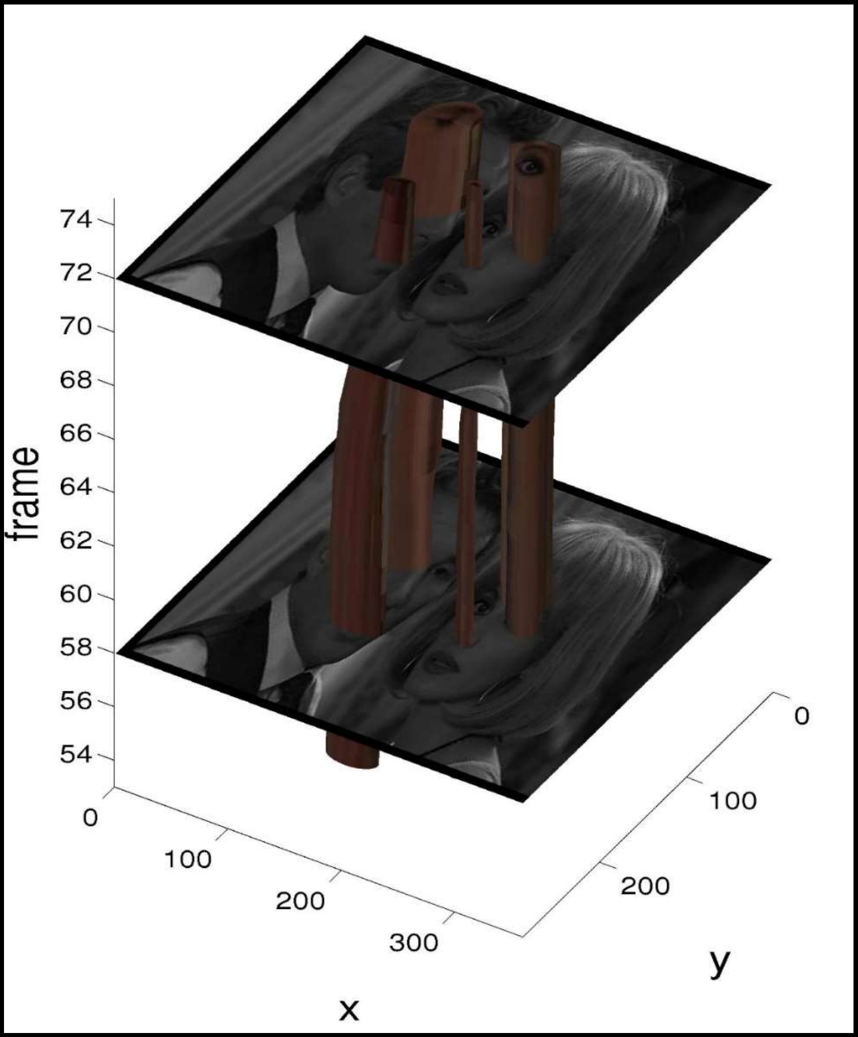
Recognizing people



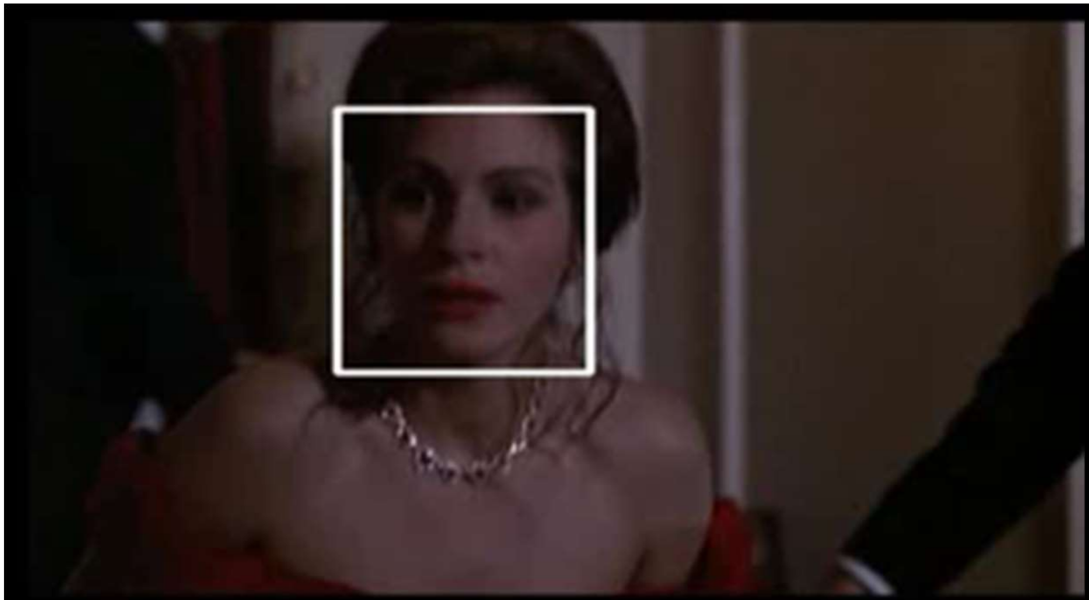
(Sivic, Everingham, Zisserman, 2005)

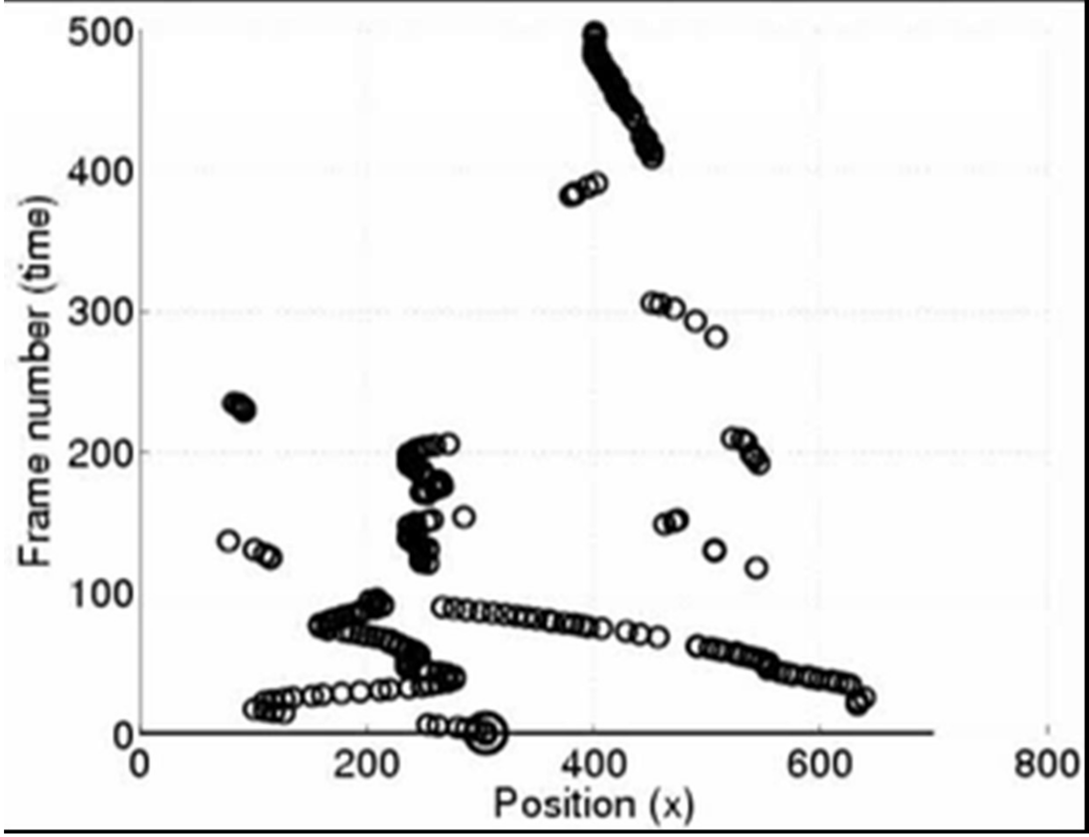
Faces:
Region tubes for tracking faces



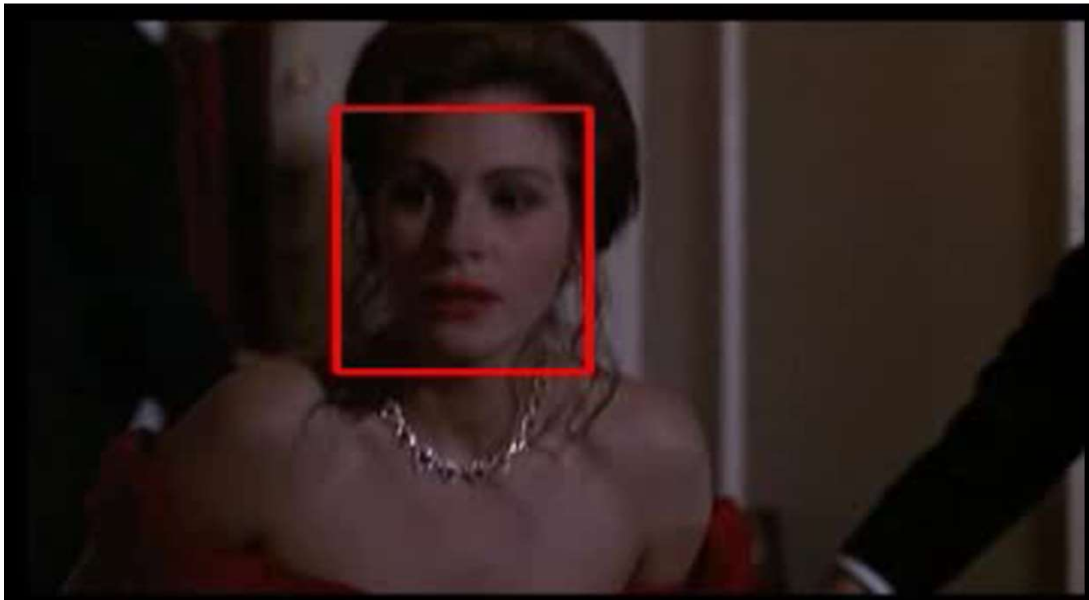


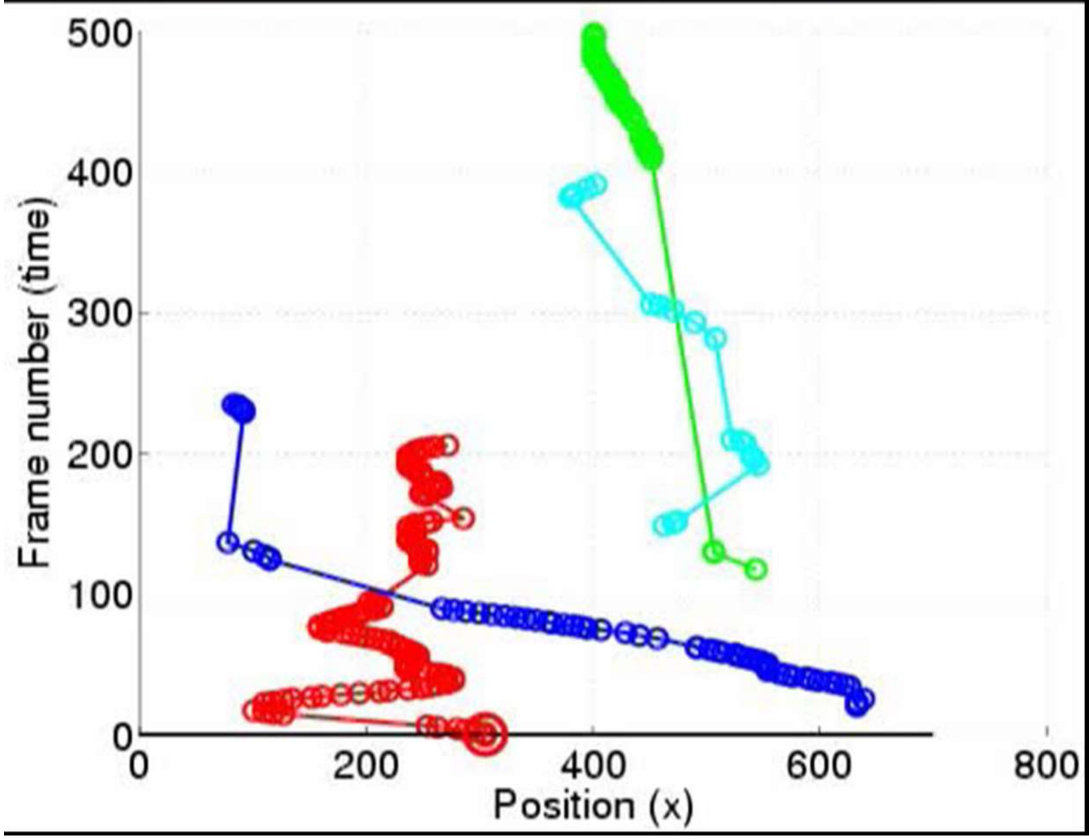
[Sivic, Everingham and Zisserman, 2005]





Raw face detections

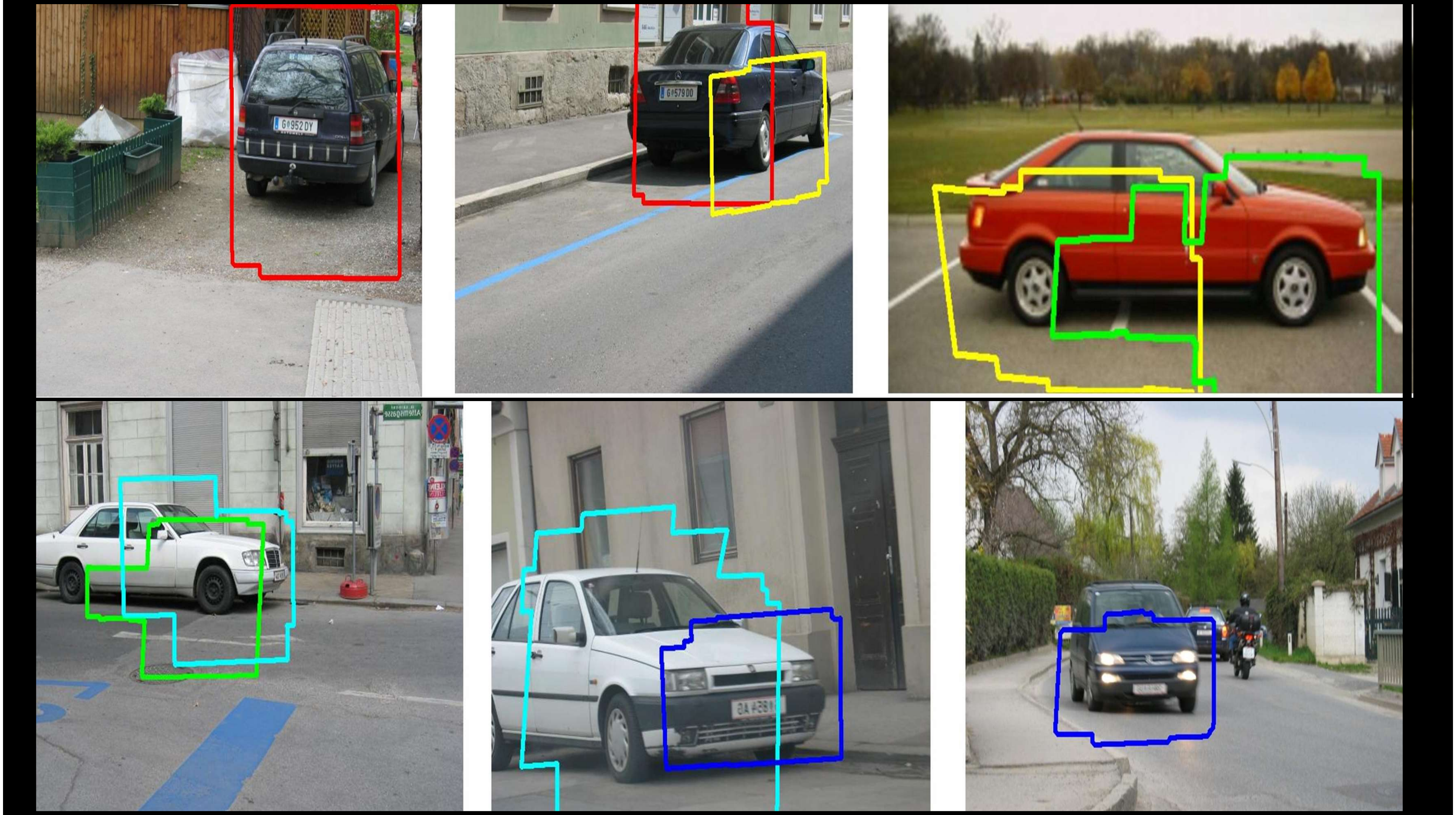




Tracking by detection and recognition

Connected face tracks

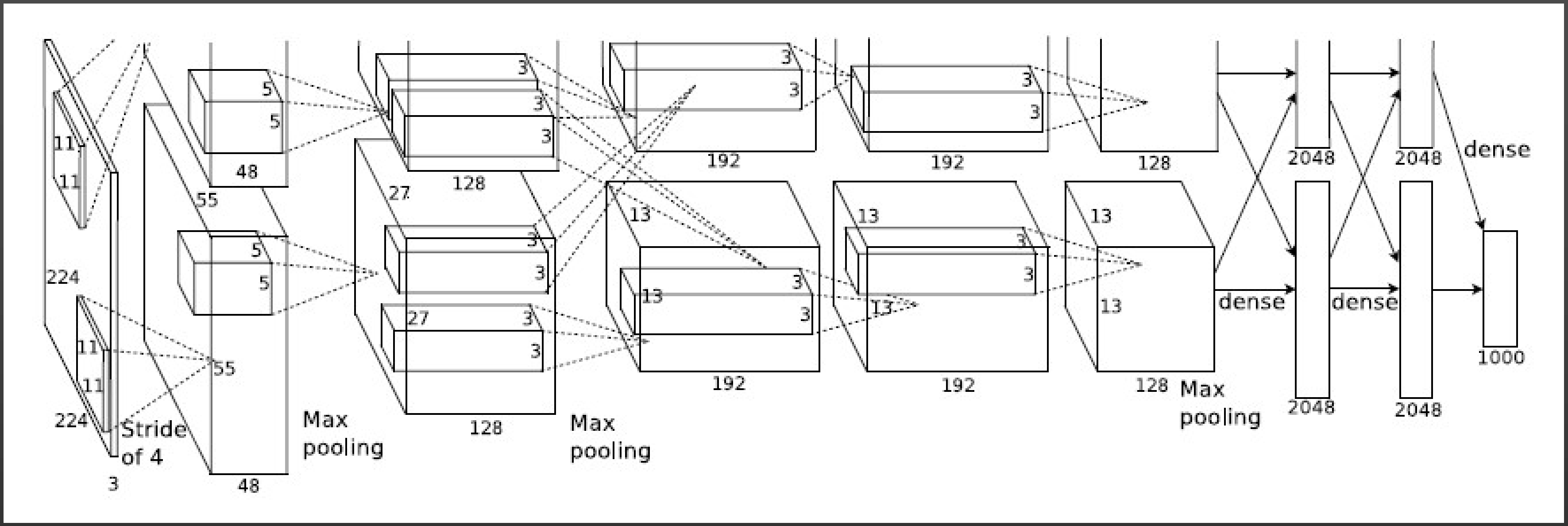
Recognition



(Kushal et al., 2007)

Convolutional neural networks

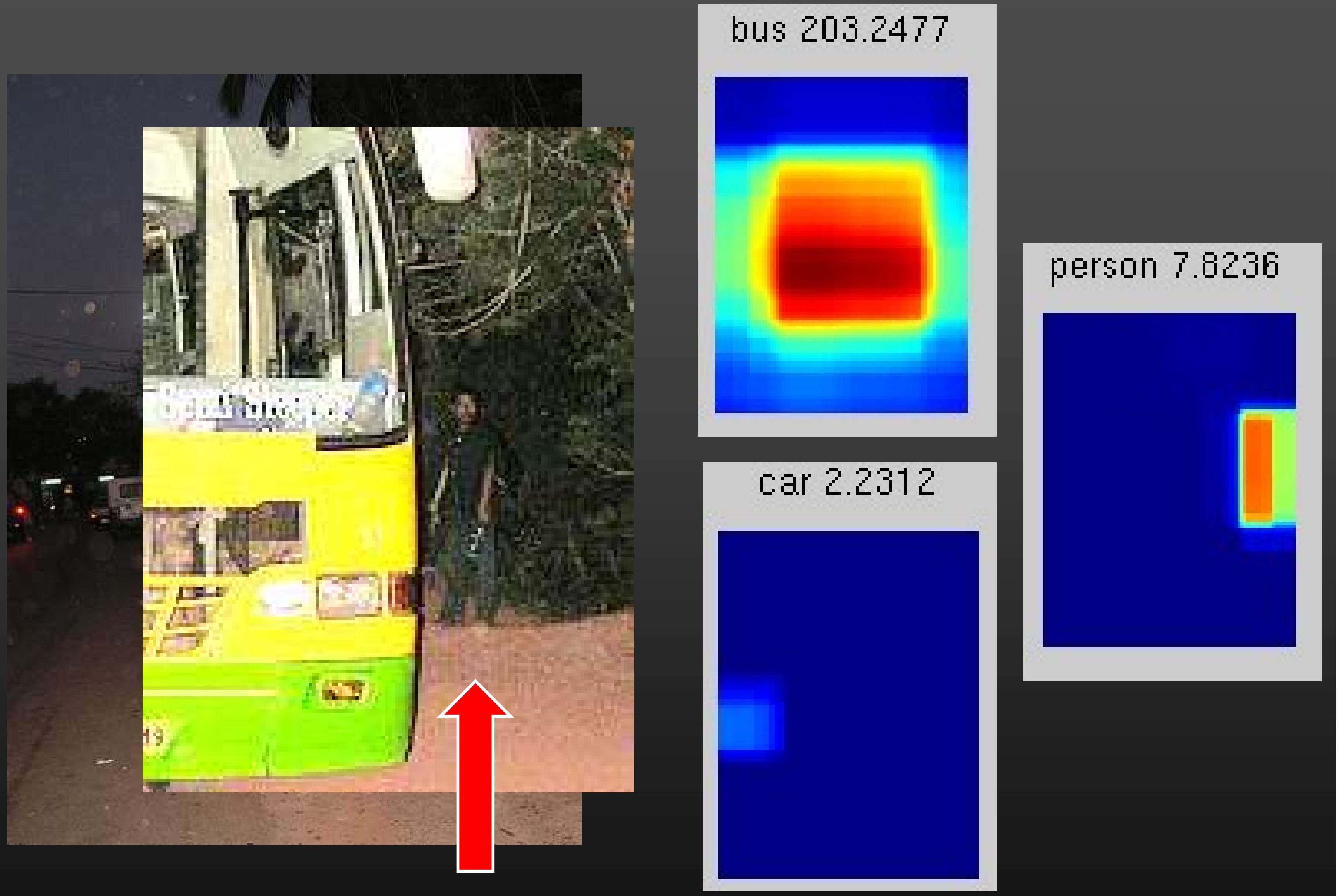
[Krizhevsky et al. NIPS'12]



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.

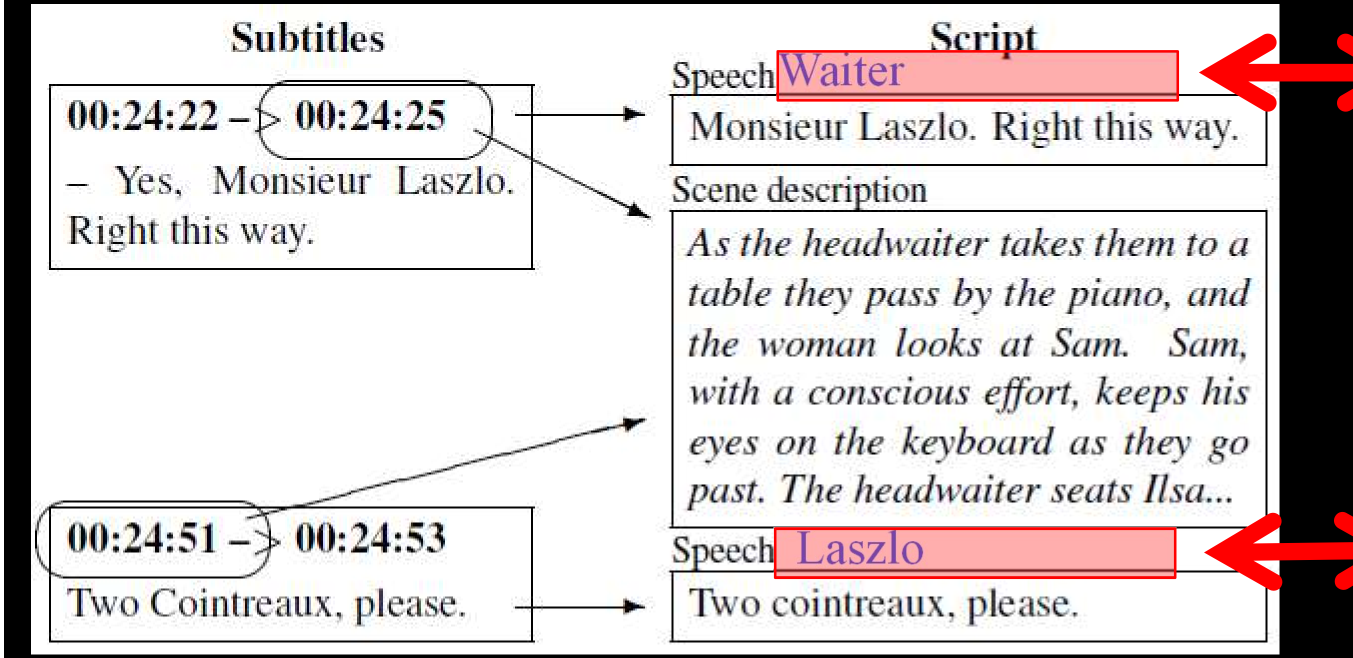
Some results



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

Automatic learning from video scripts

Input: Videos with aligned shooting scripts.



Subtitles

00:24:22 –> 00:24:25

– Yes, Monsieur Laszlo. Right this way.

00:24:51 –> 00:24:53

Two Cointreaux, please.

Script

Speech Waiter

Monsieur Laszlo. Right this way.

Scene description

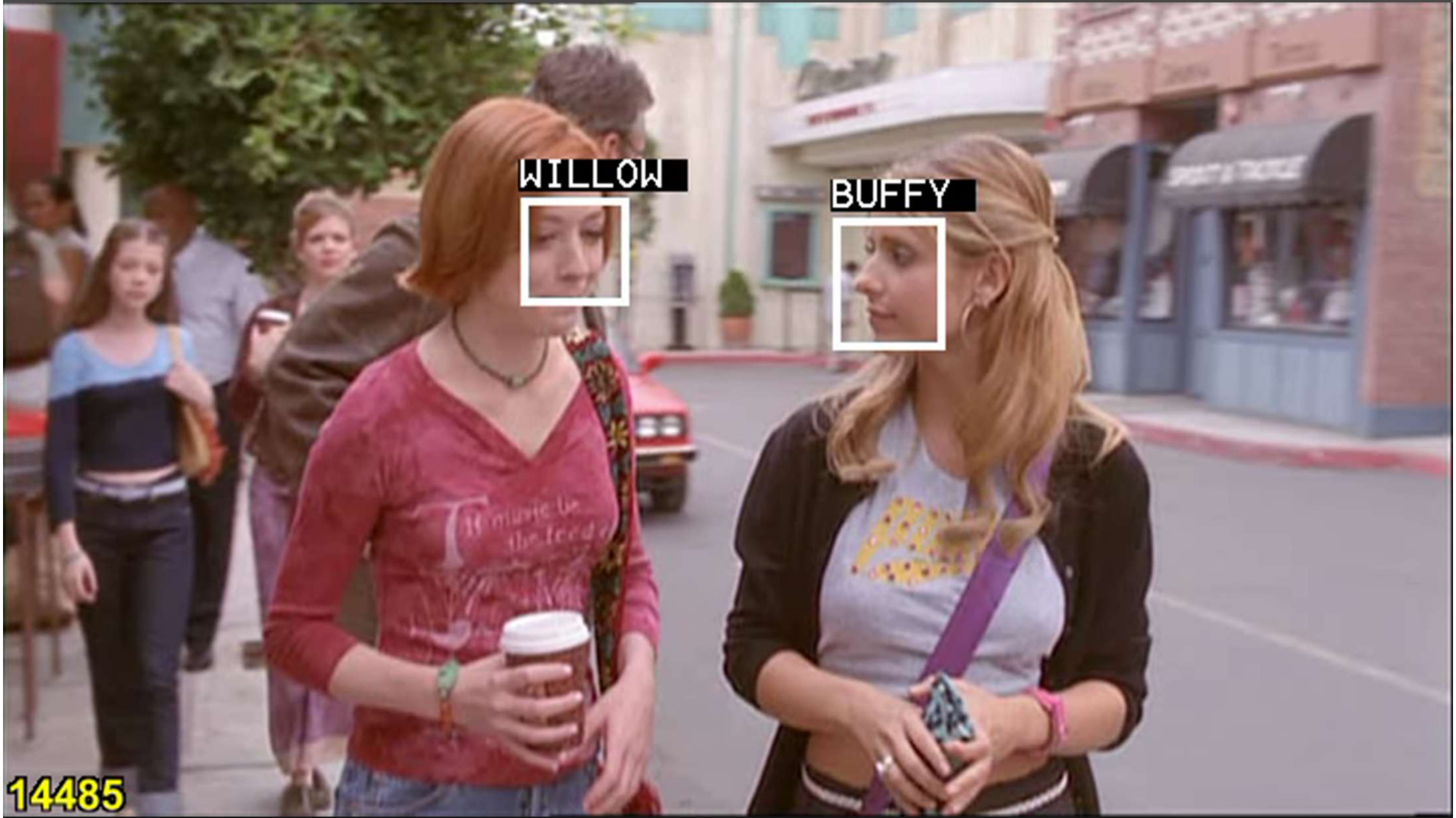
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Speech Laszlo

Two cointreaux, please.

Output: Recognizer for each character in the video

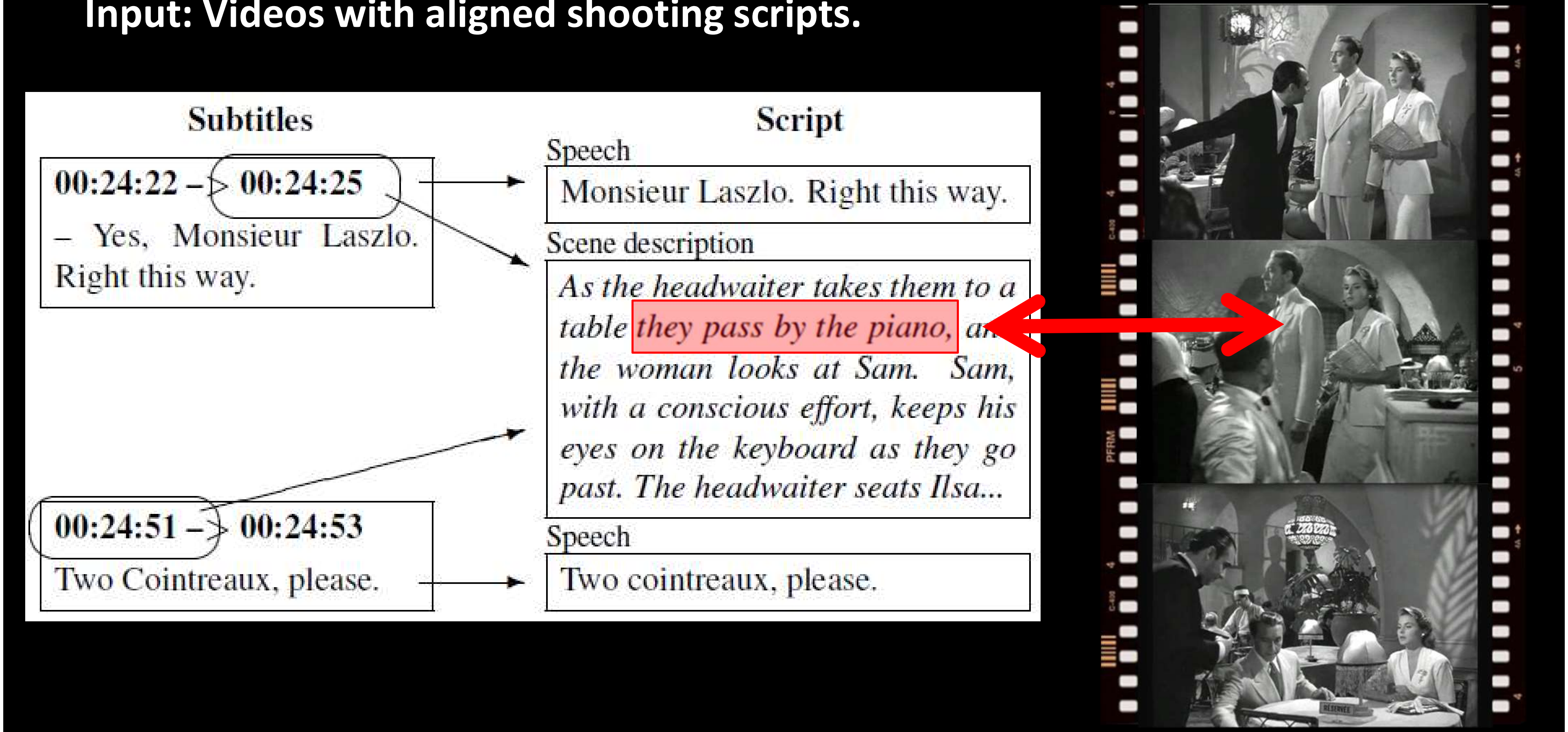
Recognizing people



(Everingham, Sivic, Zisserman, 2009)

Automatic learning from video scripts

Input: Videos with aligned shooting scripts.



Subtitles

00:24:22 –> 00:24:25

– Yes, Monsieur Laszlo. Right this way.

00:24:51 –> 00:24:53

Two Cointreaux, please.

Script

Speech

Monsieur Laszlo. Right this way.

Scene description

As the headwaiter takes them to a table they pass by the piano, a the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Speech

Two cointreaux, please.

Output: detector of human actions.

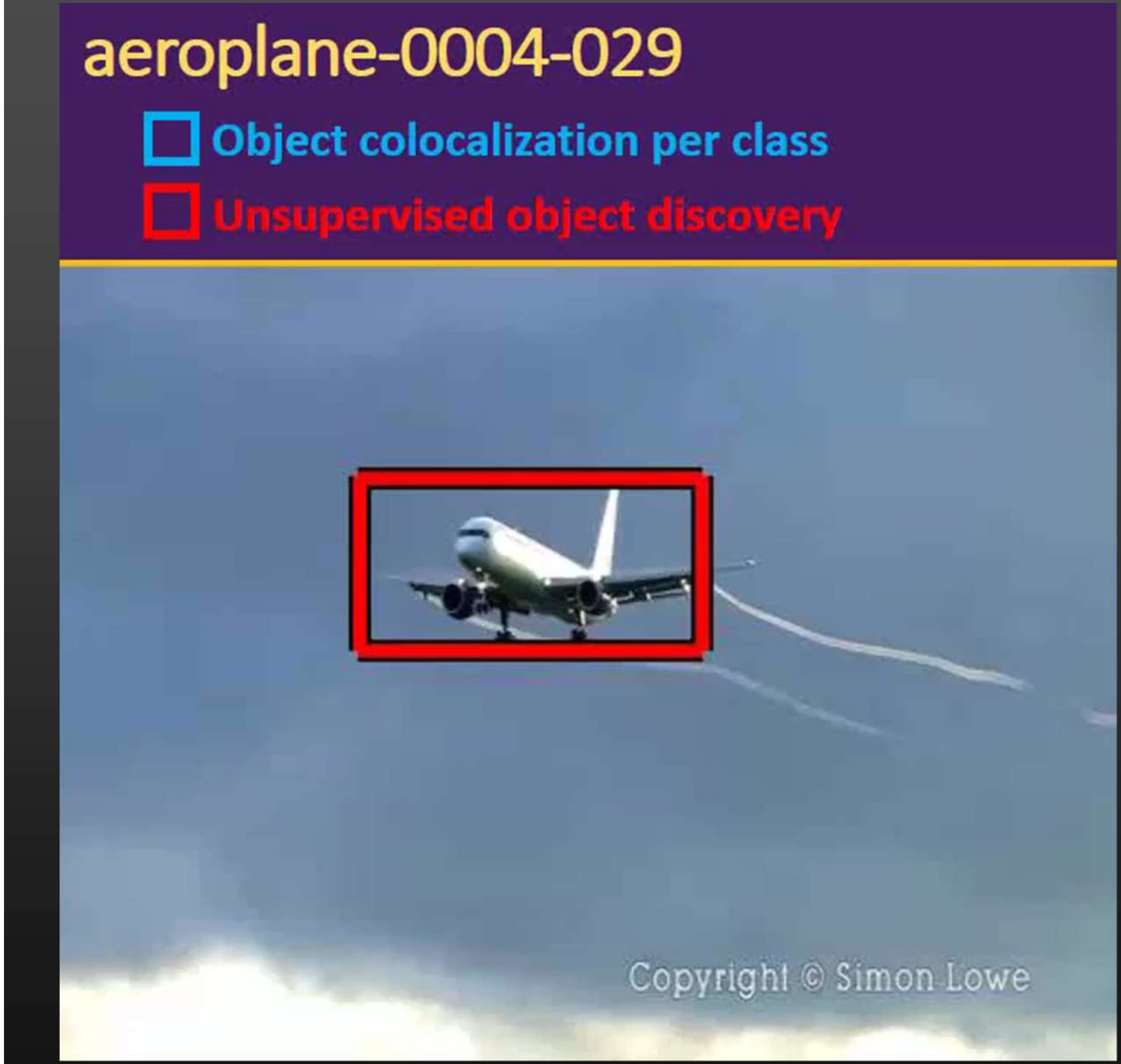
See also [Laptev, Marszałek, Schmid, Rozenfeld 2008]

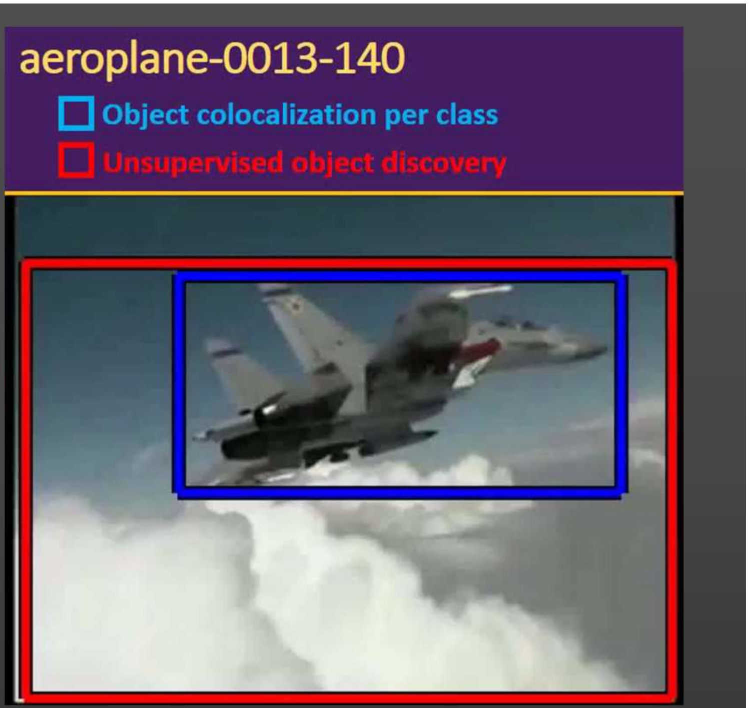
Weakly-supervised video interpretation

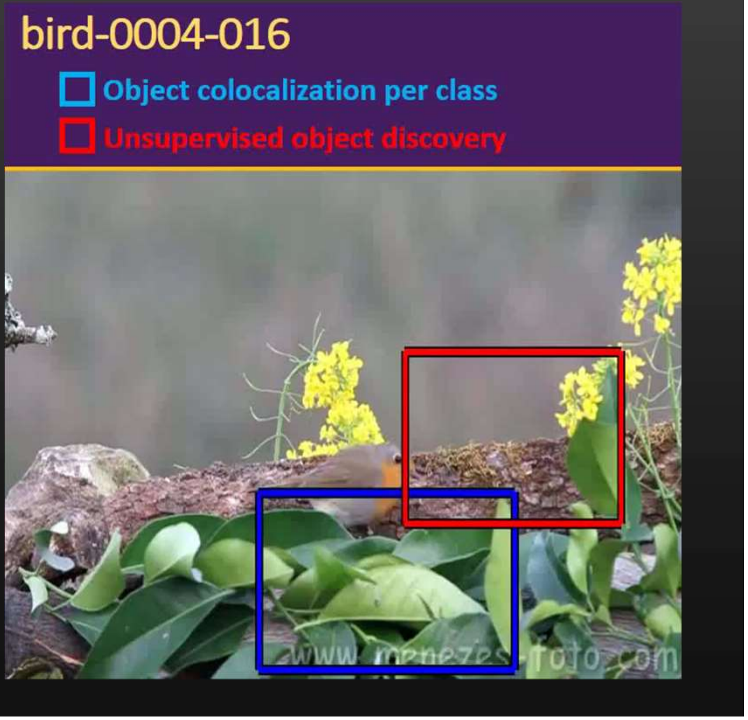
Clip number 0101

(Bojanowski et al., 2014)

Unsupervised object discovery


aeroplane-0004-029
□ Object colocalization per class
□ Unsupervised object discovery
Copyright © Simon Lowe

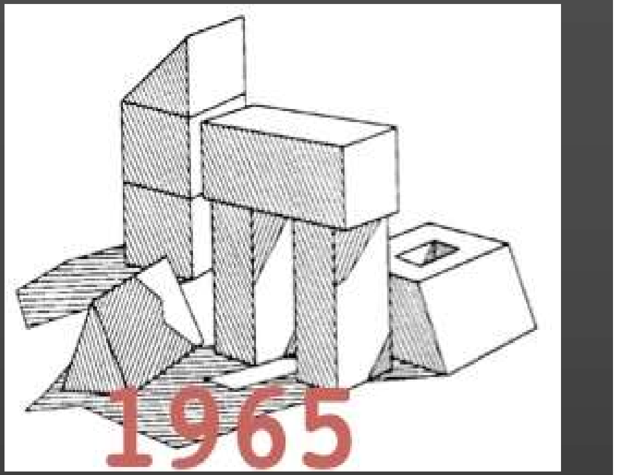

aeroplane-0013-140
□ Object colocalization per class
□ Unsupervised object discovery

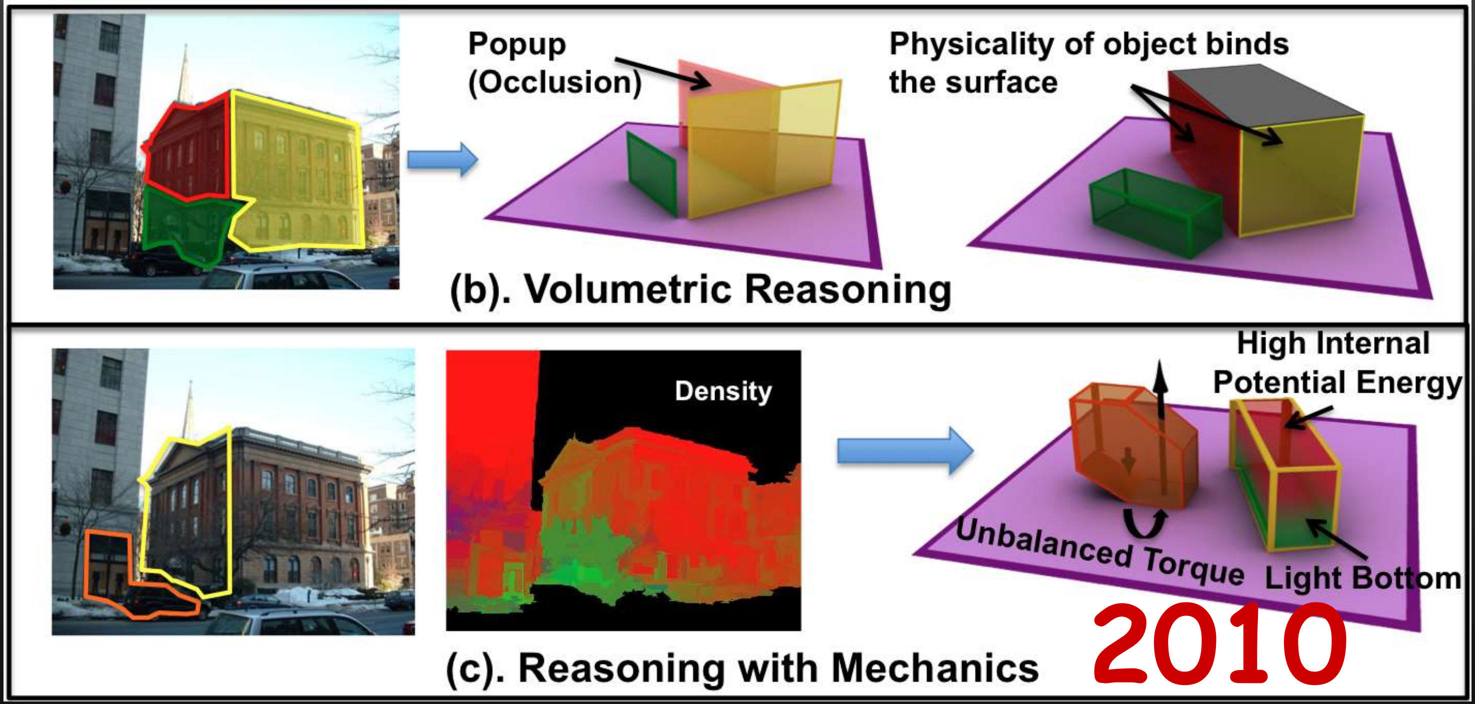

bird-0004-016
□ Object colocalization per class
□ Unsupervised object discovery
www.menezes-foto.com

(Suha et al., 2015)

What about scene understanding?

The blocks world revisited

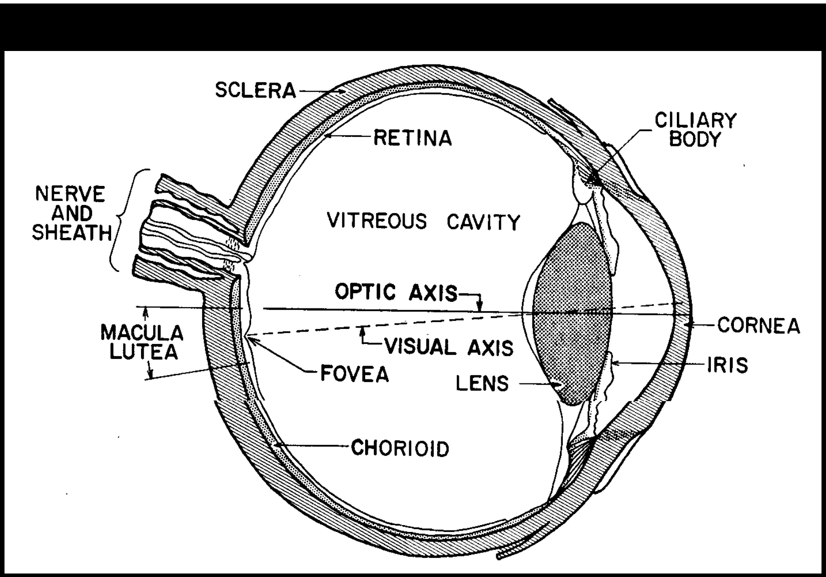




(Gupta, Efros, Hebert, ECCV'10)

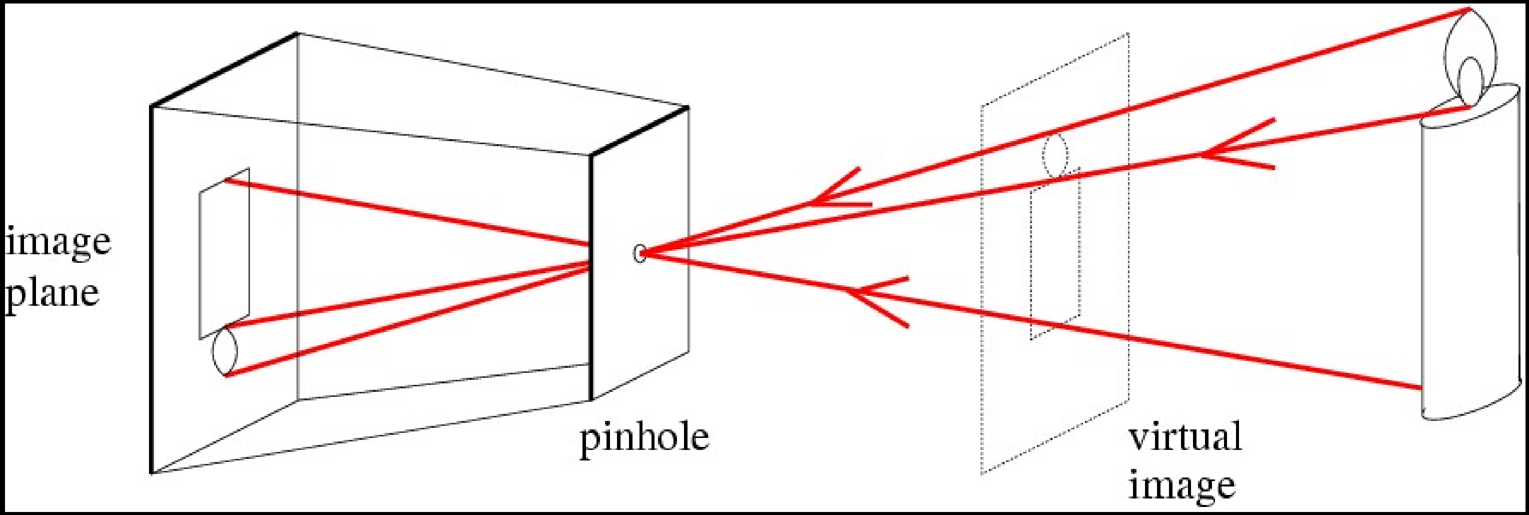
Camera geometry and calibration I

- Pinhole perspective projection
- Orthographic and weak-perspective models
- Non-standard models
- *A detour through sensing country*
- Intrinsic and extrinsic parameters

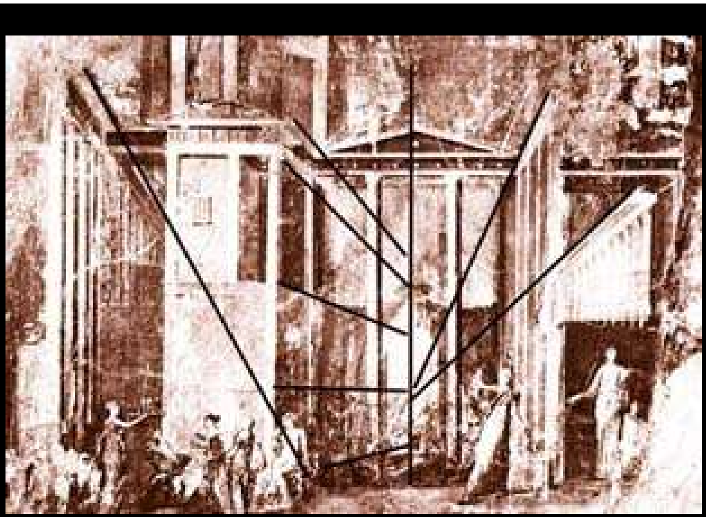


Animal eye: a looonnng time ago.

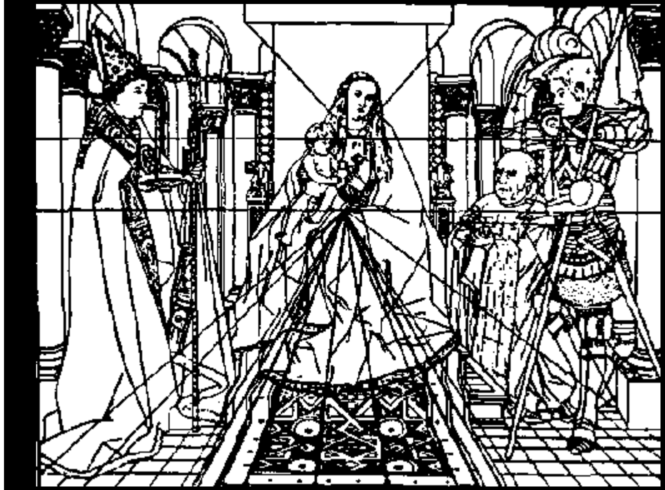
Photographic camera: Niepce, 1816.



Pinhole perspective projection: Brunelleschi, XVth Century.
Camera obscura: XVIth Century.

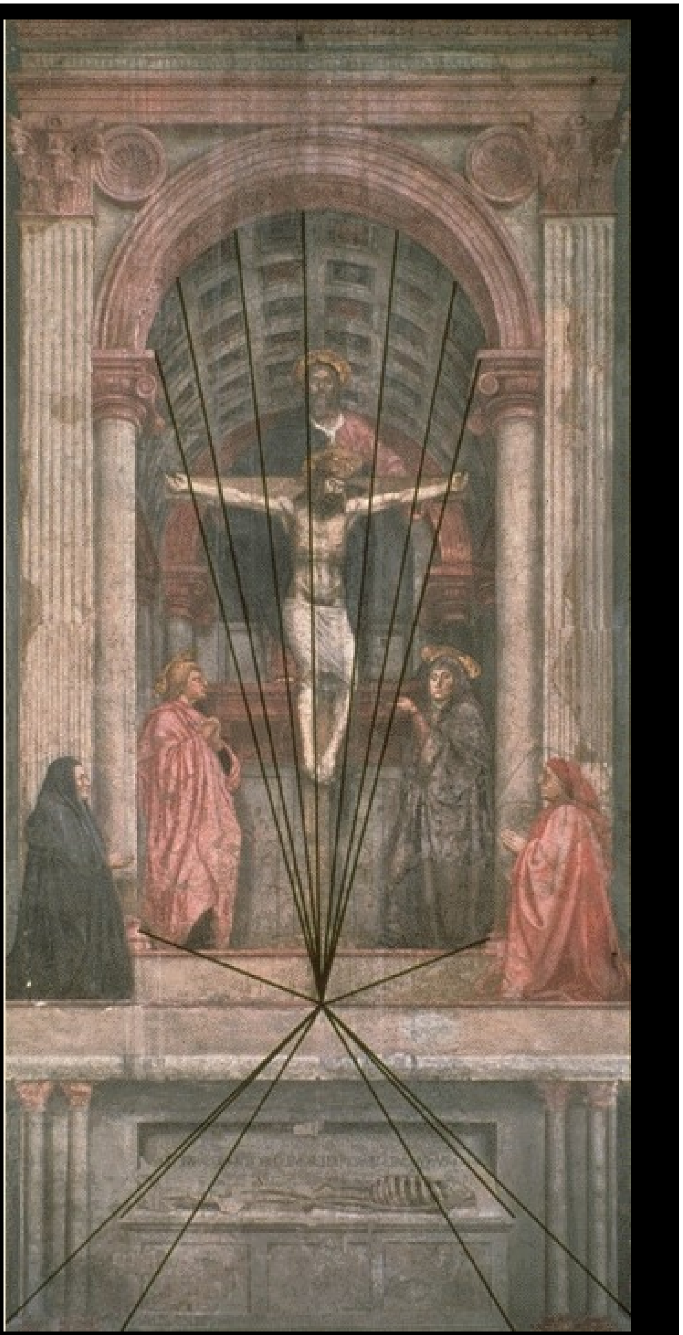


Pompei painting, 2000 years ago

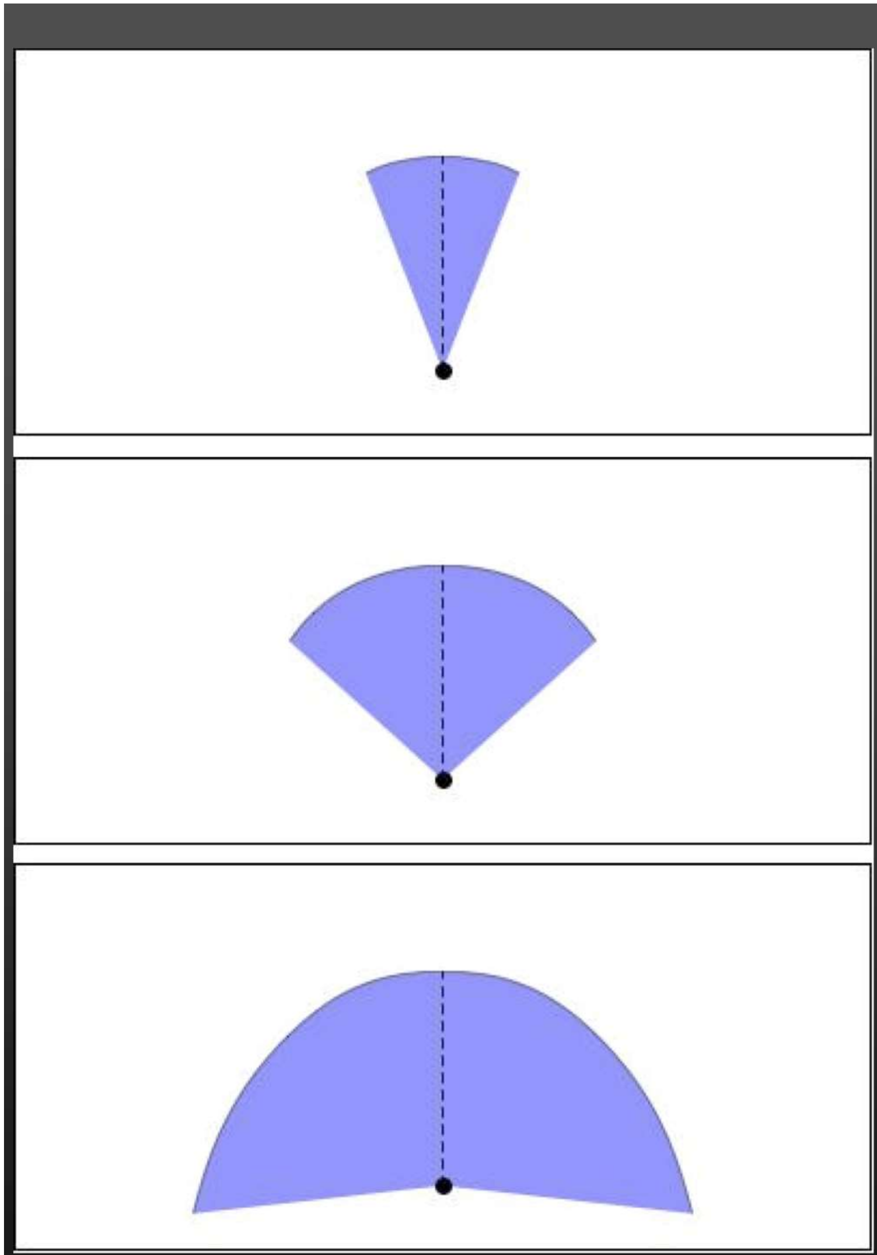


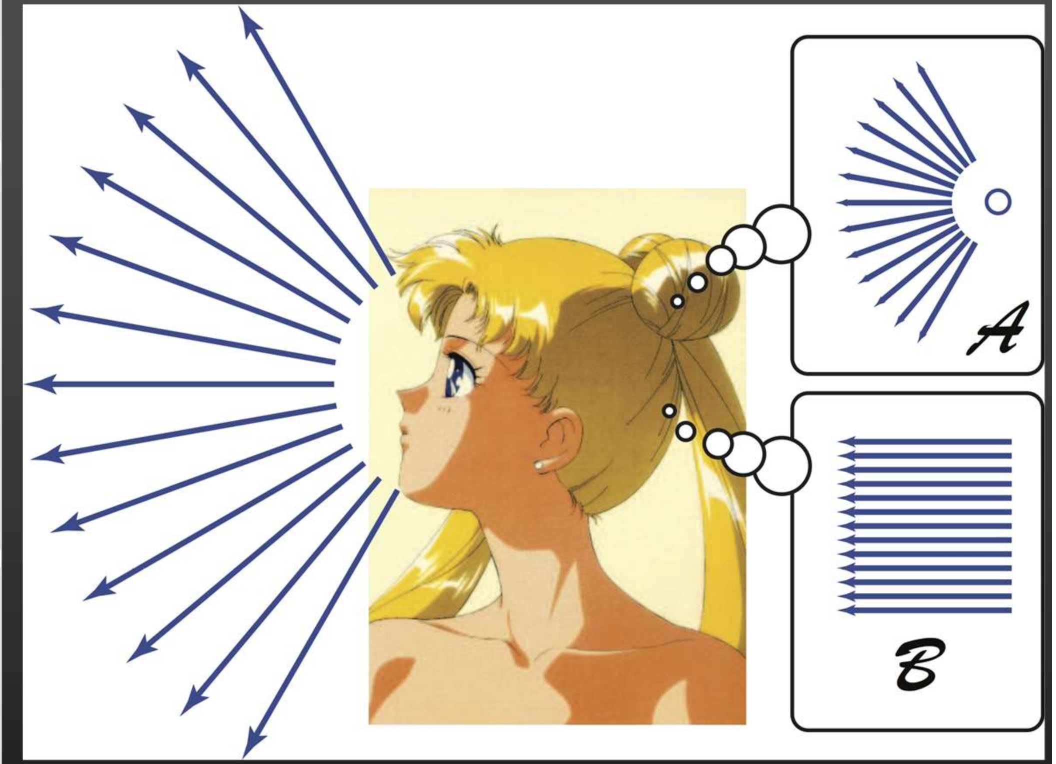
Van Eyk, XIVth Century

Brunelleschi, 1415

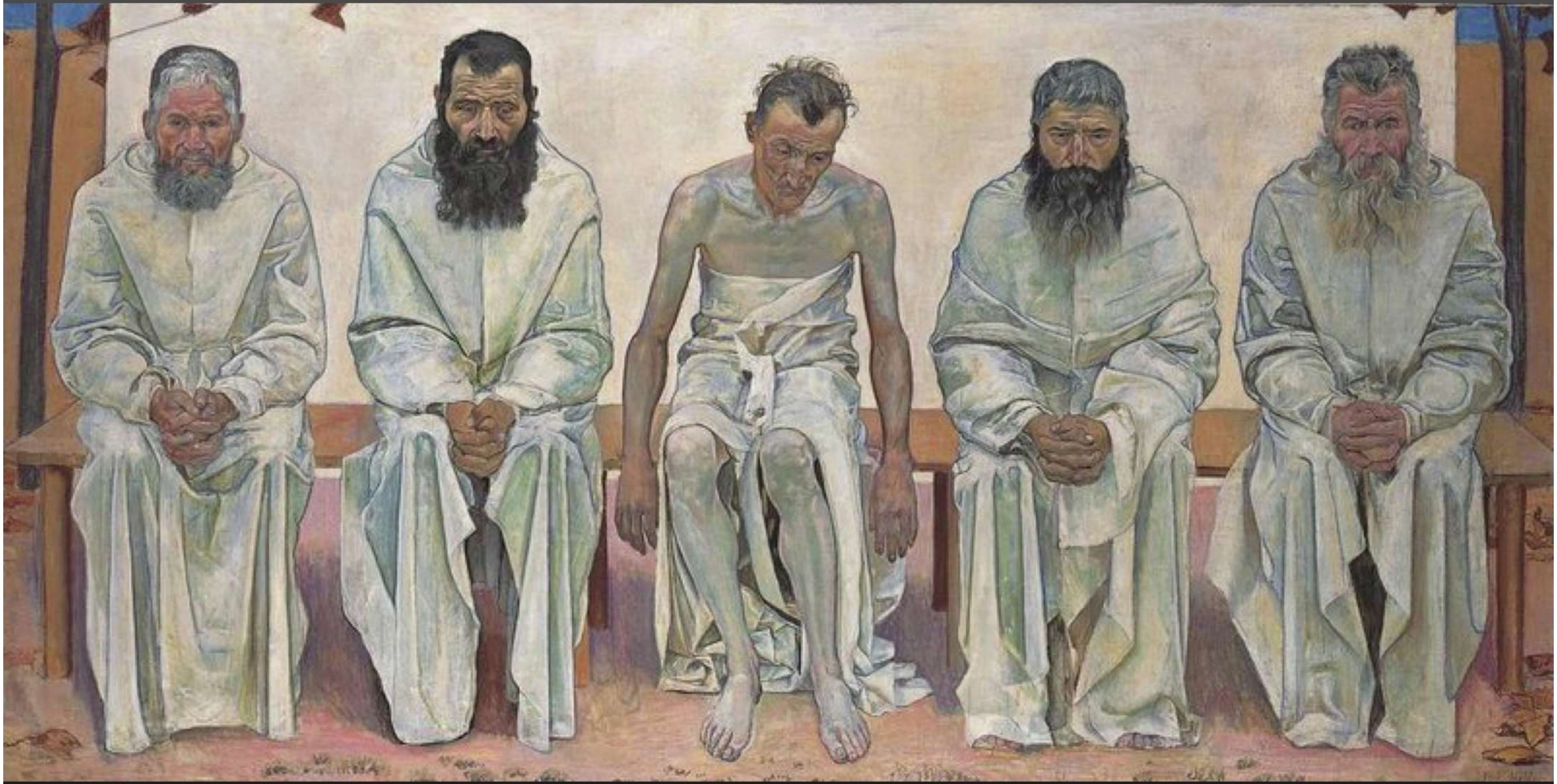


Massaccio's Trinity, 1425

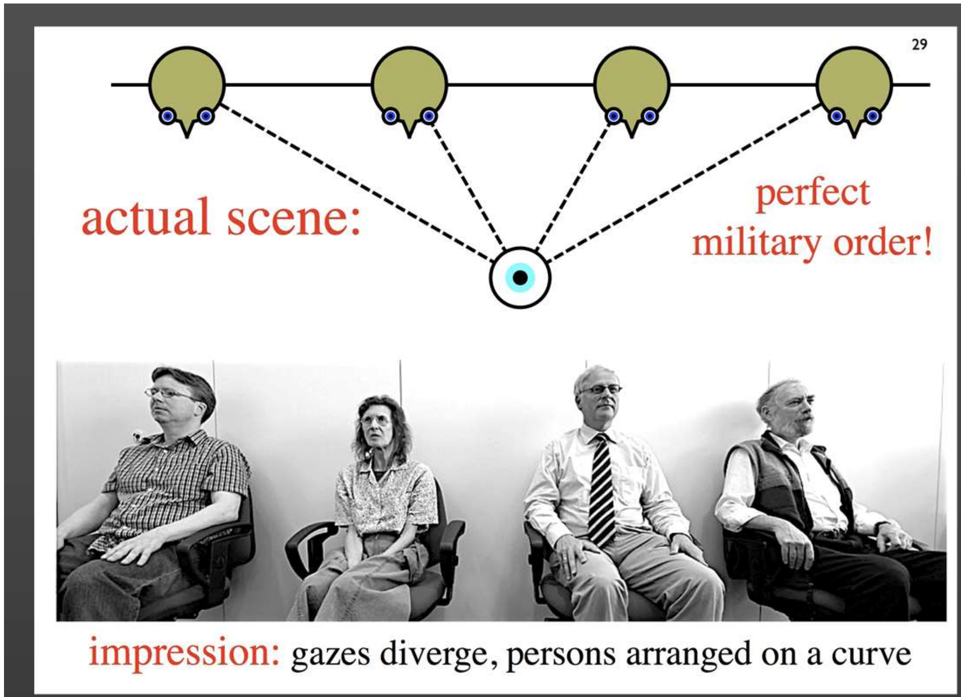




Most people don’t experience the divergence of visual rays in a veridical manner. This is fine. [Koenderink]



Ferdinand Hodler

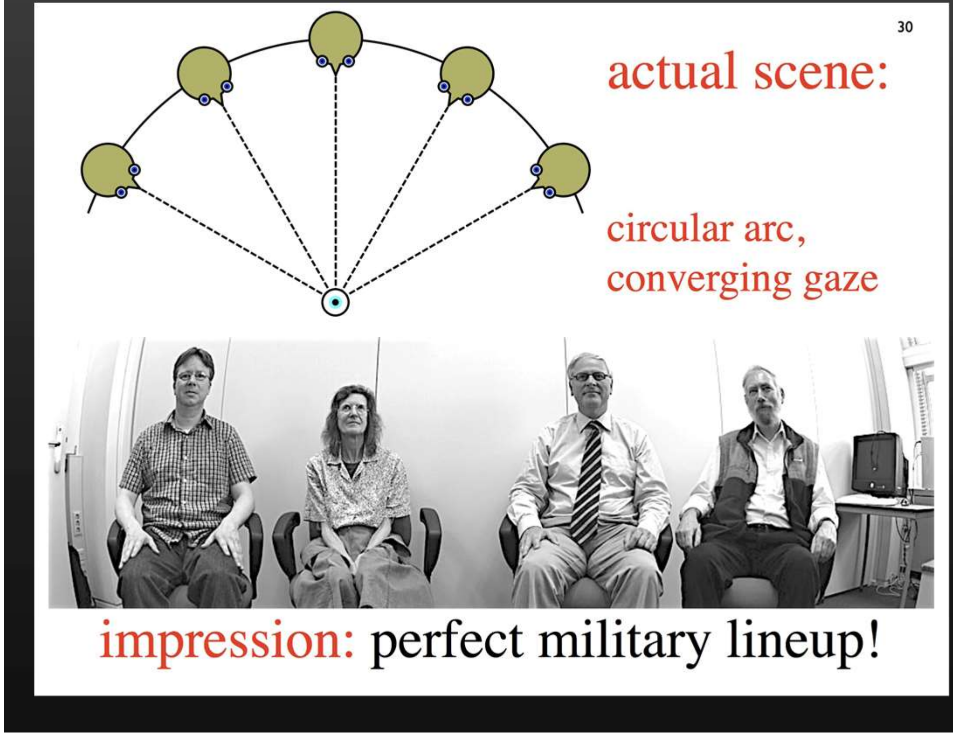


29

actual scene:

perfect military order!

impression: gazes diverge, persons arranged on a curve

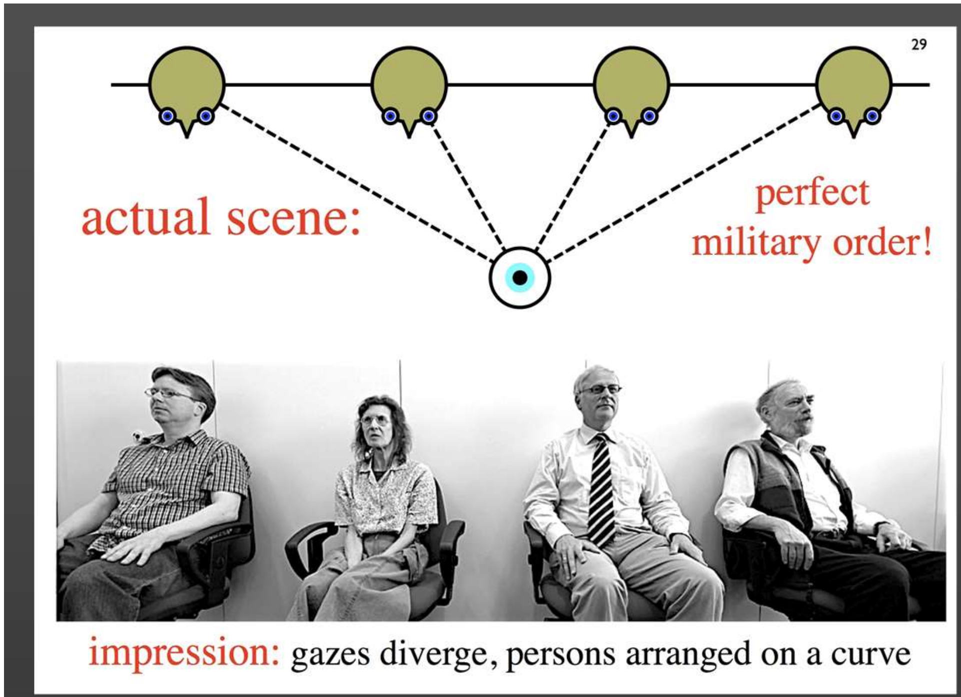


30

actual scene:

circular arc, converging gaze

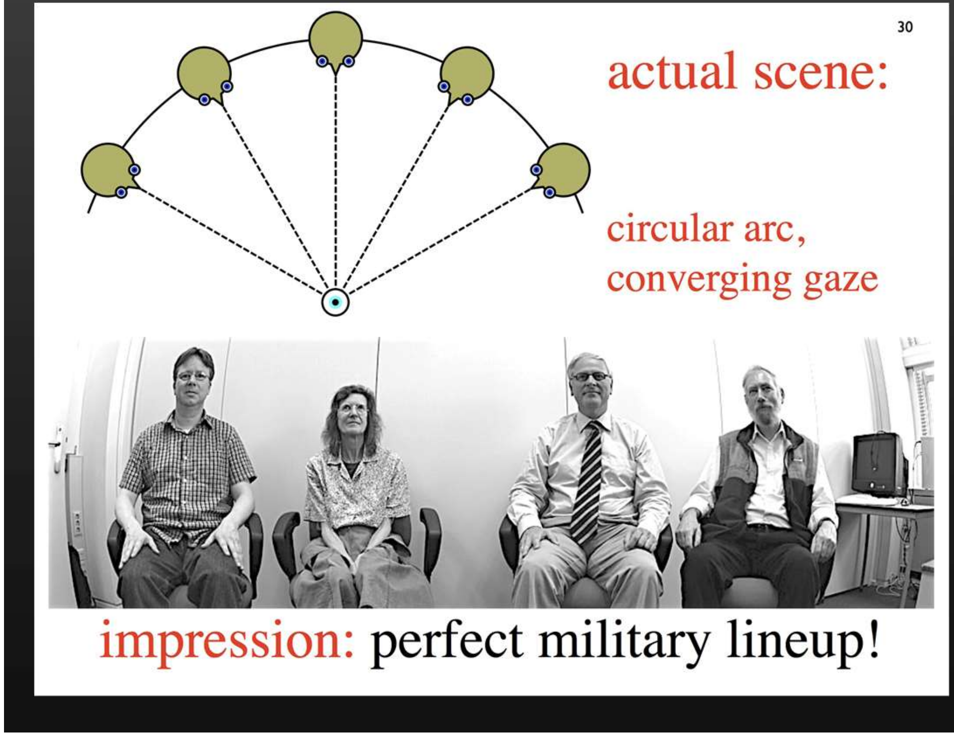
impression: perfect military lineup!



actual scene:

perfect military order!

impression: gazes diverge, persons arranged on a curve

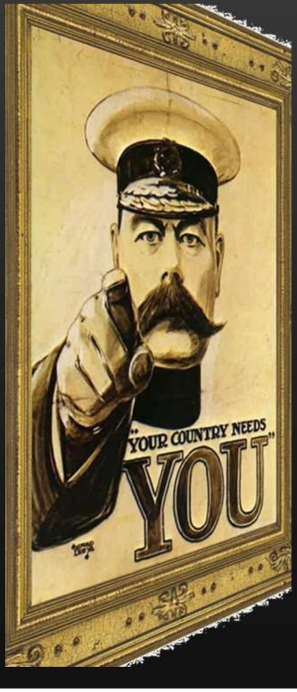


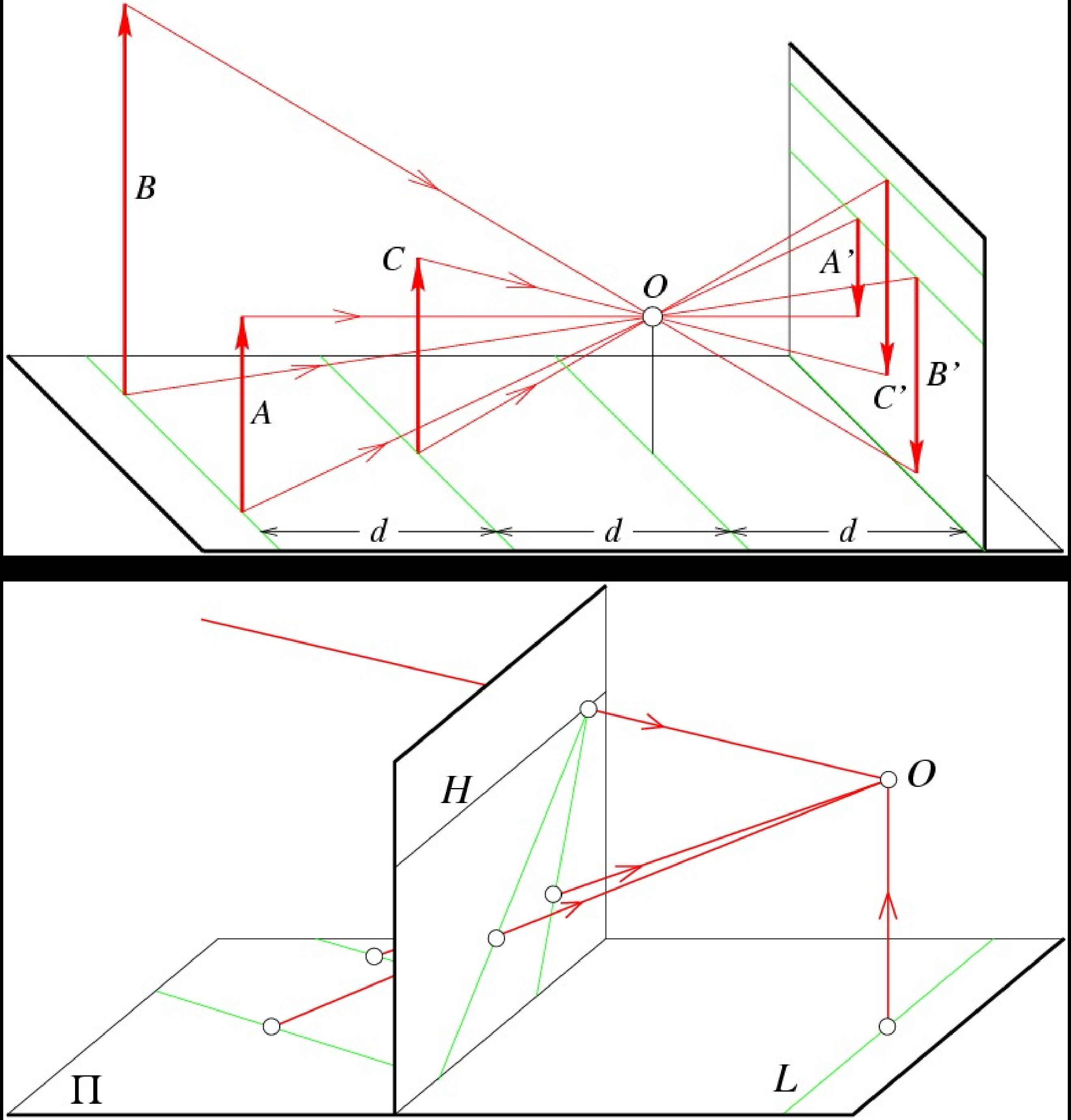
actual scene:

circular arc, converging gaze

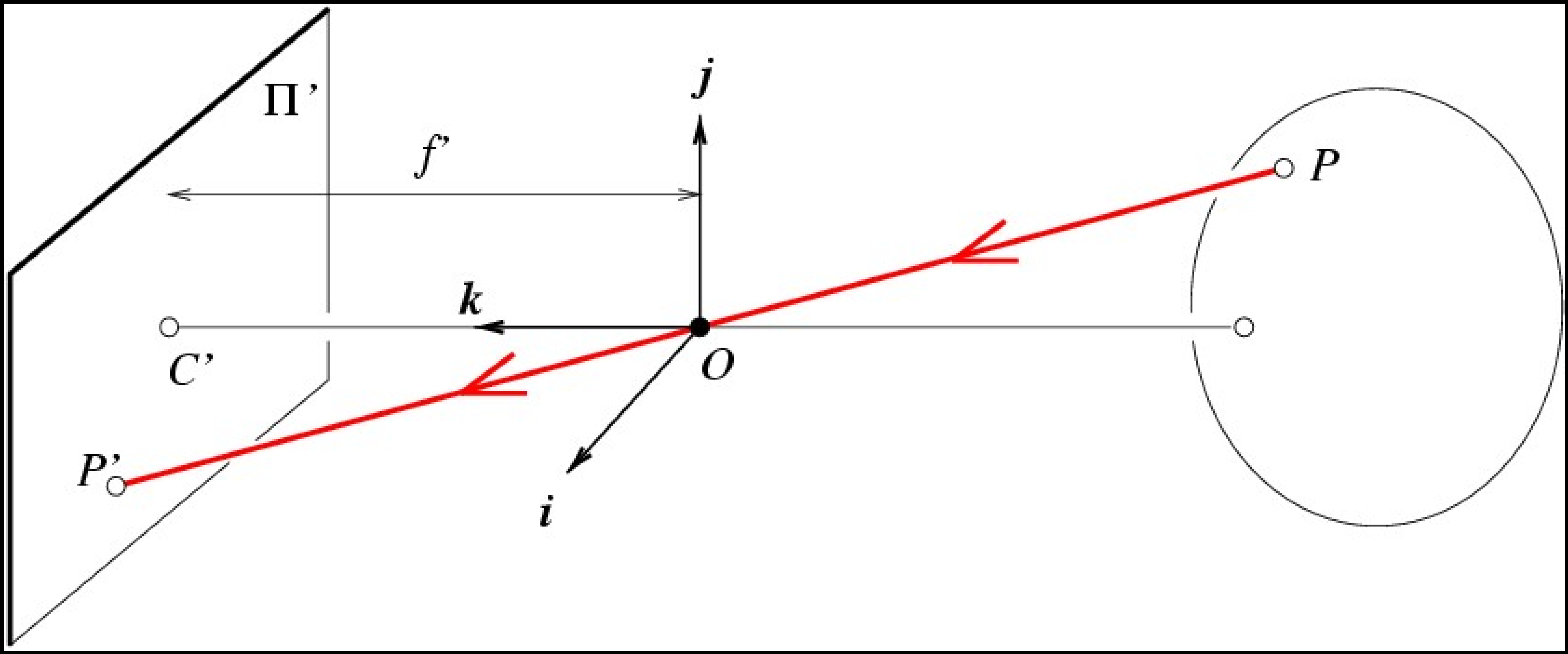
impression: perfect military lineup!

"YOUR COUNTRY NEEDS
YOU"




B
C
O
A'
A
C'
B'
d
d
d
H
O
Π
L

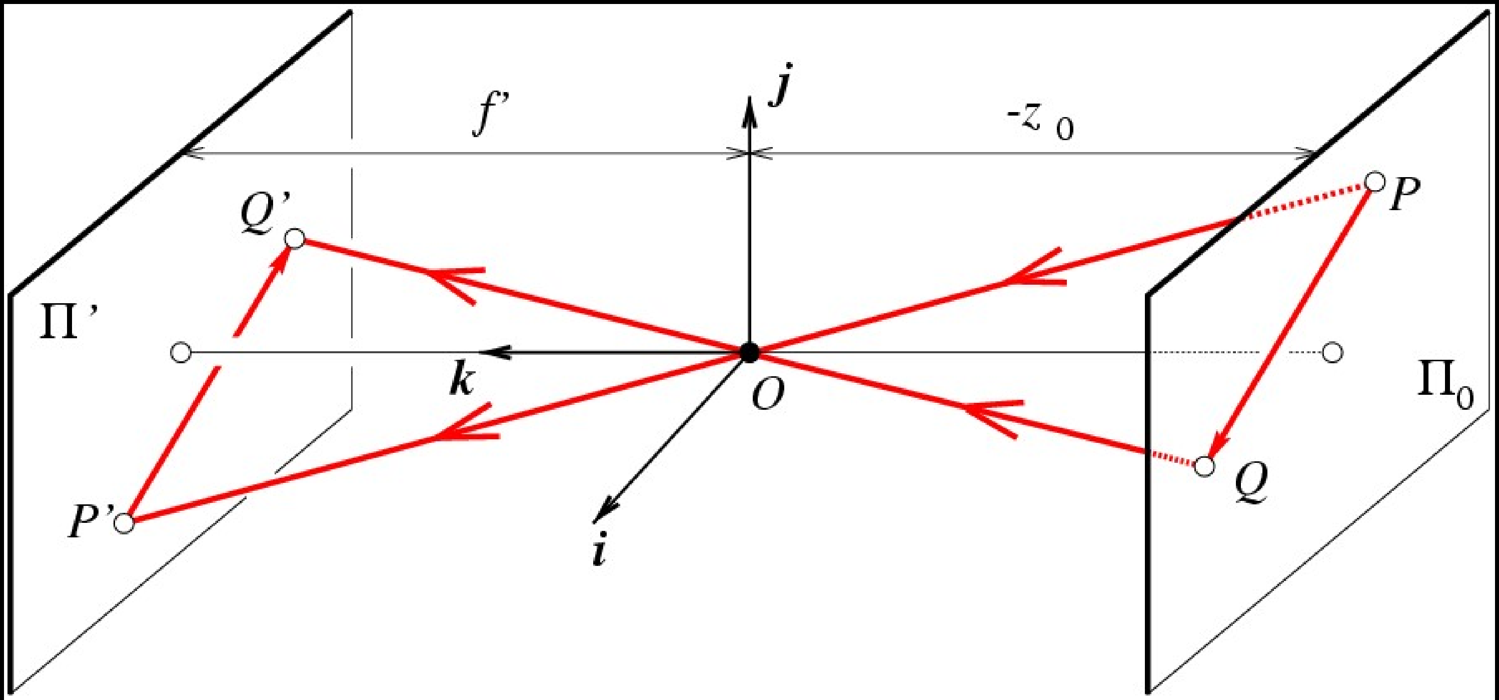
Pinhole Perspective Equation



$$\begin{cases} x' = f' \dfrac{x}{z} \\ y' = f' \dfrac{y}{z} \end{cases}$$

NOTE: *z* is always negative..

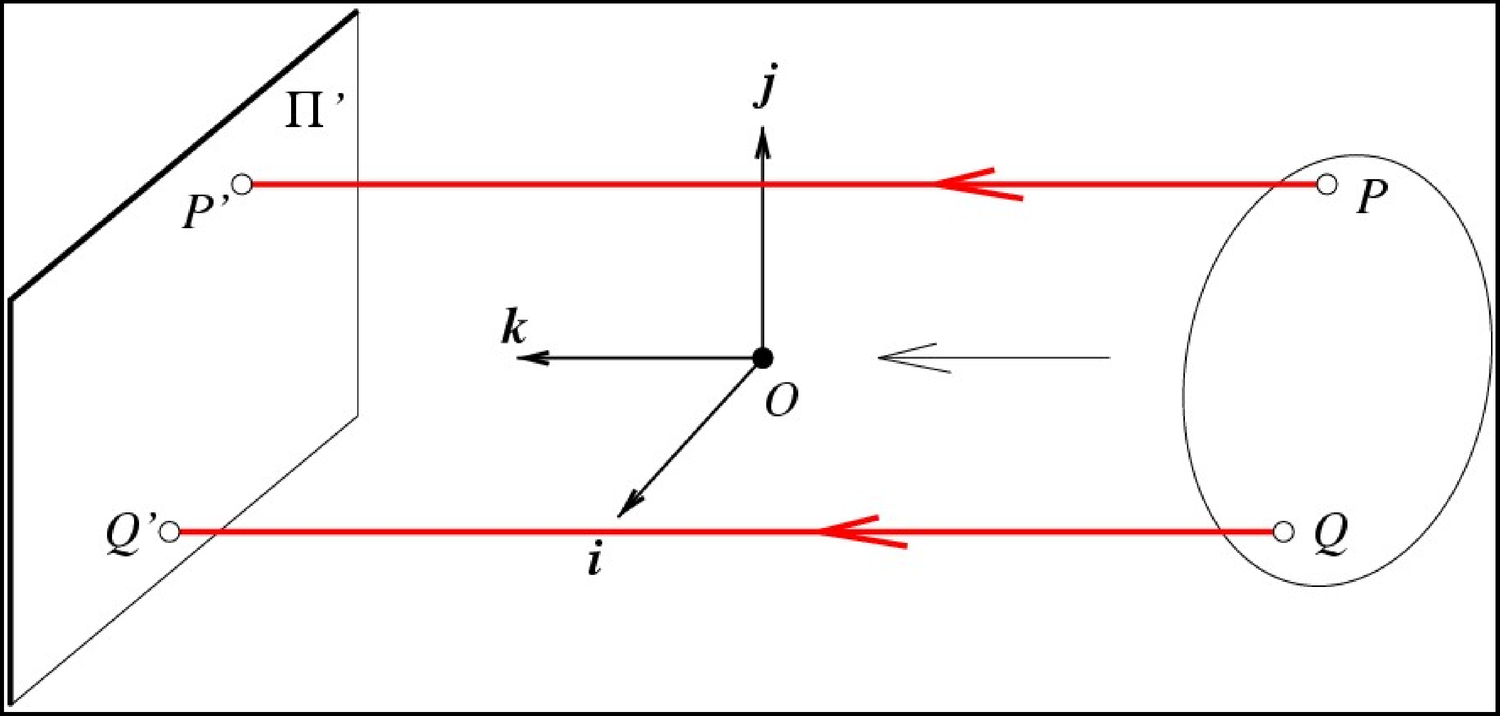
Affine projection models: Weak perspective projection



$$\boxed{\begin{cases} x' = -mx \\ y' = -my \end{cases}} \quad \text{where} \quad m = -\frac{f'}{z_0}$$ is the magnification.

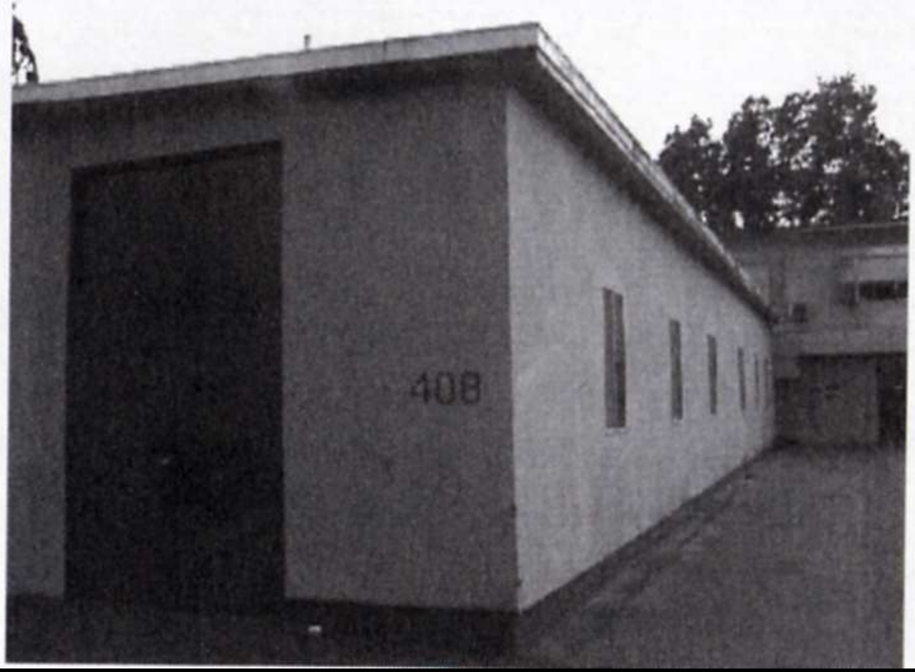
When the scene relief is small compared its distance from the Camera, *m* can be taken constant: weak perspective projection.

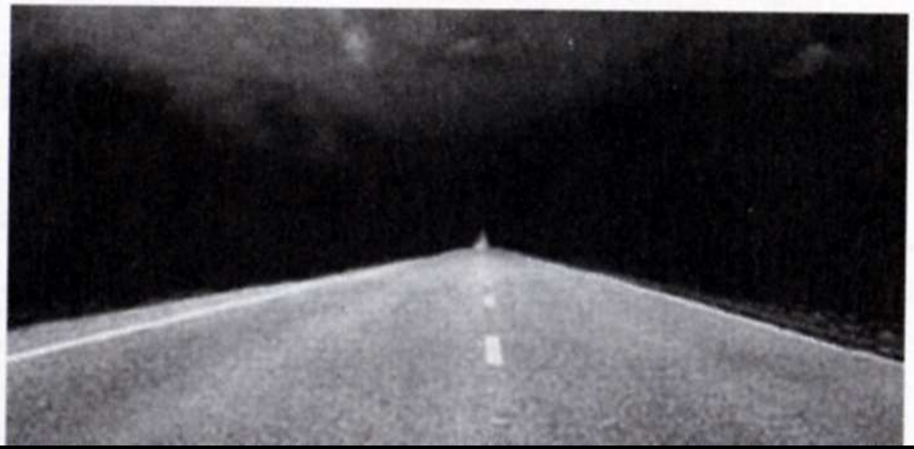
Affine projection models: Orthographic projection



$$\begin{cases} x' = x \\ y' = y \end{cases}$$

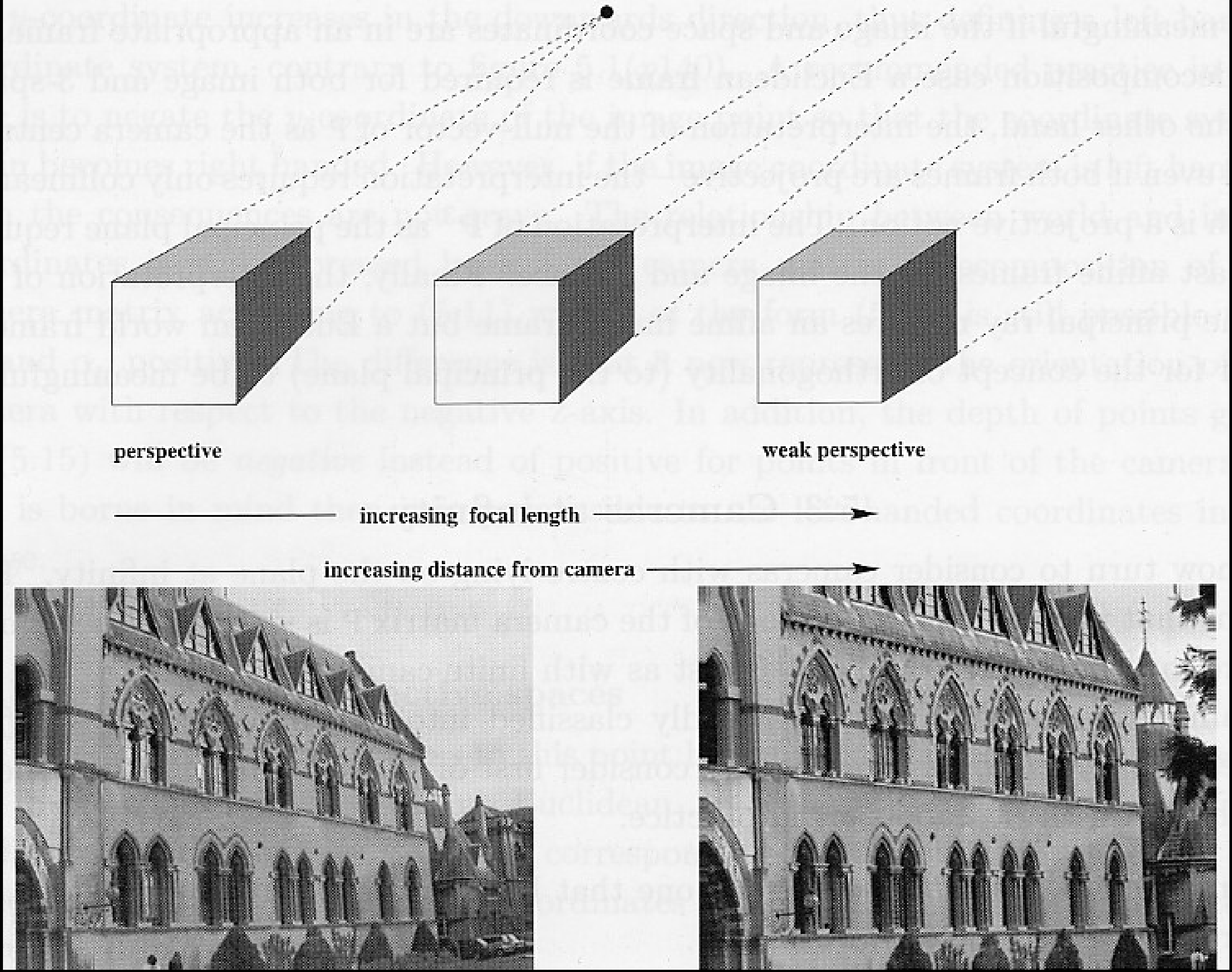
When the camera is at a (roughly constant) distance from the scene, take *m*=1.





Strong perspective:

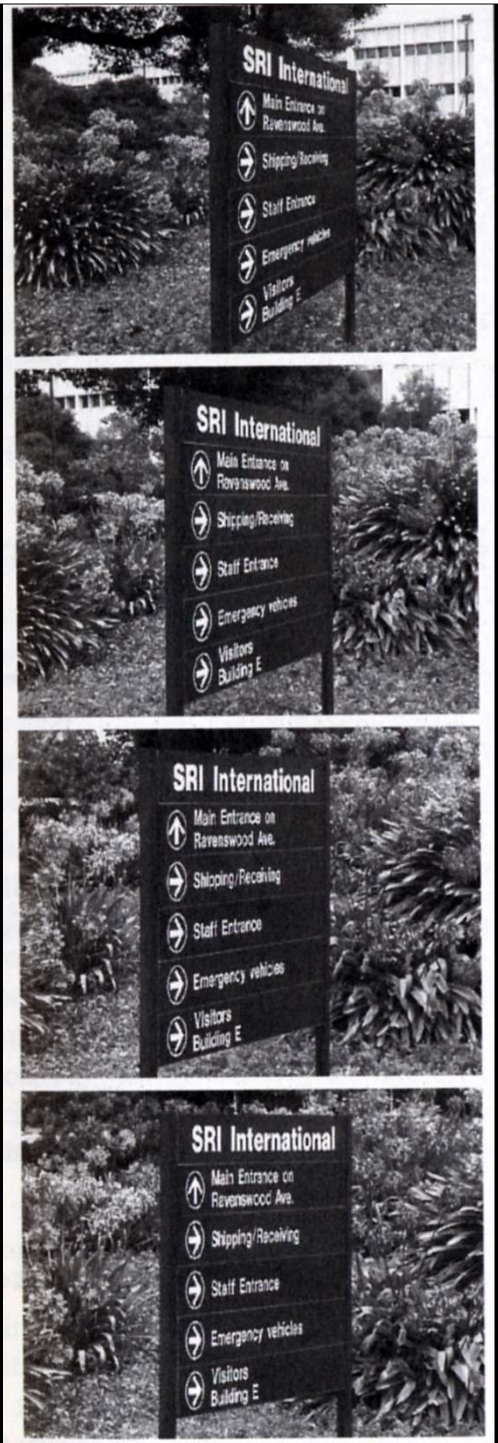
- Angles are not preserved
- The projections of parallel lines intersect at one point

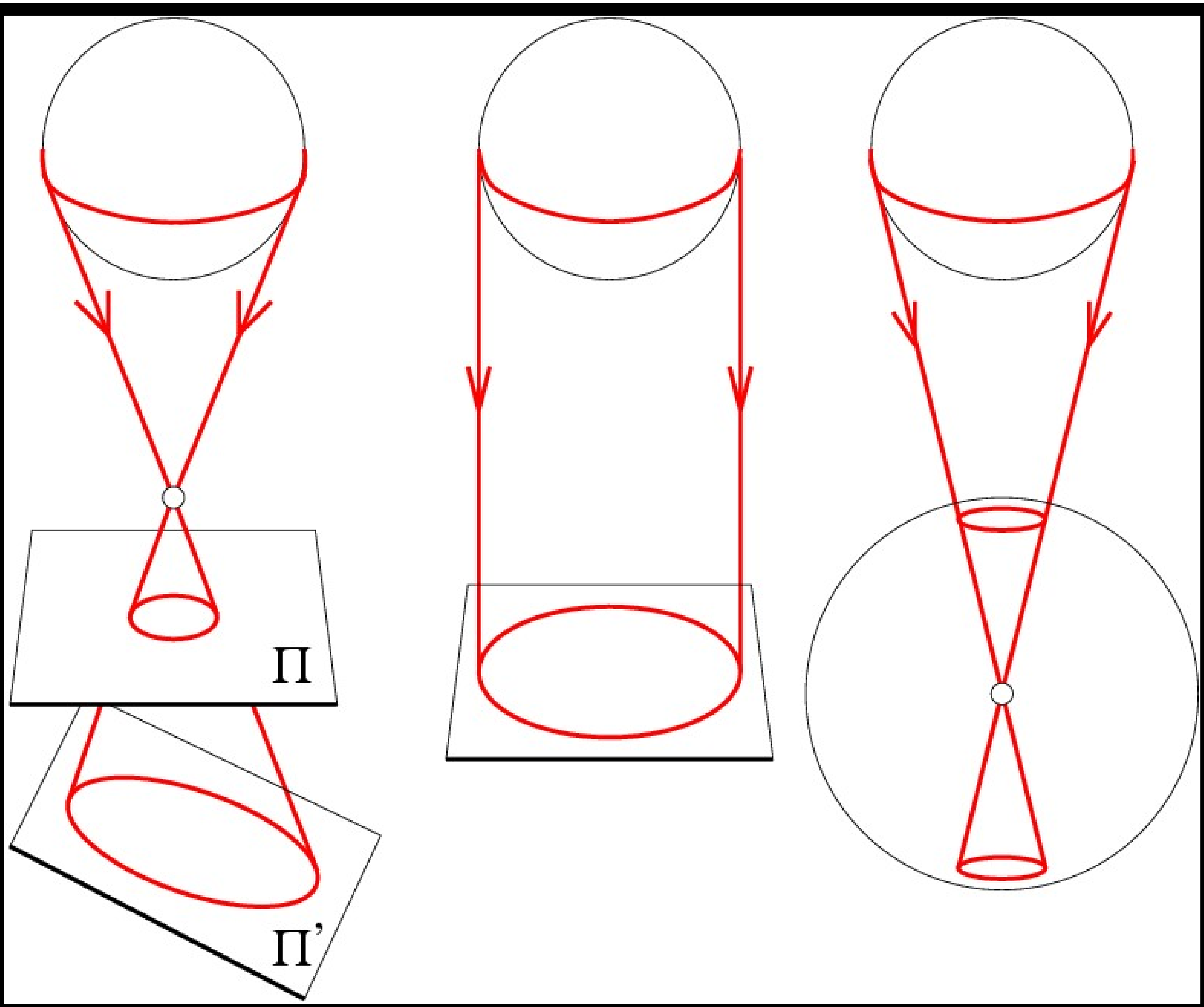


From Zisserman & Hartley

Strong perspective:
Angles are not preserved
The projections of parallel lines intersect at one point

Weak perspective:
Angles are better preserved
The projections of parallel lines are (almost) parallel





Planar pinhole perspective

Orthographic projection

Spherical pinhole perspective