

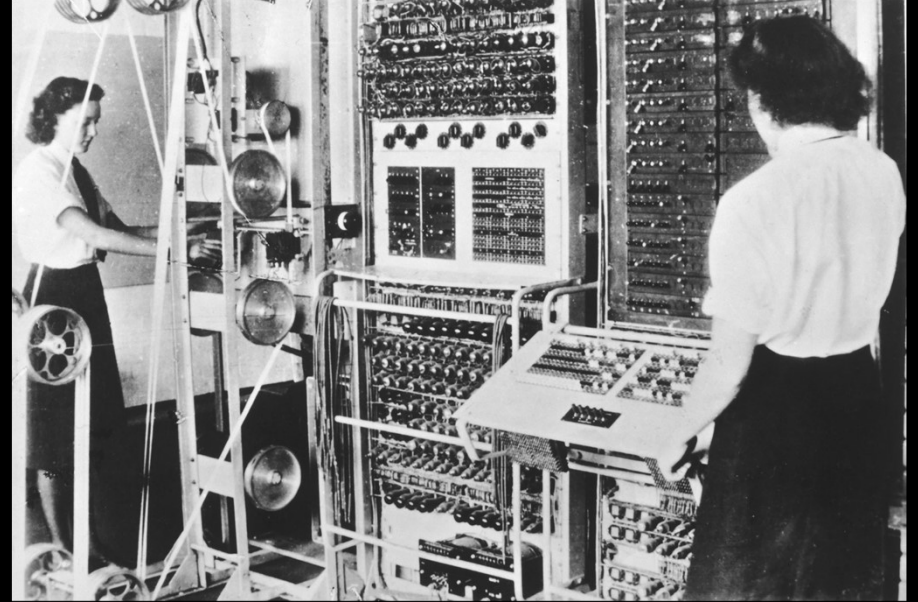
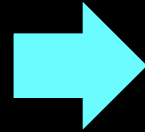
Introduction to computer vision



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”

08122635252121314133210507102023222326333642453122202526231614141320343524131413131328231714091819212527212744423326363323191924333640424044352
08133242231010244837260708101827272228343748543631272928231817161440615352181515151417261108071121242930253250473431383218192631263139394940322
09112942230507164739281118182026262630364853544234293027211816161533595448171415151414191409070816252529253754473536434024222531243135324540372
09092641394046495039292129282727272834405055573016202627221917161517272920141515151514152013080912252628223451463436403217202532223335334436311
09102343545752454839301412121528282835374752552915182527273226212020181818181918171817171814091013232628223654442735453622232532283332385341401
070918394518312238403214040813282828313339465235161825292633302625232323232628293131282413131208222727223648402735392916222432334043434748441
05081539421427223037311705091228303130343945487117142225243033302927252627262830323533292212171517263024203647382642453120252631243234454538280
08121433381714112636322105091027303129384042403110121924251713192624252930262827283229231613161623312923183240332237402716232530253331465448340
22251628314170112036312005060926282824434036301608101719171204172724223524100813112835140612171823303739303337251843442619262729243331394142270
23261923312528091531292205050922191318272929181413172221191503182425162517050912153638150506192321215771272936342938402217232828213131535042180
18191819281603071123201506051021110318232319161314161618191504111823093114000615122831100404233129222628232933293546412120252927172927424747190
05182414181606091223161411720221503151315131212091212121612070612214291504021007233310040421193525242212223243343402617172733342933473833100
05132513161614171918161507131619180406131607091515161721201711071031132817020410092734110506111024272927252122203450393741261719364142504837100
09102615151211141515161506111413140710181918182318101728181921111332081410040709093429070508030112231907251826263148403638303228181736363630070
11082417141309121313141412151719161109101010121508041322201917101635110505070806083313040722230804131506211830303338303535252326313340453821050
57222117091417221622211625232020192014132019100500000307140602051313040304060605153614091114171904020205181215042824212632312624233635364022050
48241320091319351719191524272425262922182212030201030306090505081306020303040405121412141004050802090804071111081711142330282929263739363219301
08041022091312352328210907101410060509121109040708090907060805060707090806050504070703060607050604101610040407685103072134332722213237383617363
05040413060708342219080301000202050508131107131914110810071209020408131210111110080407080825000508041519050211858703032546443311162631343116362
02050803030407100911040002010000060508172238161104020101020304050609111111101108020102071010010709051713030403201205043129233312122425272118372
0805050304050503040604000000404000203071119030503010202030305060708090911101111030101021316030913041411020403072006033322131712142725241616262
0502010304151403060813120806111404010100104050804020203040506060707080913101117140500021507011220041410020203061804042912081511202522241416232
04040304041616132019131317251214150302030306050504040404050606060707101918121219261101010901011623070802030203071303052207050506162308151021242
05060810101517050602010309181313190703030406060606050406050506071016263020141216382606030200001517020203051417121103051014120706202204101124262
14171714091405021401020305060503121201020406060606060505060605071223363823111219373316090503060905020406051619150702020304090205231603061424333
101412240915030410040405050302010403020203050606060606050607181020344729171723353520100606070709050505071217060300010101000006210305121335623
13090707081006040303020203080400010303030306050606050812050606070202233730181725475025090606061215070606091107000302010202010002060308160870996
151409070403020101010101080400020304050404050607050504030606020311291020191725444319090606071214120807161708010101020202010000000108131162976
21090302010101010000000000101020304050606060606060911103040709030402030717192227231613110909081012110825351606000002020100000000005061525331
080212070000000000001020303040505050607070808080708173223040829080040608111825354134251711100906070909294219110400020201010202020100000610050
01020503010101010203020304040606060708080808090907040708180104131100021634242148444437272015110907050707284424150802030100010102020100010101071
0206020404030508120705030405060607080909100909090604030202020105160105171309104449453631241913110805060518432326140704010000000010101000101010
0002091107040711130904061006060707080809101010100905030302020206161513021210044757494135282216151306050513392128241009040302000000000000101000
0000030505060404020302051807070707080808091010100806020201010318042635020400195453514437292318121609040509293326261209070202010100000010100000
0000010405060300000202181405060606060707080910090606030202010127090304010101275349494739312318121015070504153336220813070402010100000010201000
000000102040100000100140902030303040404040506070404020201010113060100000010936373535292116130906080703040916301604130504030101000000000501000

- 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:

- $1000 \times 1000 \times 24 \times N$ bits;
- matching n features against n features costs $n!$;
- shadows, highlights, texture..

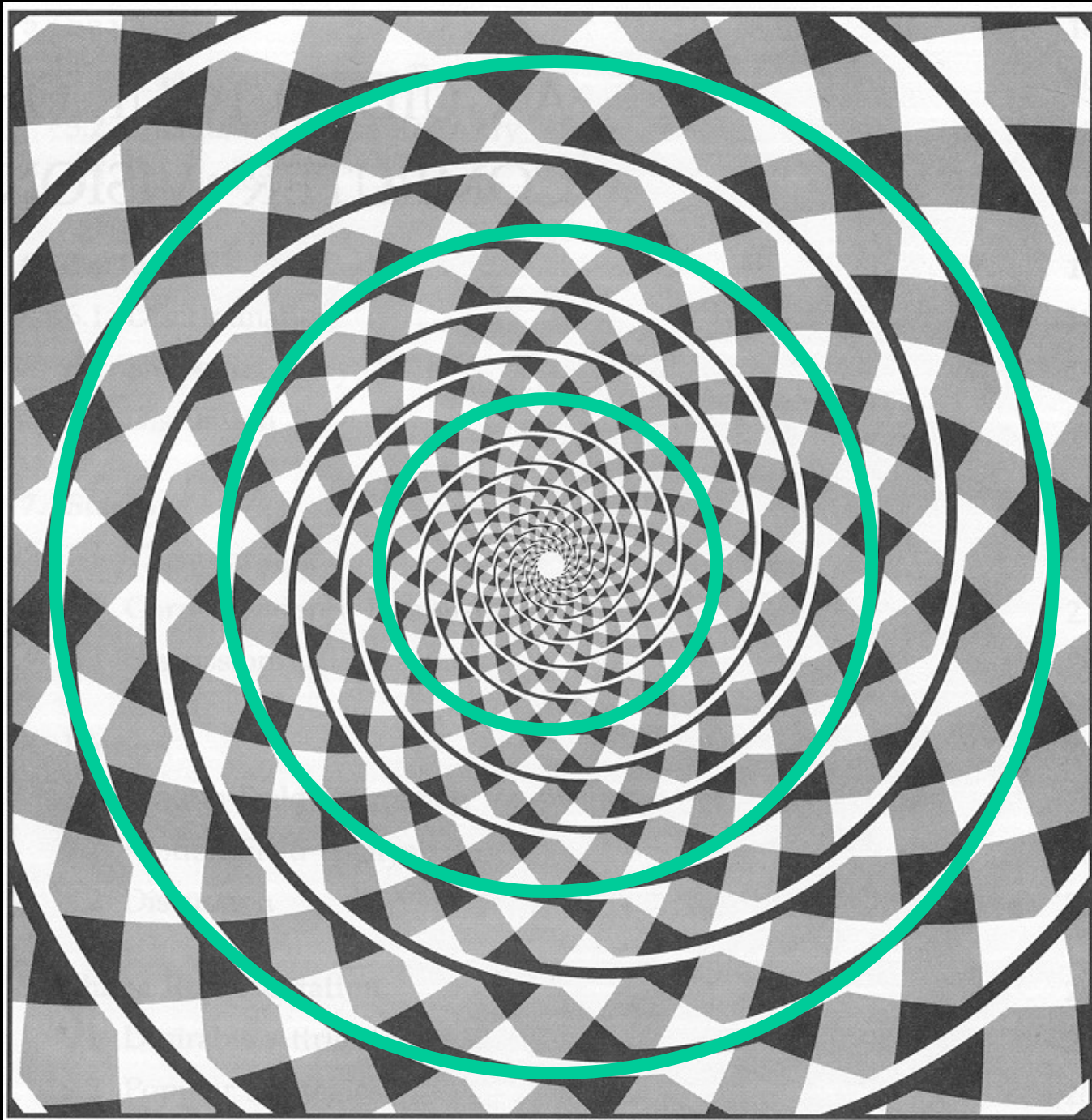
Too little information:

- 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)

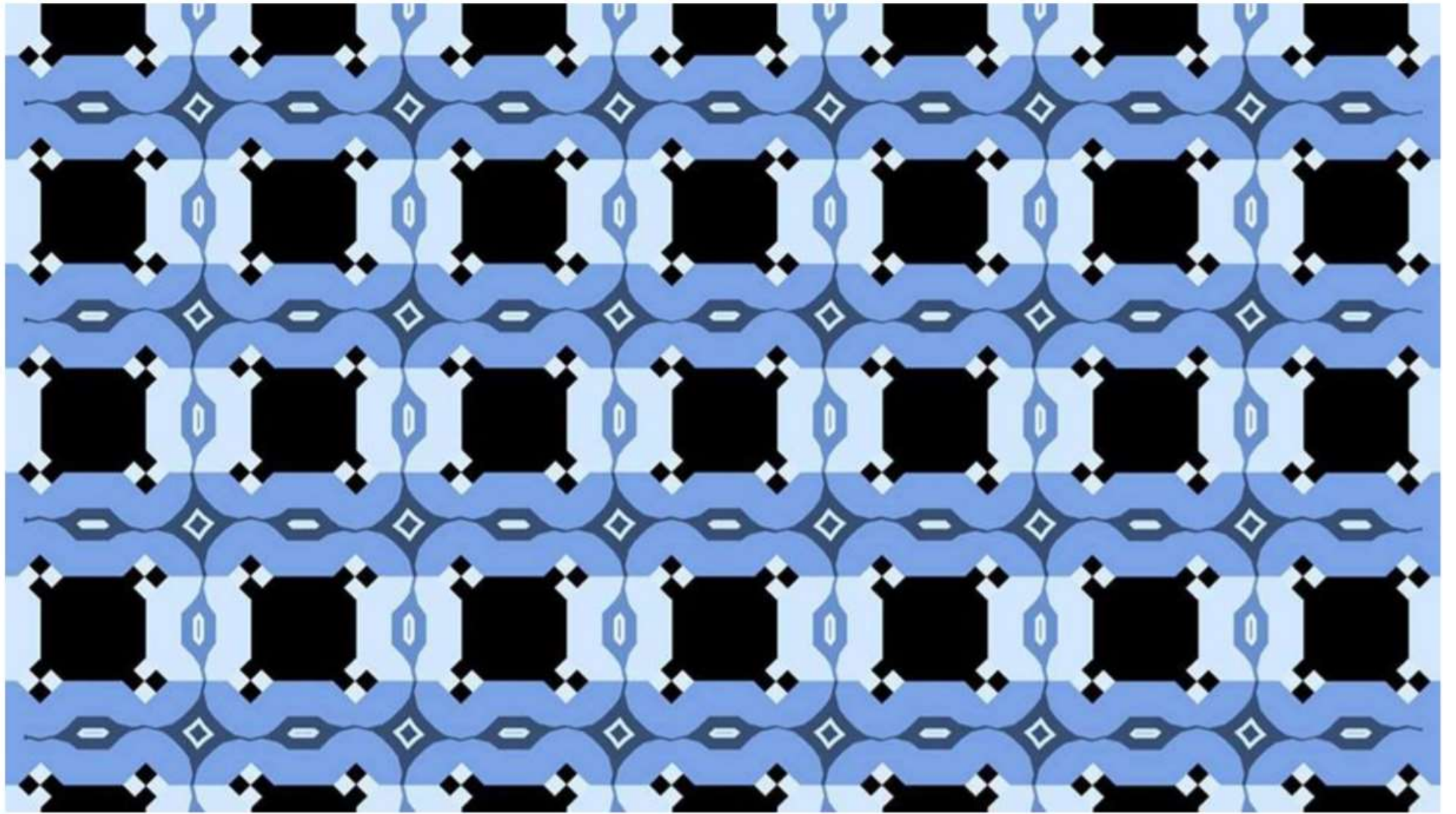


<http://go.funpic.hu>

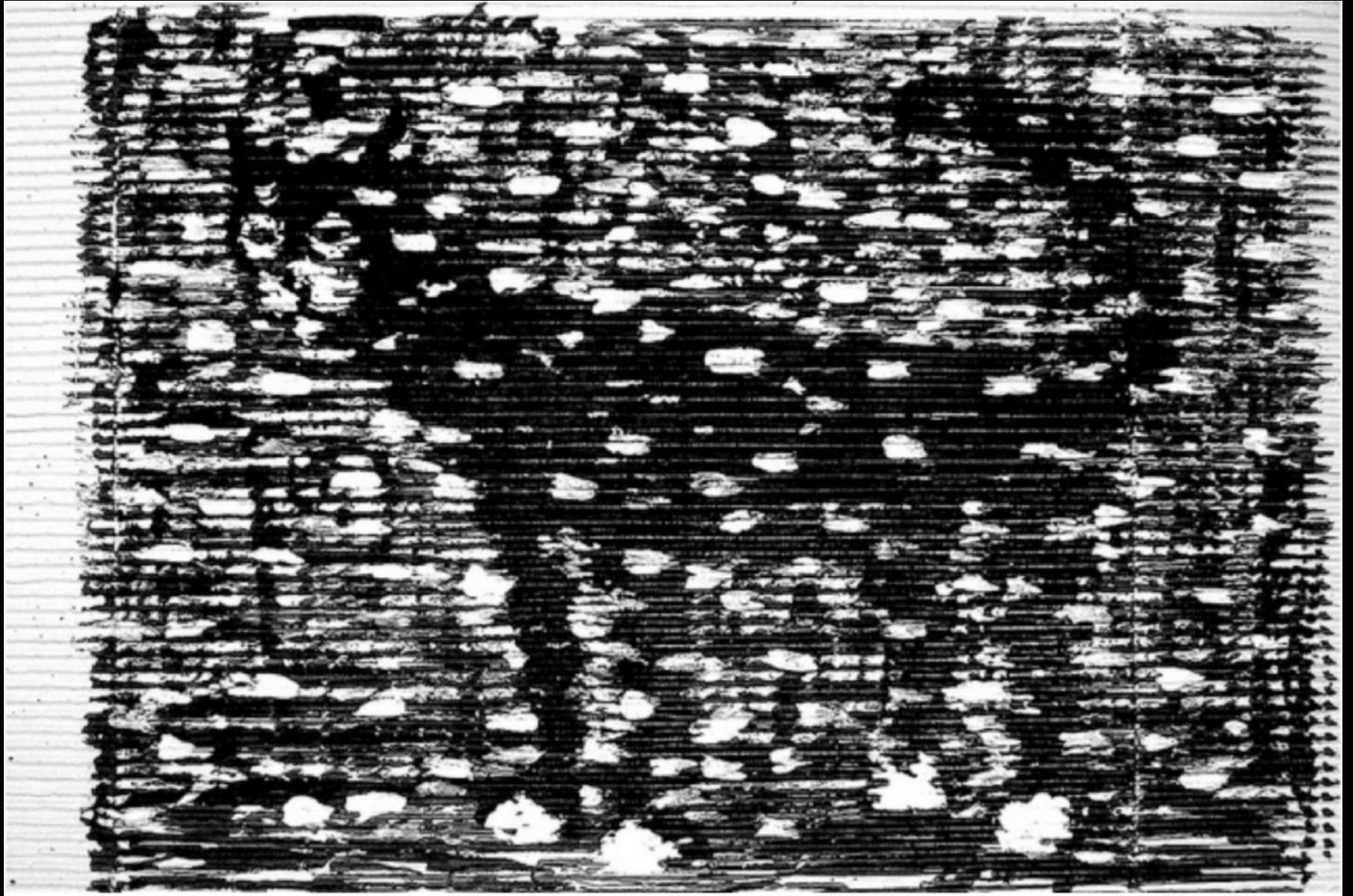
J.J. Koenderink, www.gestaltrevision.be/en/resources/clotcrans-press

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)

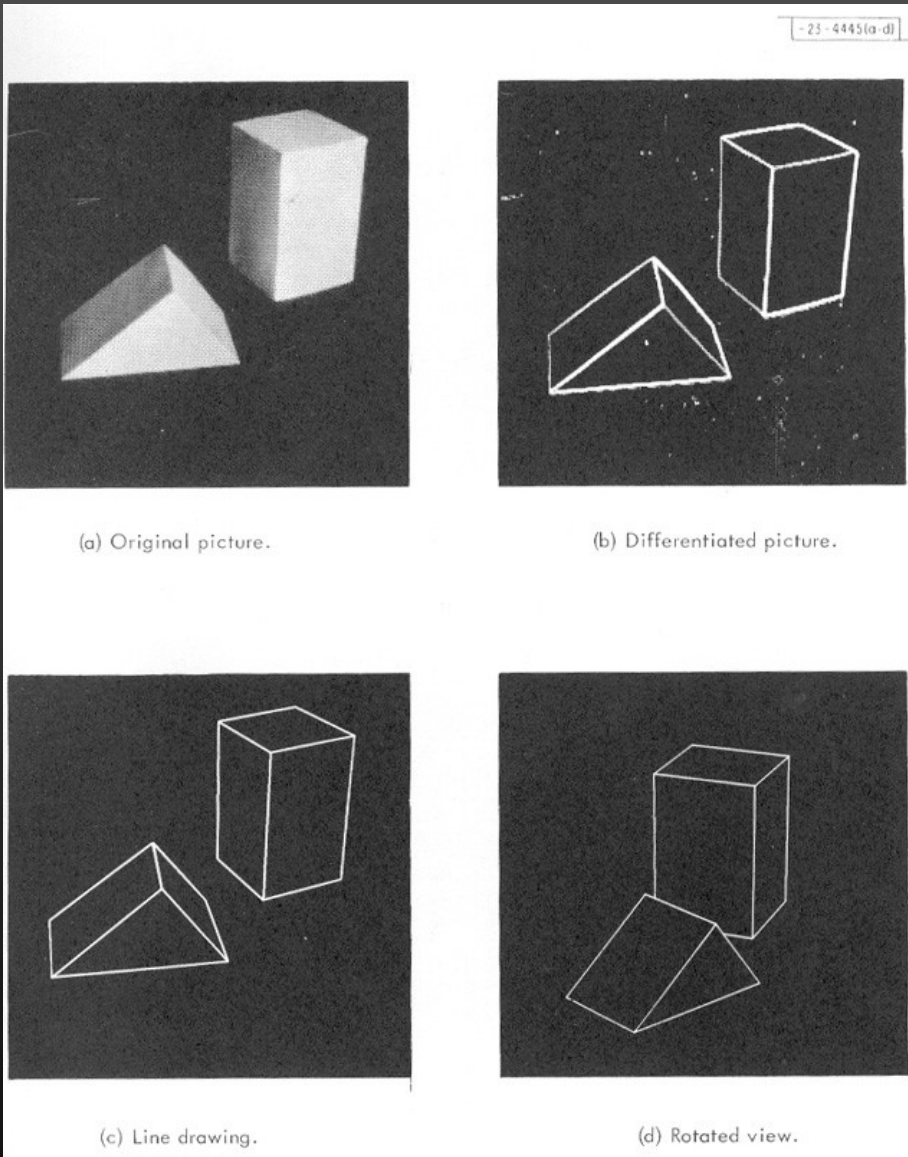


(Franco Mattichio)



(Pau Buscato)

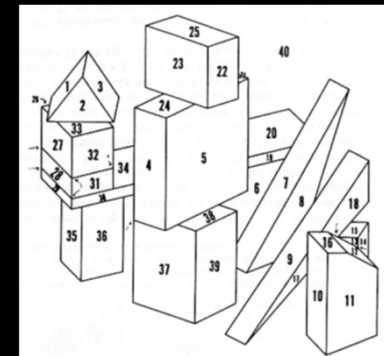
Origins of computer vision



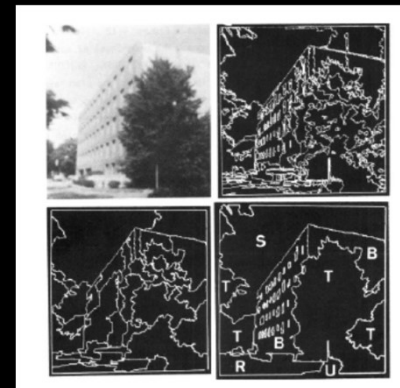
L. G. Roberts, Machine Perception of Three Dimensional Solids, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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:

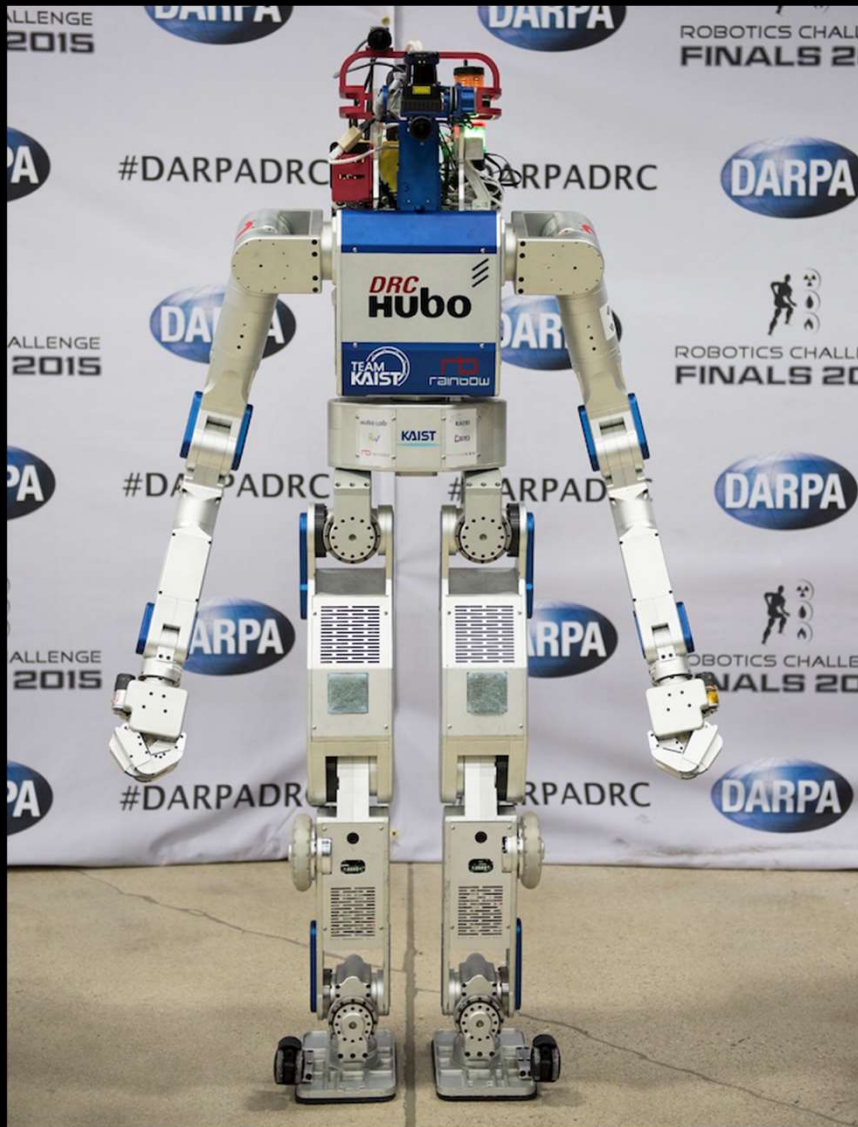
- Manufacturing: inspection, bin picking;
- Defense: ATR, photogrammetry, surveillance;
- Robotics: navigation, visual servoing.

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



CMU's Chimp

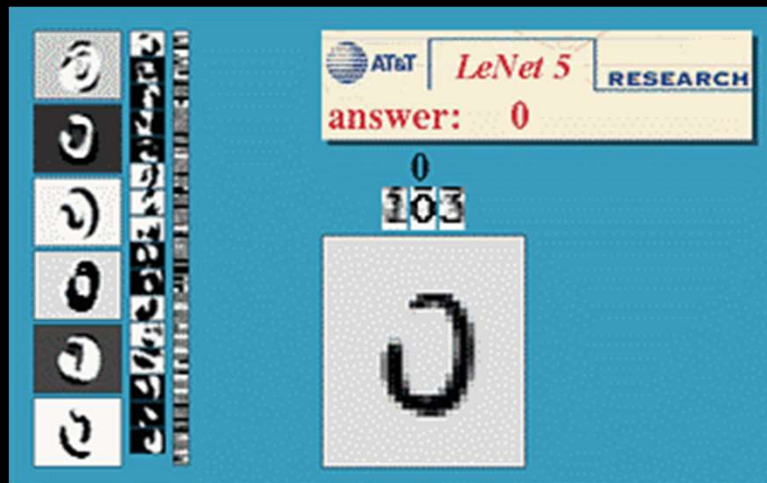
How vision is used now

- Examples of recent real world applications

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

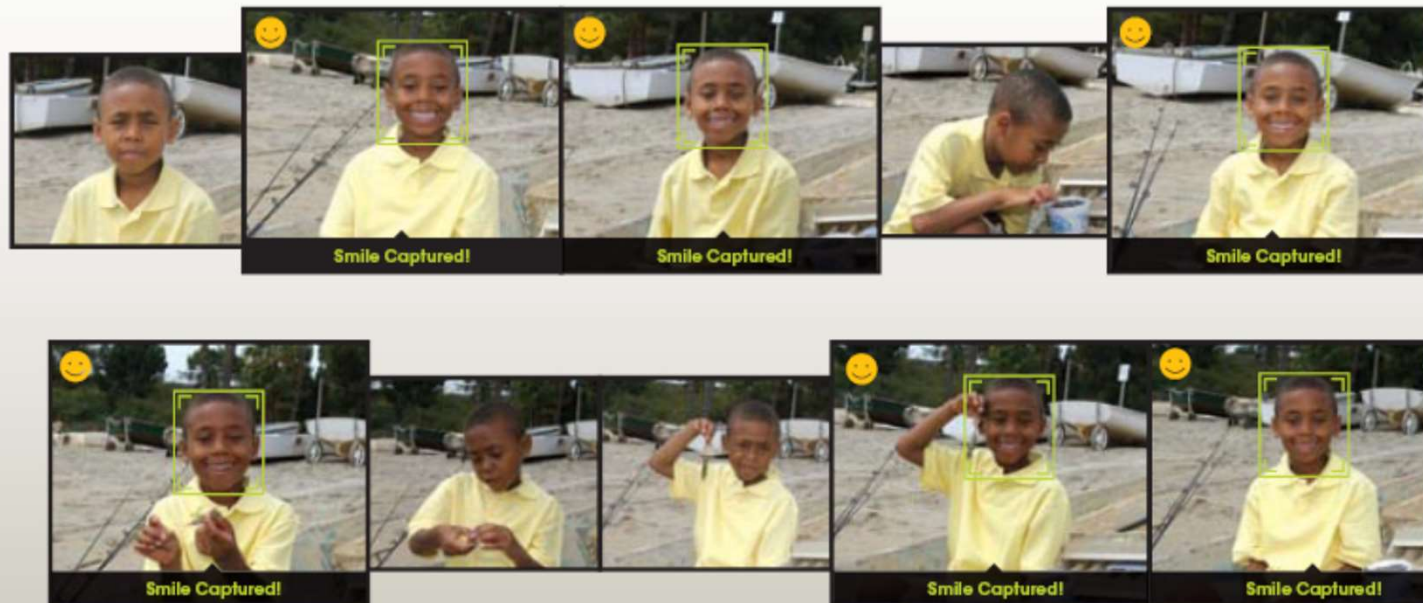


- All digital cameras detect faces

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

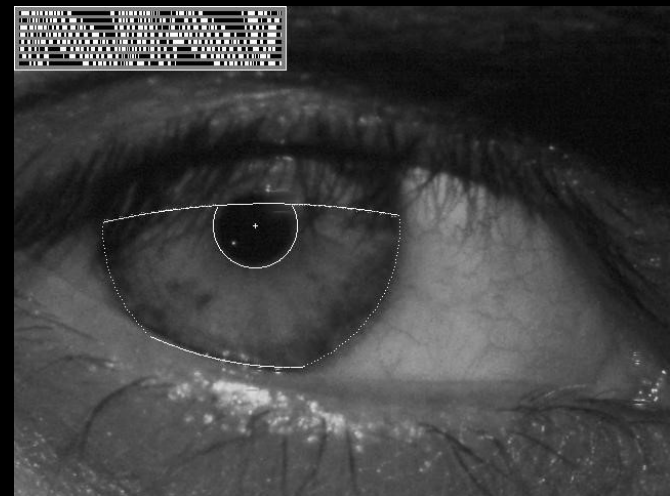


(Agarwal et al. 2009)

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 Caribbean, 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

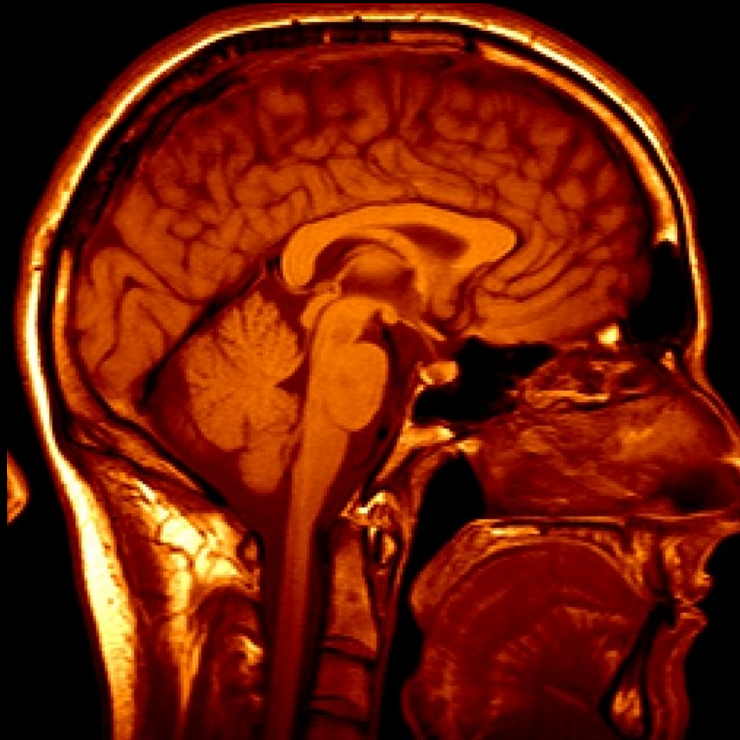
Sports



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

The screenshot shows the Mobileye website interface. At the top, there are two navigation tabs: "manufacturer products" (active) and "consumer products". Below the tabs is the slogan "Our Vision. Your Safety." and a top-down diagram of a car with three camera views highlighted: "rear looking camera", "side looking camera", and "forward looking camera".

Below the car diagram are three main product sections:

- EyeQ Vision on a Chip**: Accompanied by an image of the EyeQ chip and a "read more" link.
- Vision Applications**: Accompanied by an image of a pedestrian being detected and a "read more" link. The text below the image reads "Road, Vehicle, Pedestrian Protection and more".
- AWS Advance Warning System**: Accompanied by an image of a circular display showing a car icon and a "0.8" reading, with a "read more" link.

On the right side of the page, there is a "News" section with two headlines:

- "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"

Below the news section is an "Events" section with two headlines:

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

At the bottom of the events section is a "read more" link.

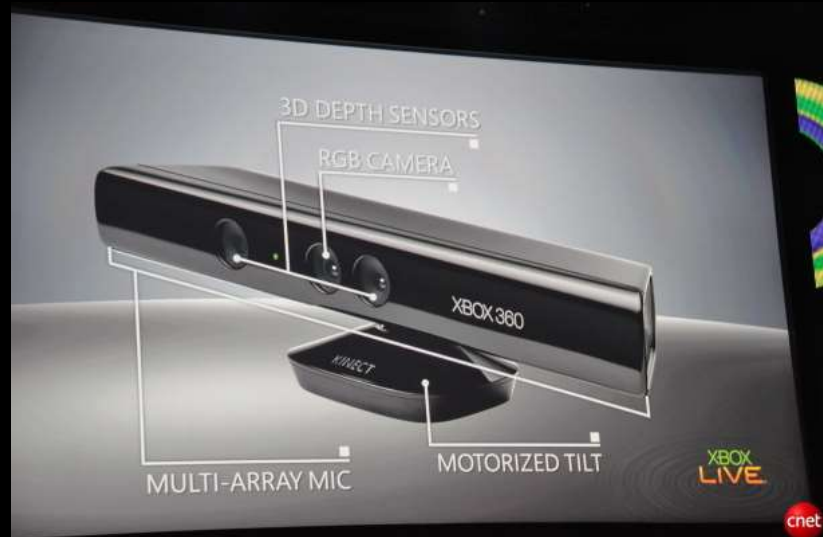
- Mobileye
 - Market Capitalization: 11 Billion dollars
 - See also CVPR 2016 keynote



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=8CTJL5IUjHg>
- 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.

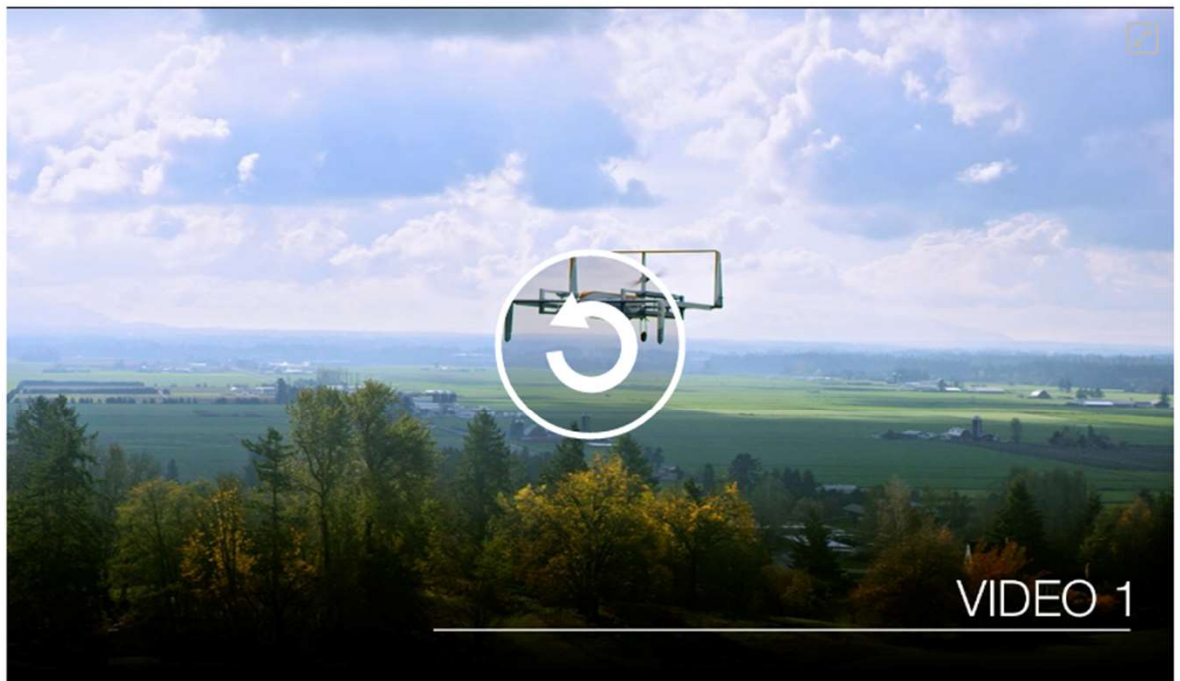
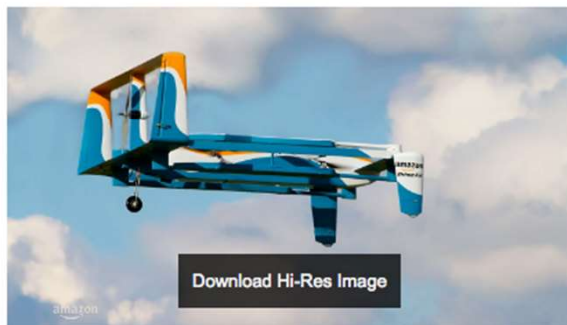
Vision systems (JPL) used for several tasks

- 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. Papadopoulos, "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/cloutcrans-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

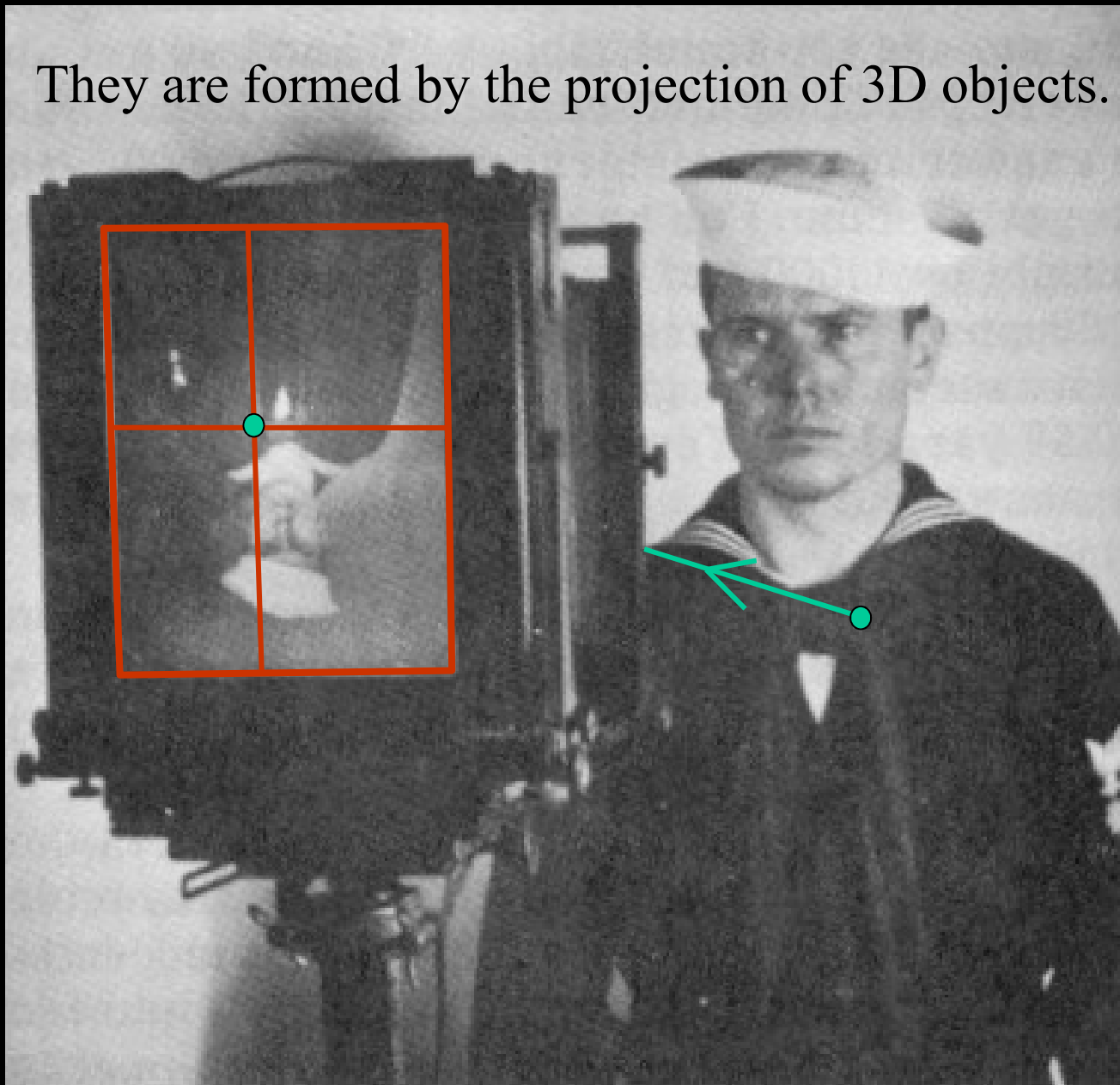
Programming assignments + final presentation

Course outline:

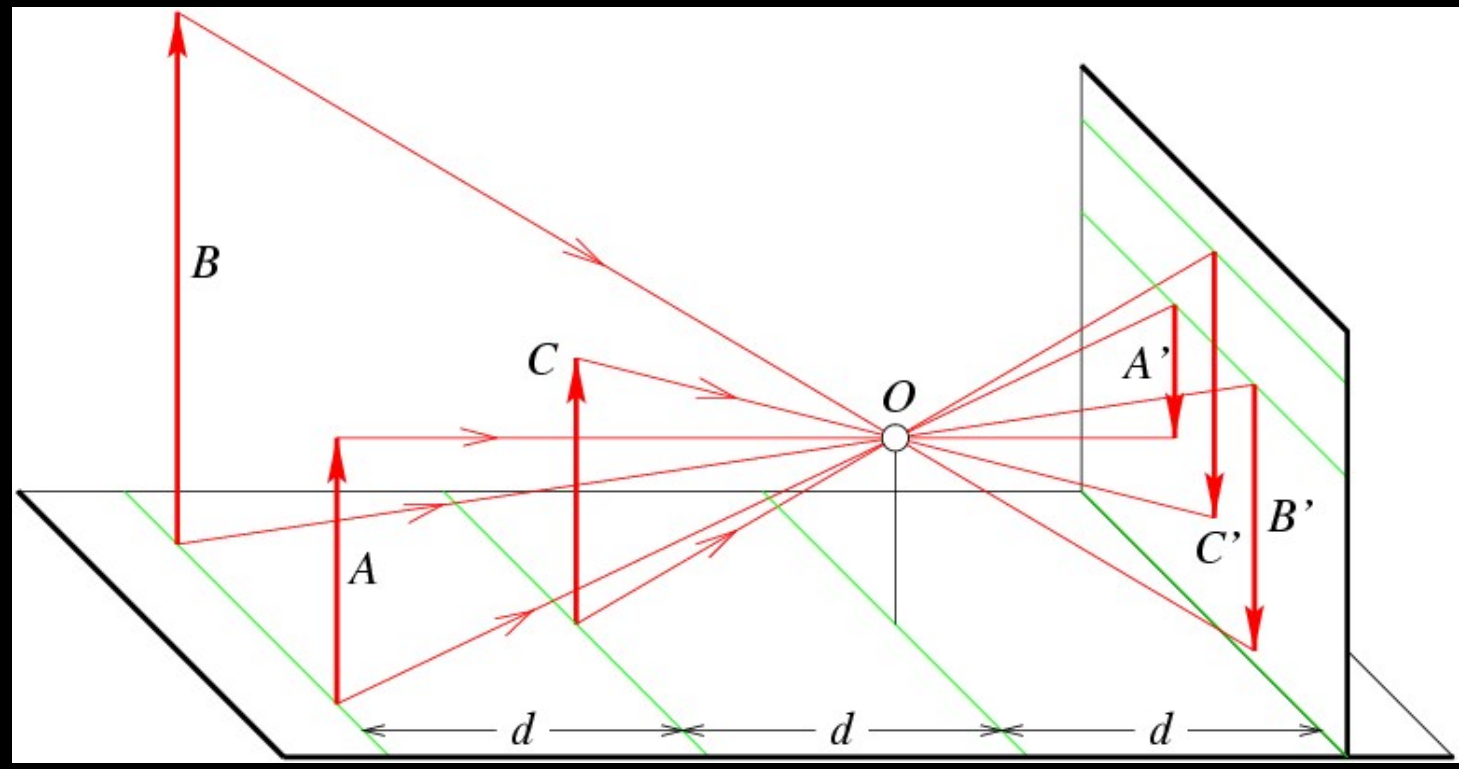
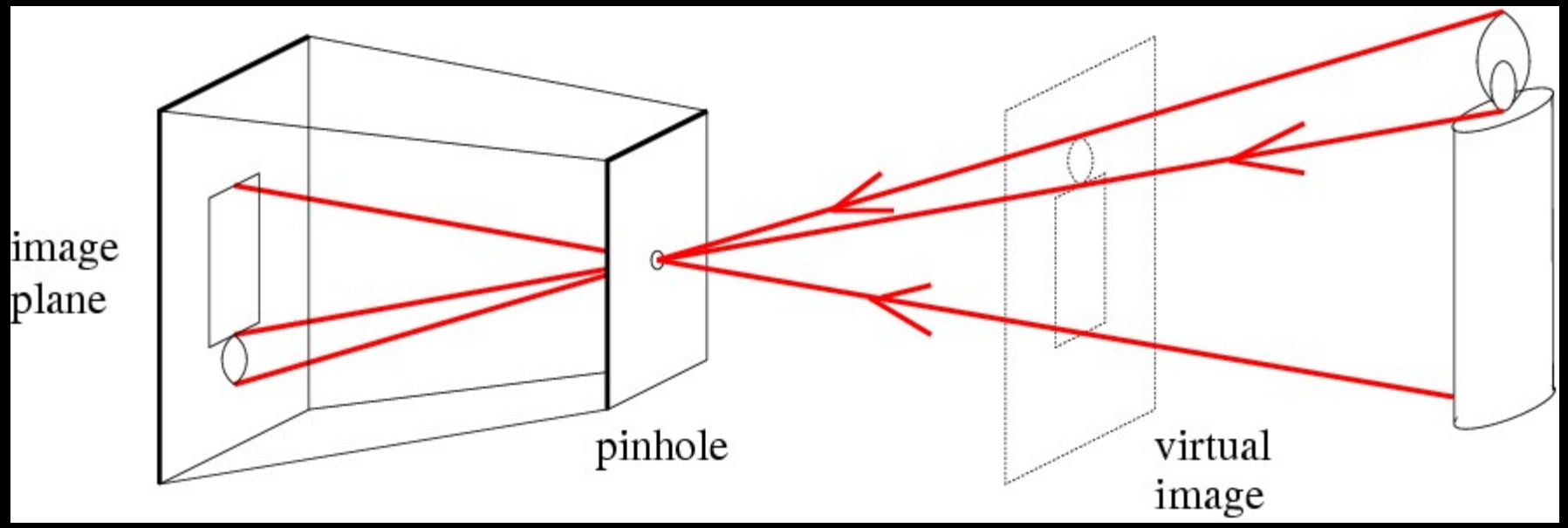
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

They are formed by the projection of 3D objects.



Images are two-dimensional patterns of brightness values.



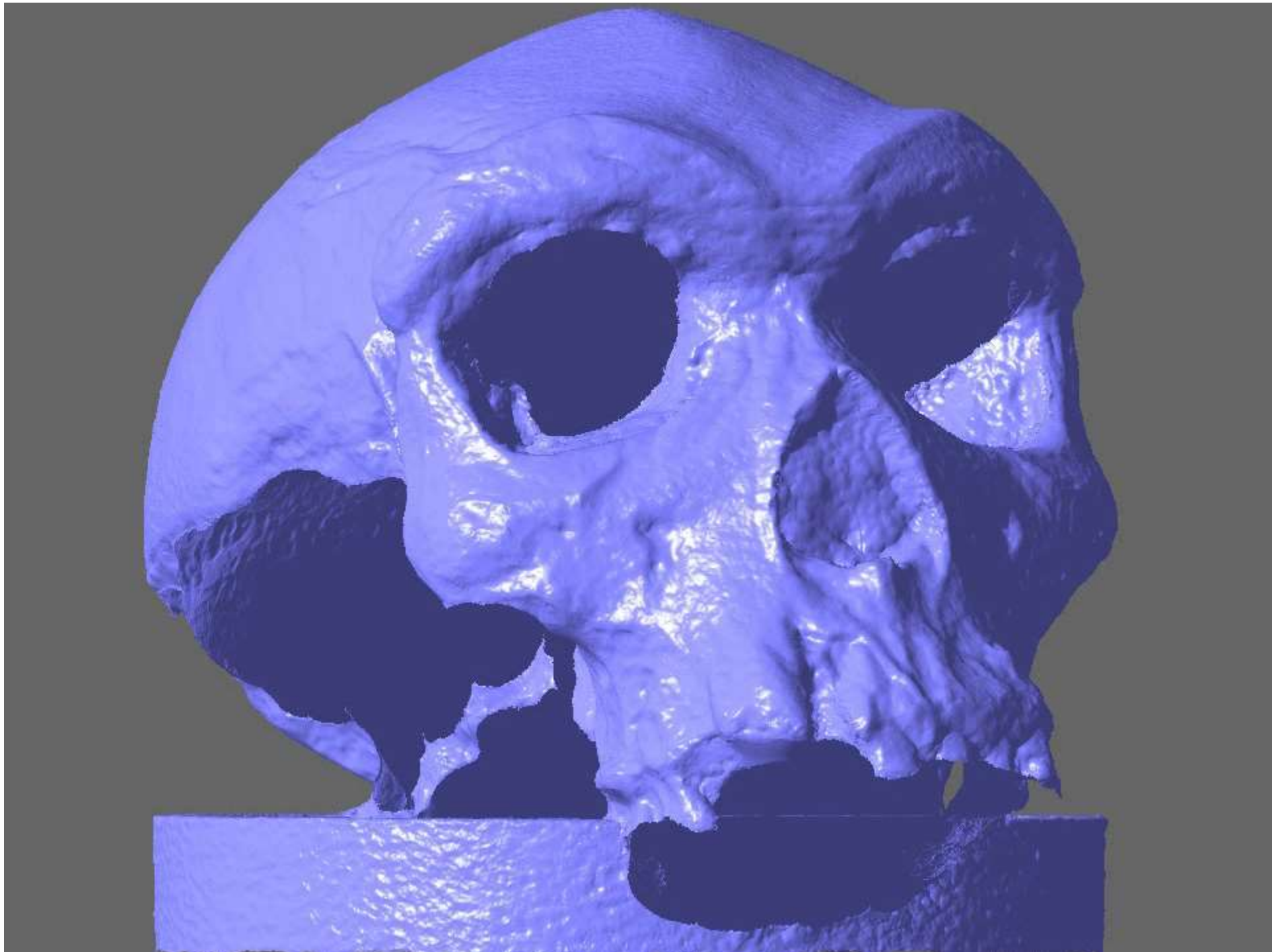
High-fidelity multi-view stereopsis

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

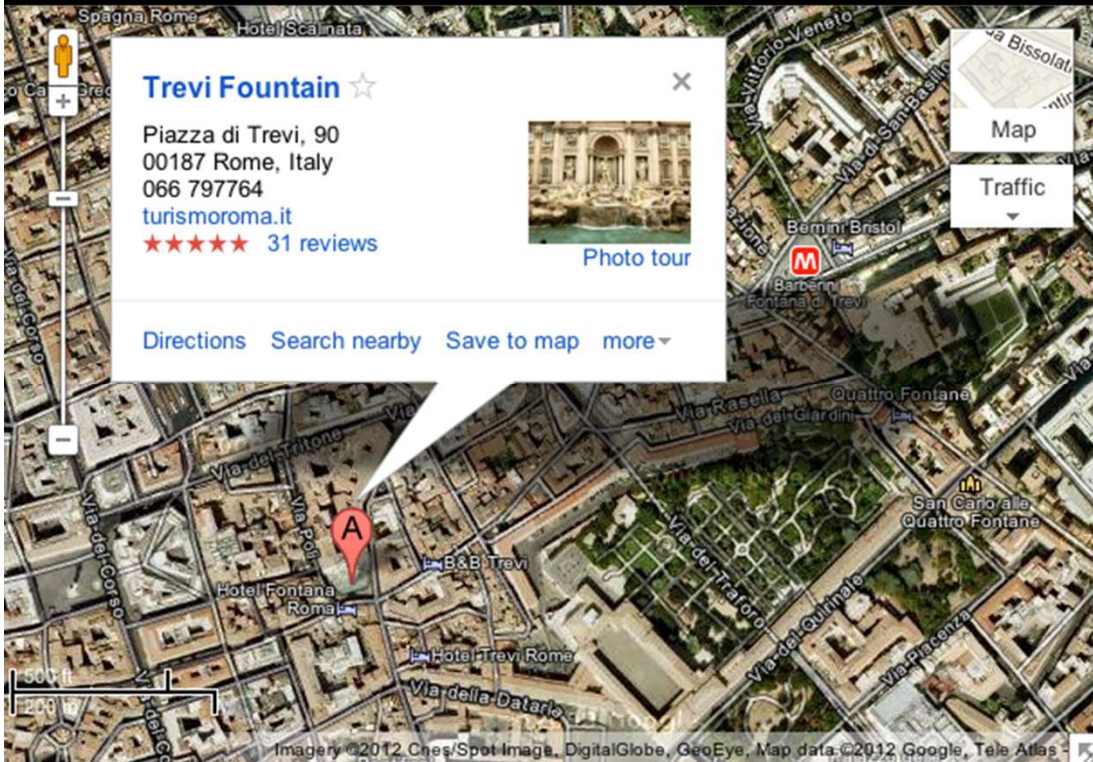
<http://www.cs.washington.edu/homes/furukawa/research/pmvs/index.html>



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)



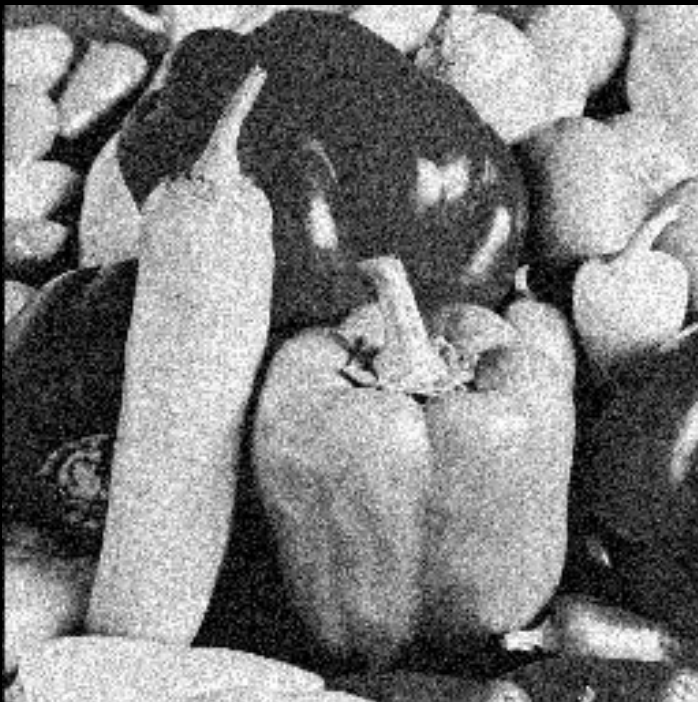
- Google Maps Photo Tour
- Lucasfilm
- Weta Digital

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. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

Filtering



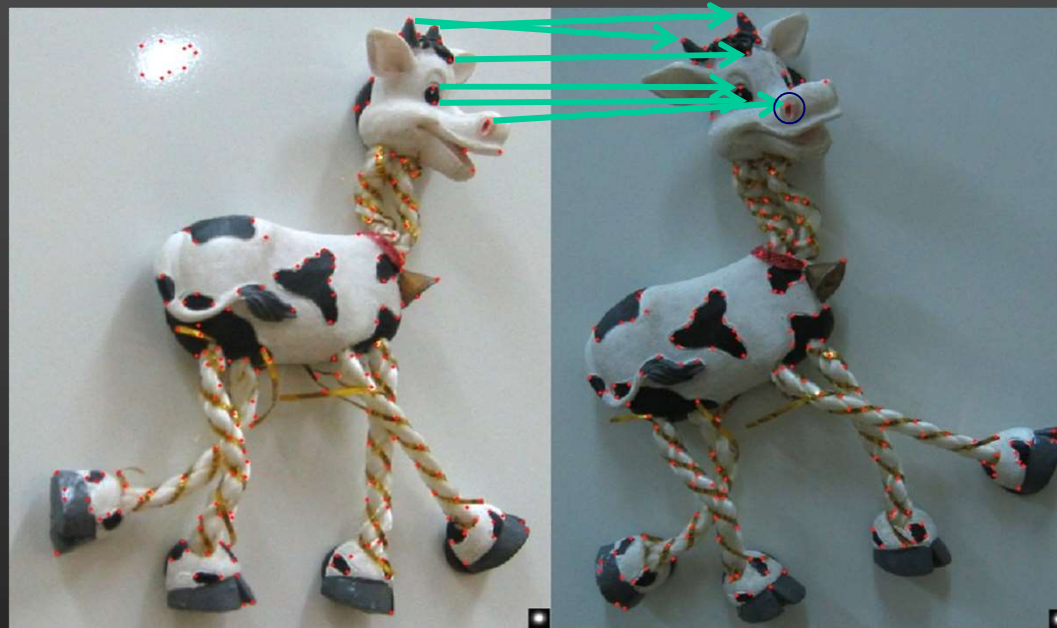
Edge Detection



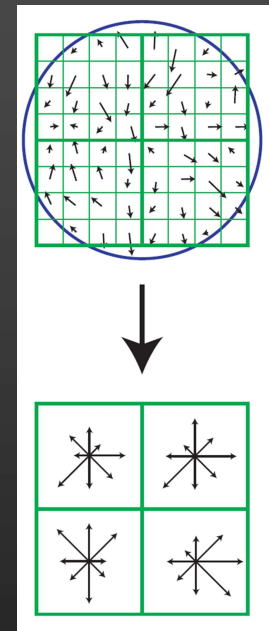
Edge Detection



Interest points and local appearance models



(Image courtesy of C. Schmid)



(Lowe 2004)

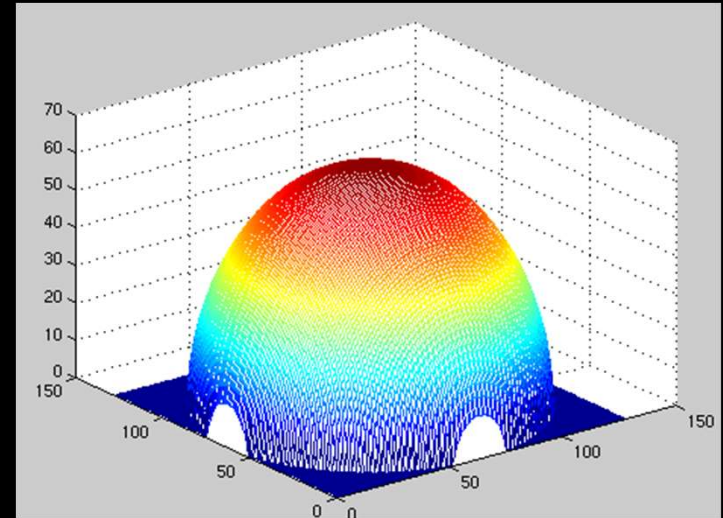
- Find features (interest points)
- Match them using local invariant descriptors (jets, SIFT)

Course outline:

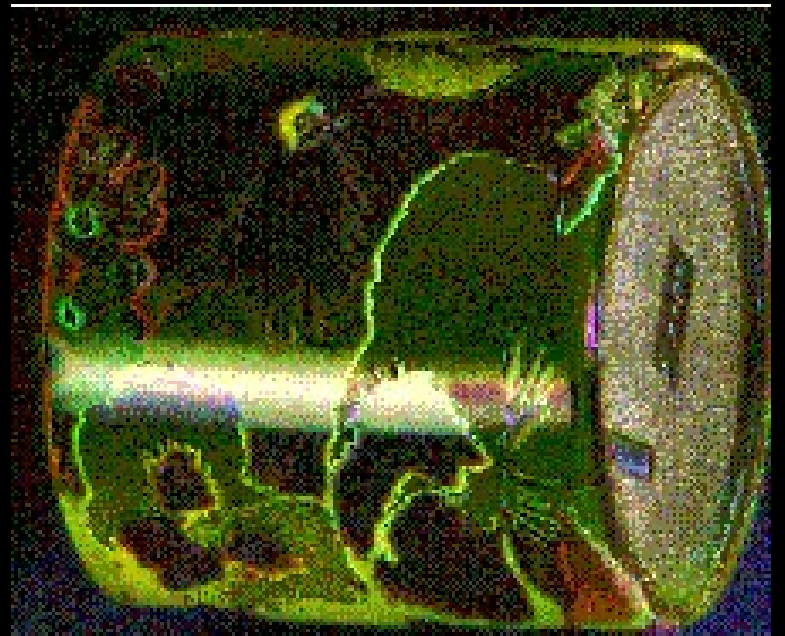
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

Radiometry/Shading



Color

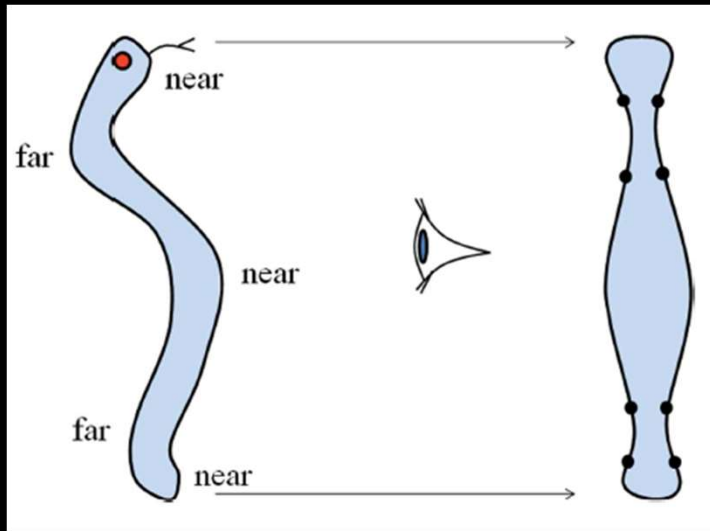


Course outline:

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

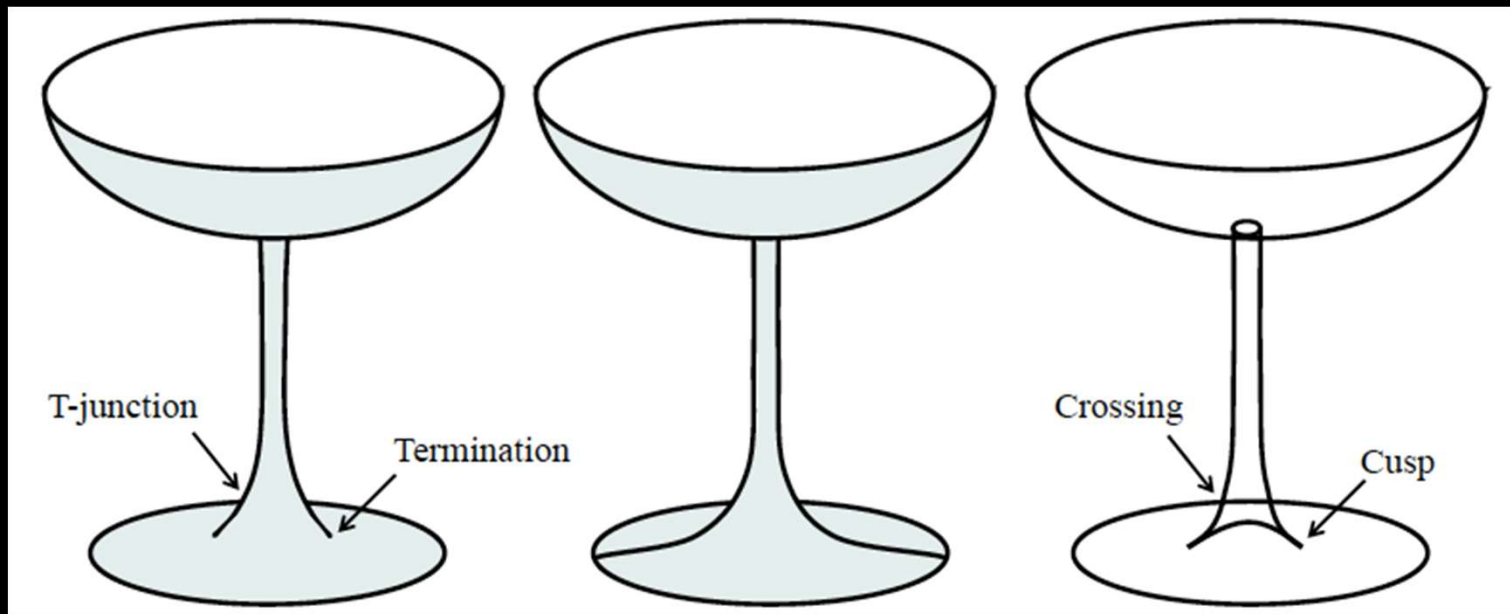
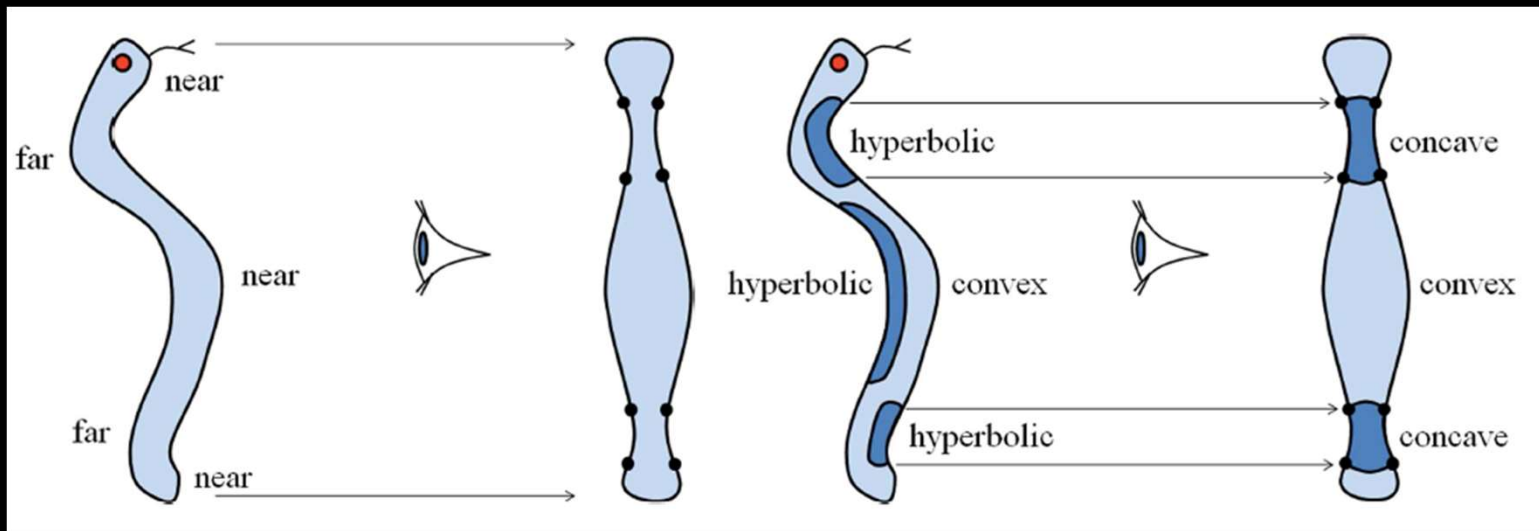
Programming assignments + final presentation

One-view (differential) geometry



(Marr & Nishihara, 1978; Koenderink, 1984)

One-view (differential) geometry



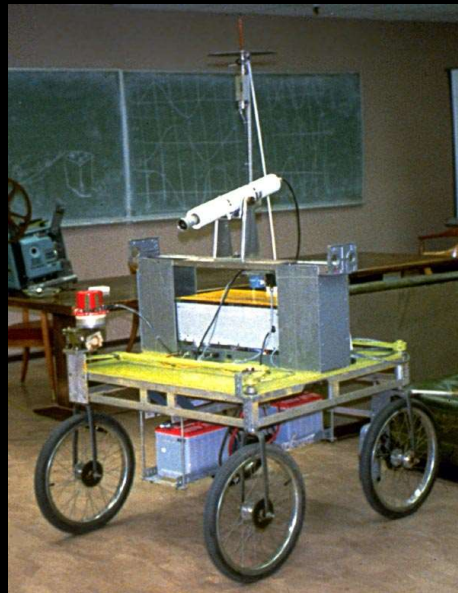
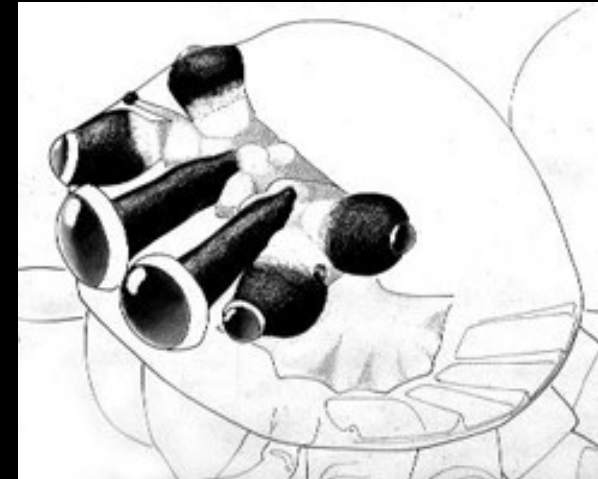
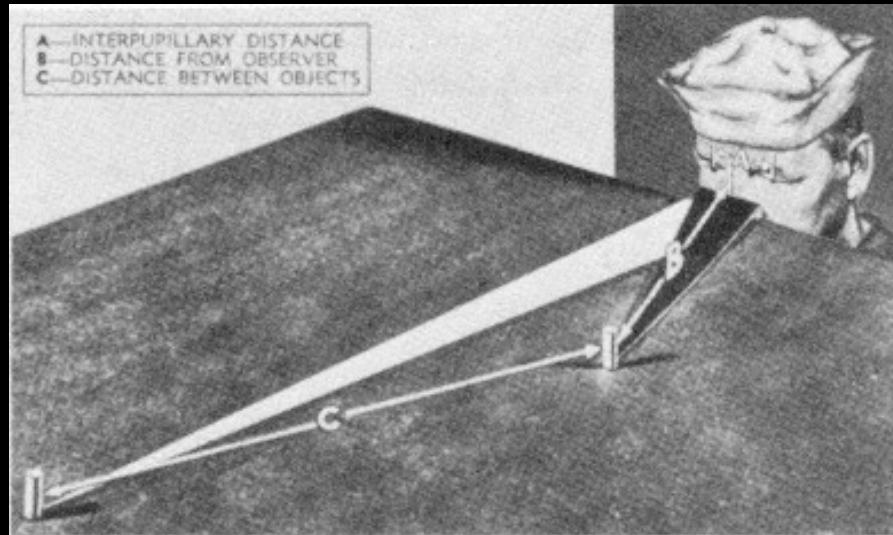
(Marr & Nishihara, 1978; Koenderink, 1984)

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

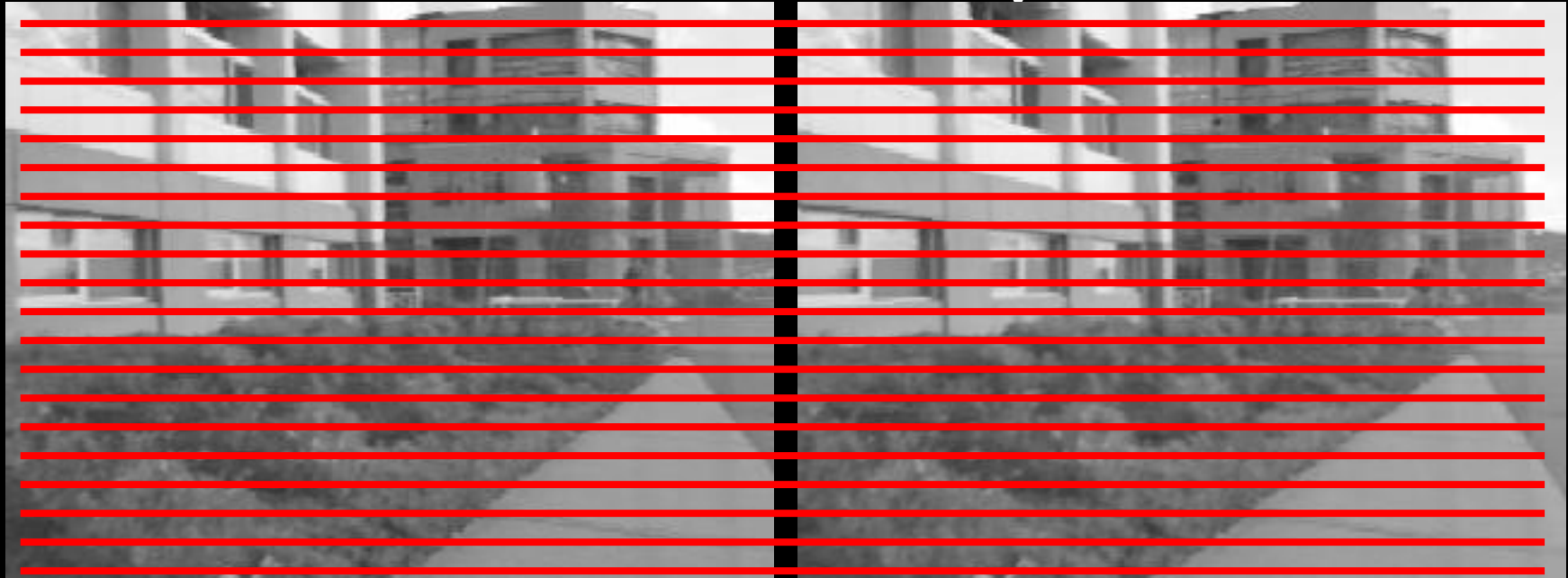


Method:

- Find correspondences
- Along epipolar lines



Two-View Geometry: Stereo



Epipolar lines for
rectified cameras

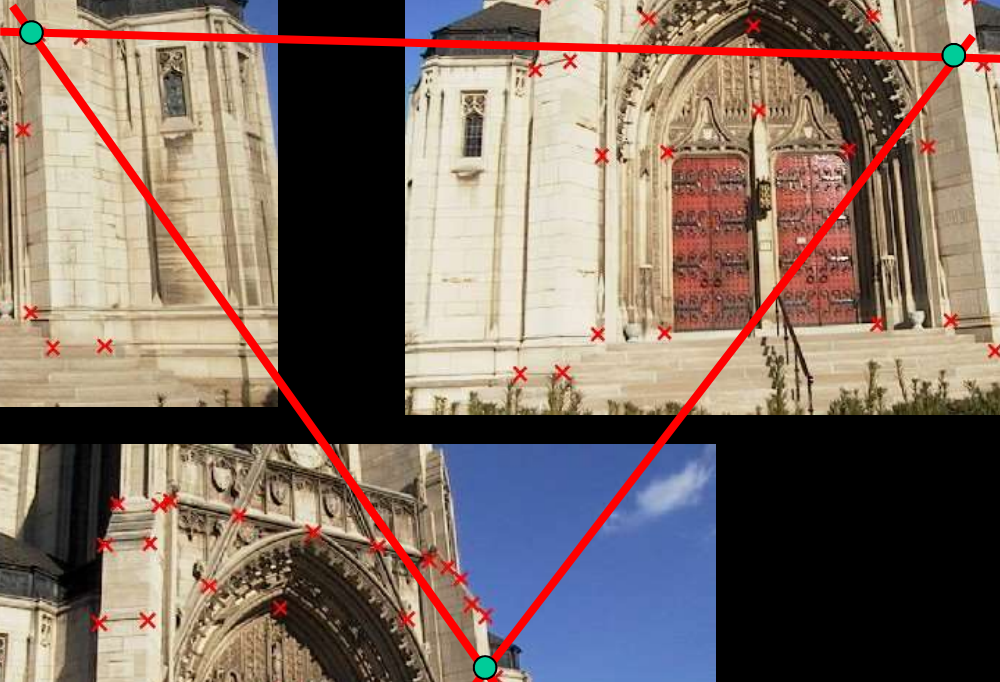
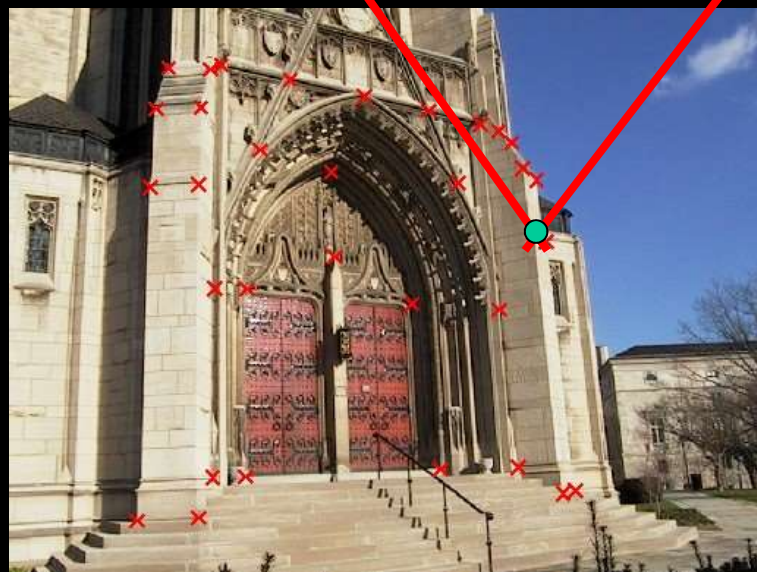
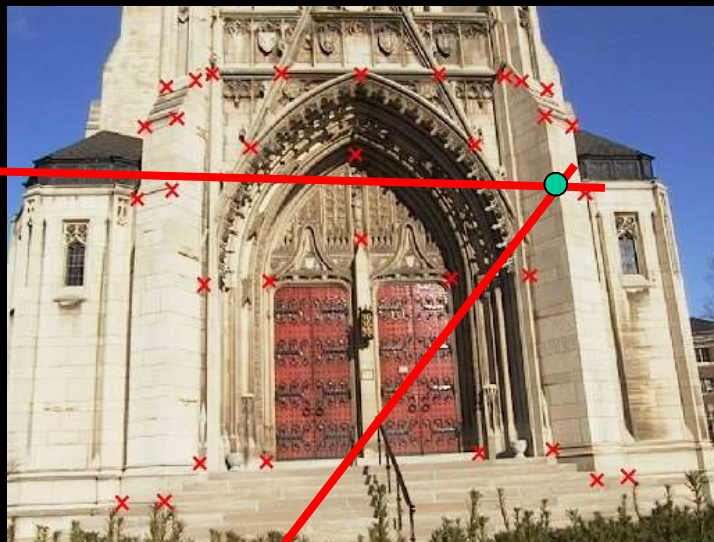
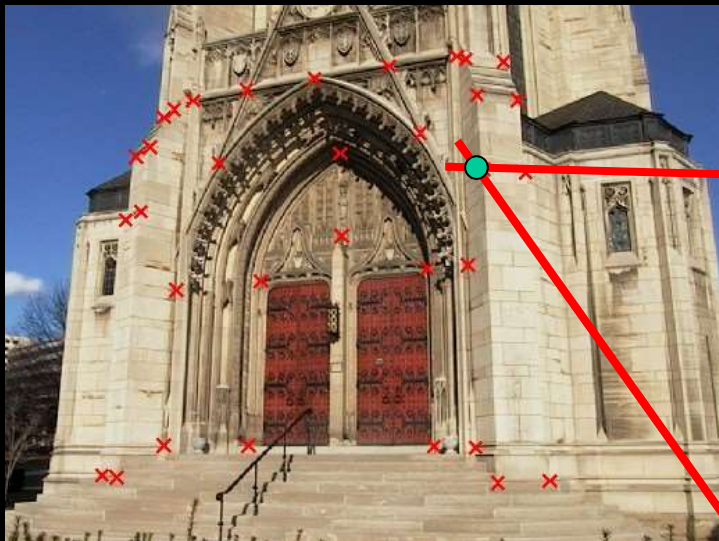


Course outline:

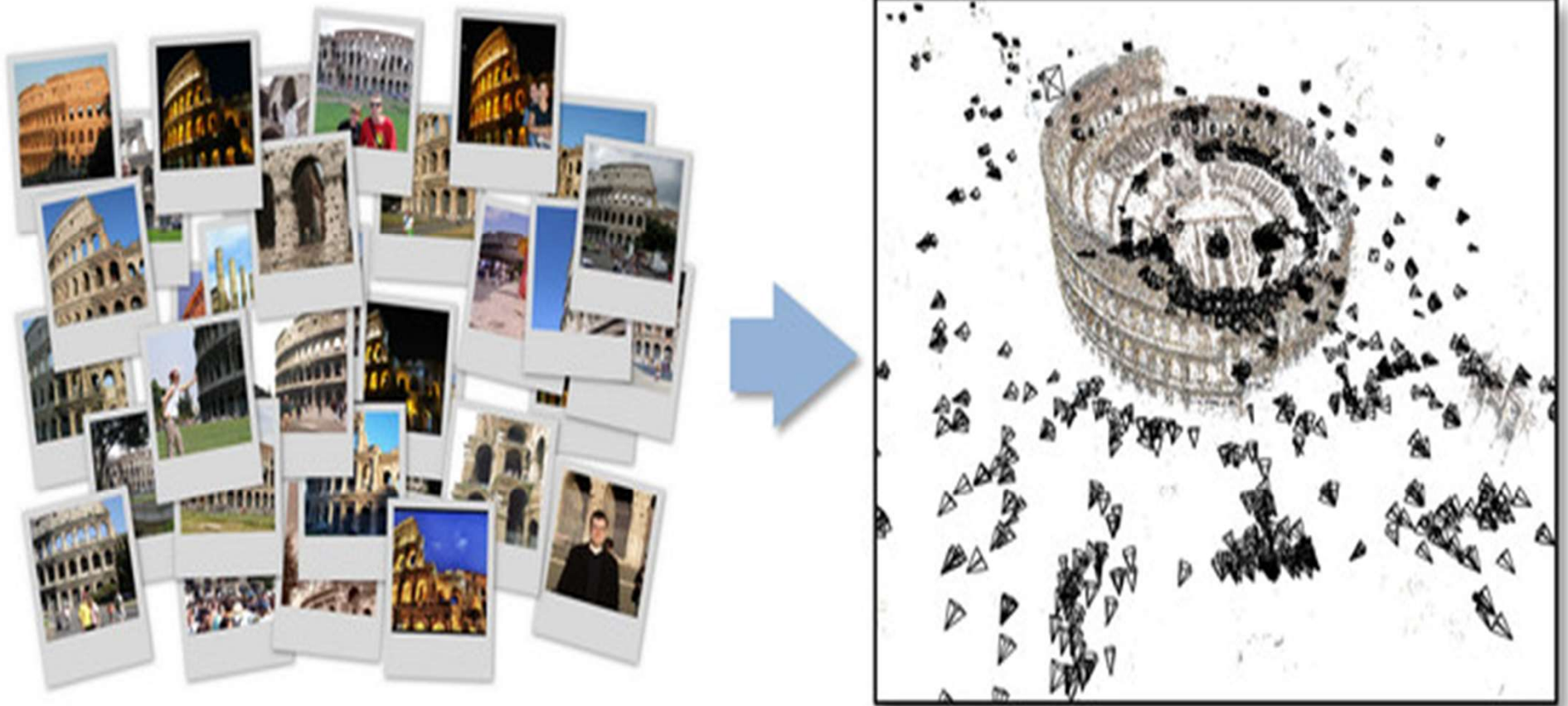
1. Camera geometry and calibration
2. Filtering, edge and feature detection
3. Radiometry, shading and color
4. One-view (differential) geometry
5. Two-view geometry and stereo
6. Multi-view geometry and stereo, SFM
7. Range data
8. Segmentation
9. Recognition

Programming assignments + final presentation

Multi-Camera Geometry



Phototourism



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



face2

400 frames

10 cameras

(Furukawa & Ponce, 2009)

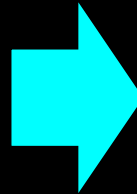


Course outline:

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. Range data
8. Segmentation
9. Recognition

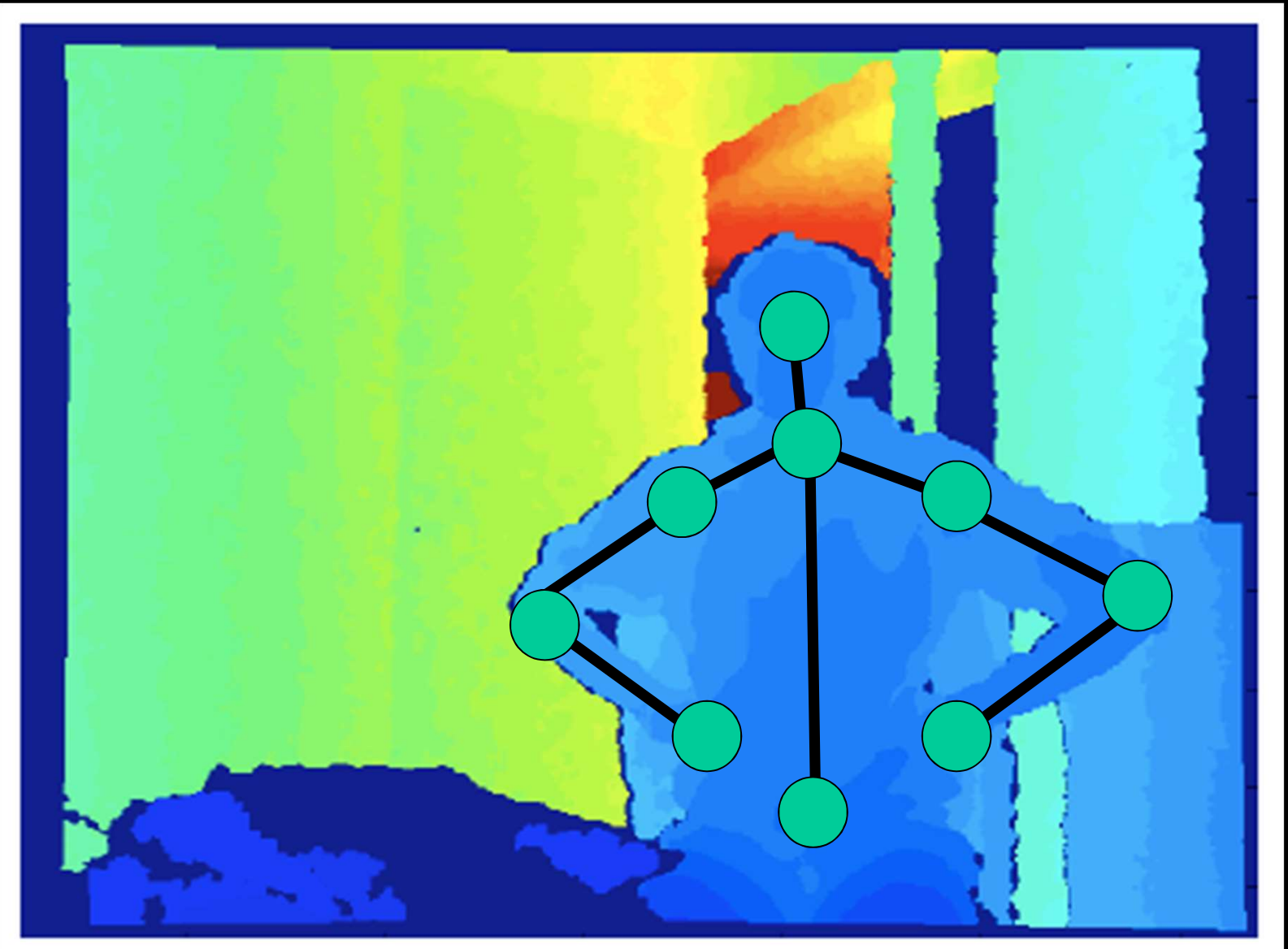
Programming assignments + final presentation

New sensors



Problem: find the 3D skeleton of people

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



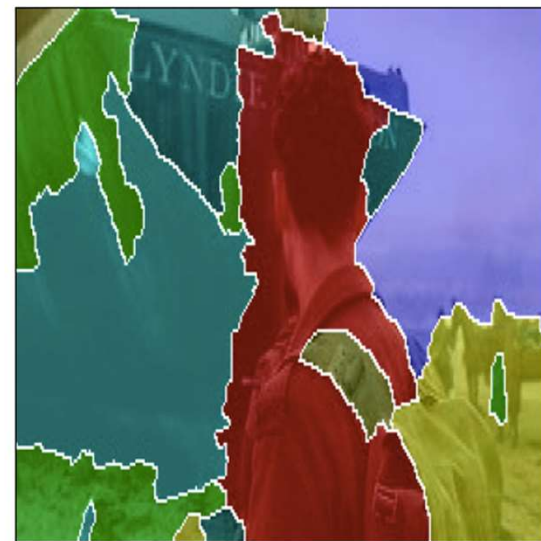
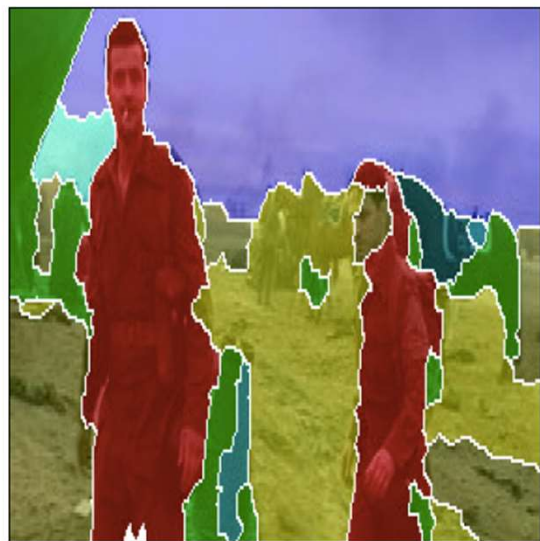
(Shotton et al., 2011)

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

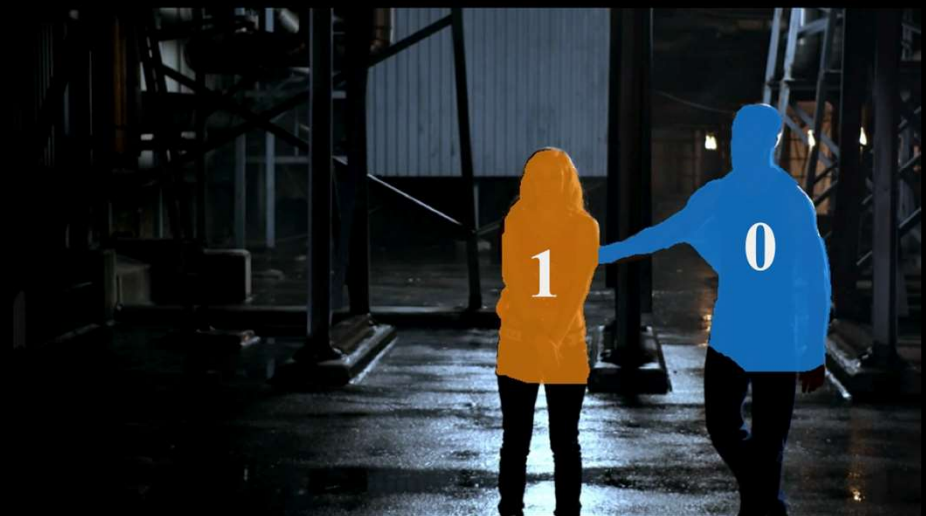
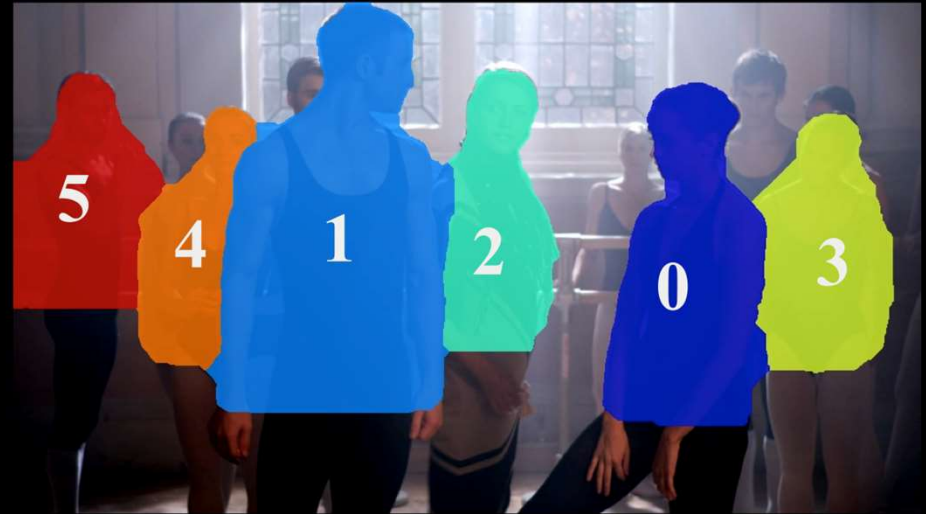
Programming assignments + final presentation

Segmentation



(Joulin, Bach, Ponce, CVPR'12)

Layered person segmentation [Seguin et al., 2015]

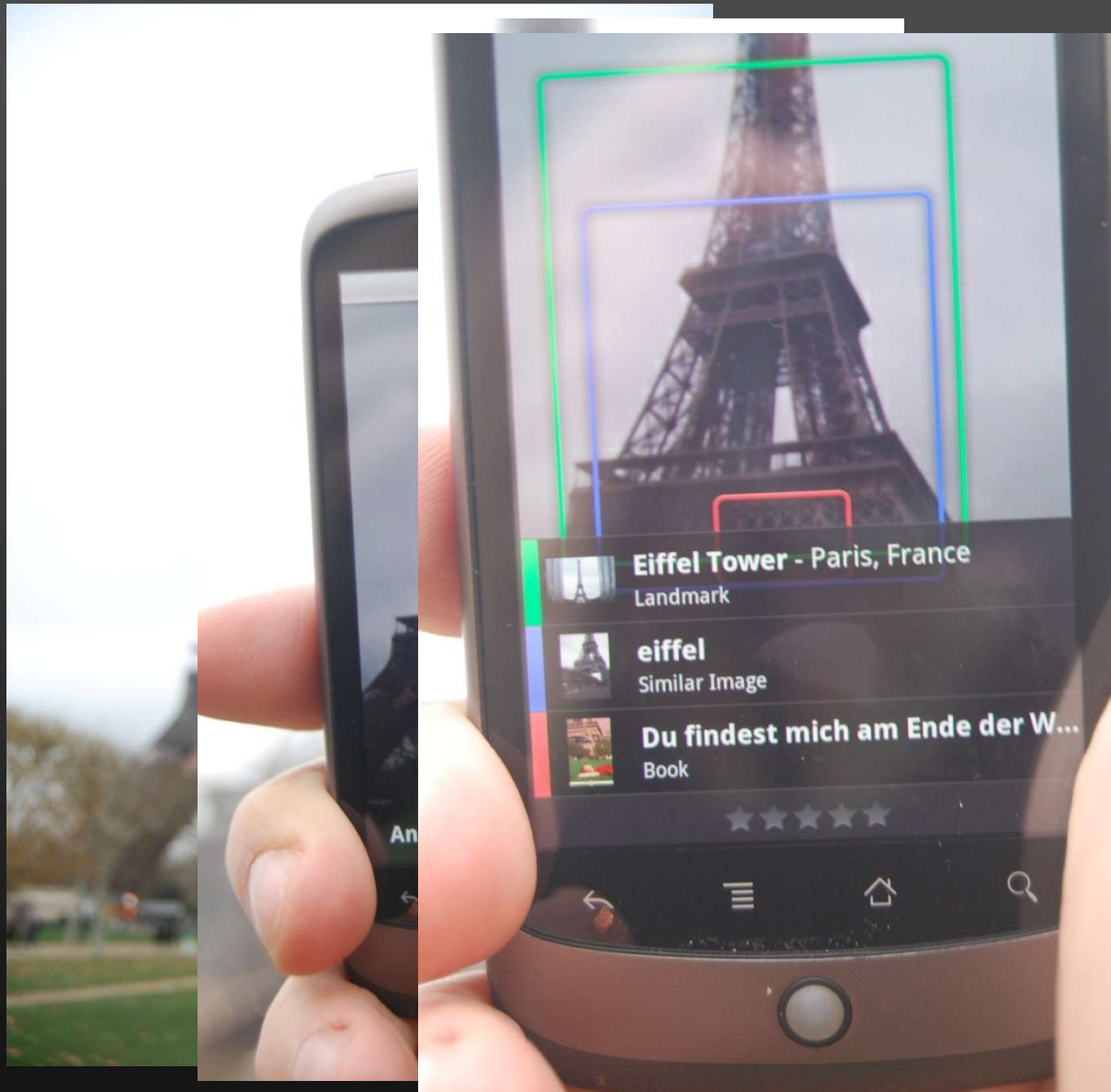


Course outline:

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

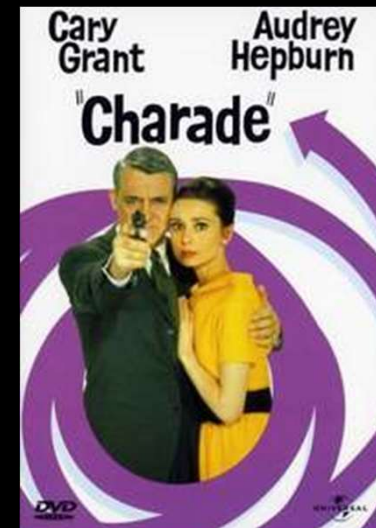


Example: Visual search in an entire feature length movie

Visually defined query



"Find this bag"



"Charade" [Donen, 1963]

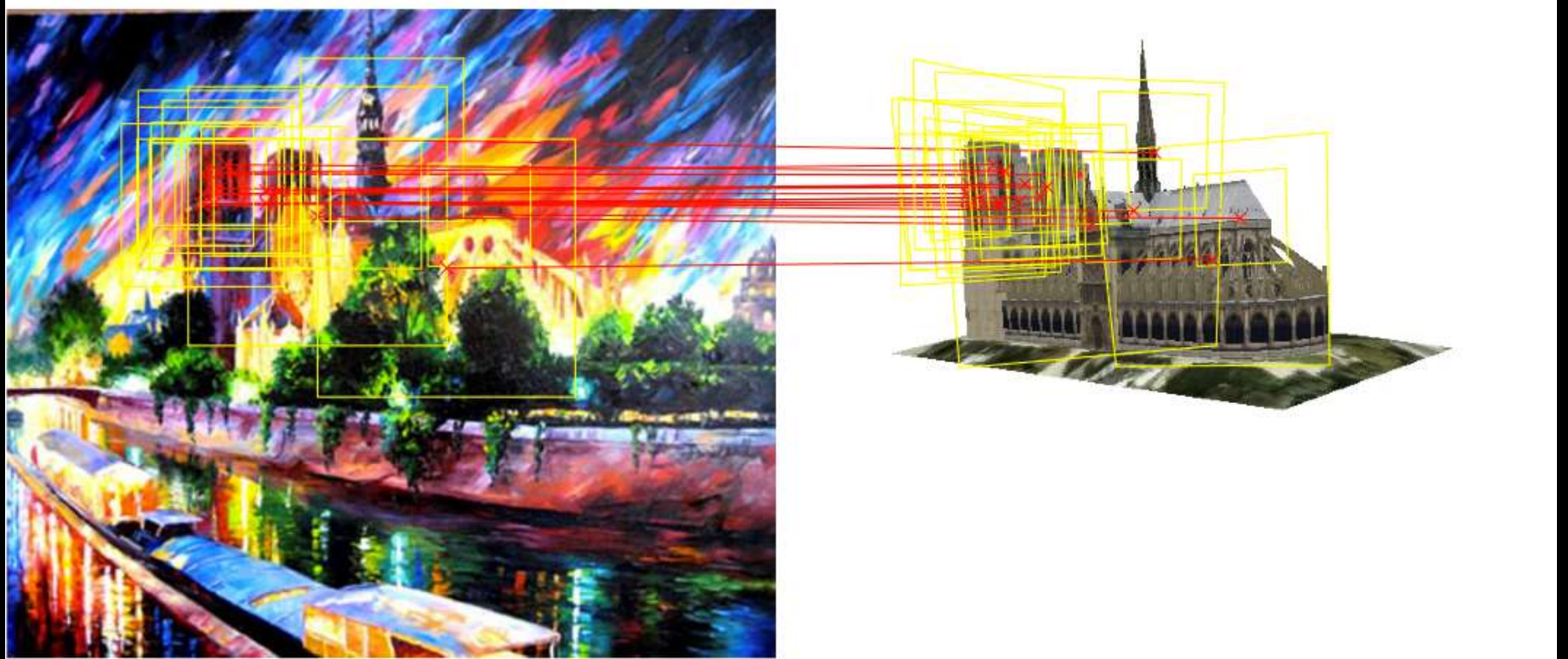
Demo:

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

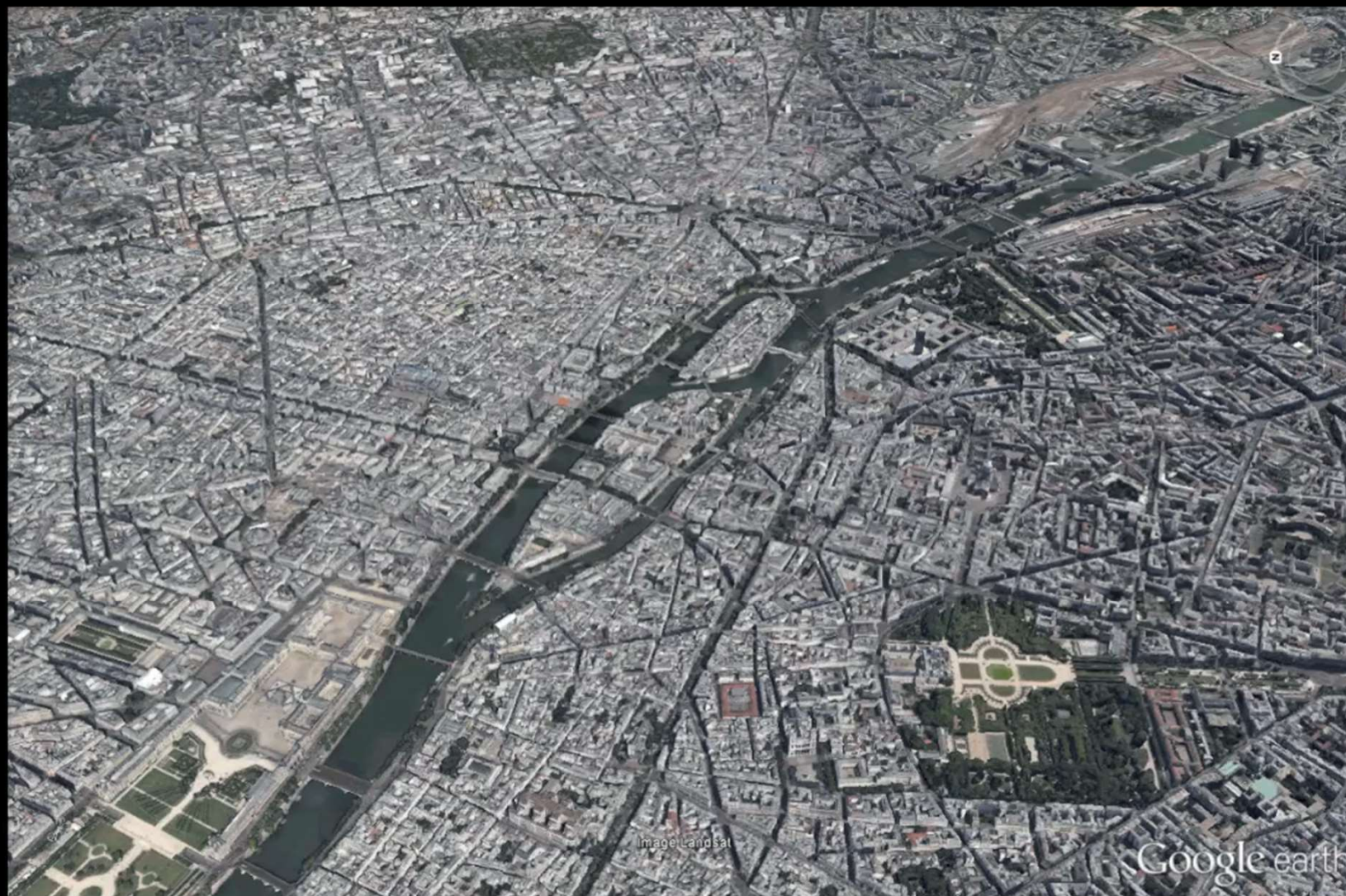
Instance level recognition: still difficult



Example: Matching non-photographic depictions



Geo-localization of historical and non-photographic depictions

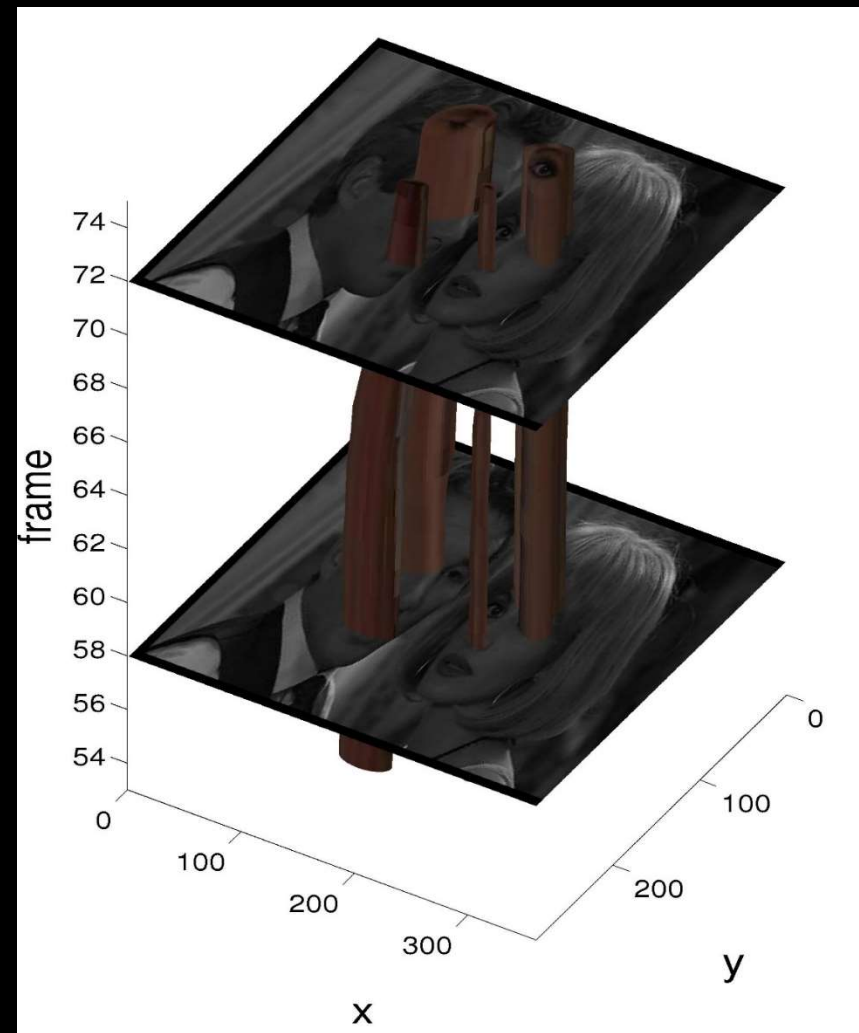


Recognizing people

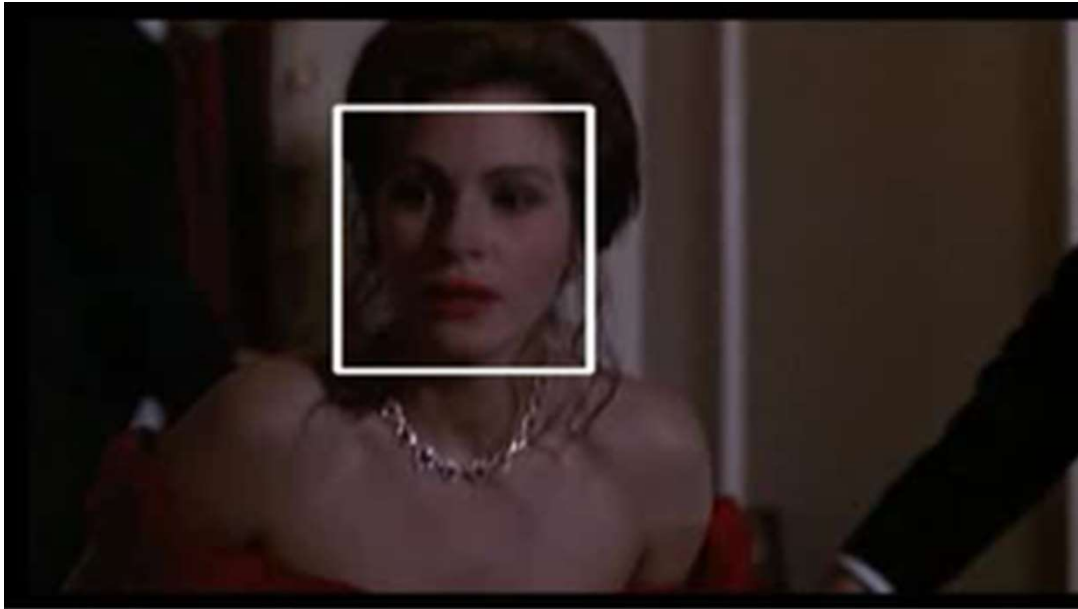


(Sivic, Everingham, Zisserman, 2005)

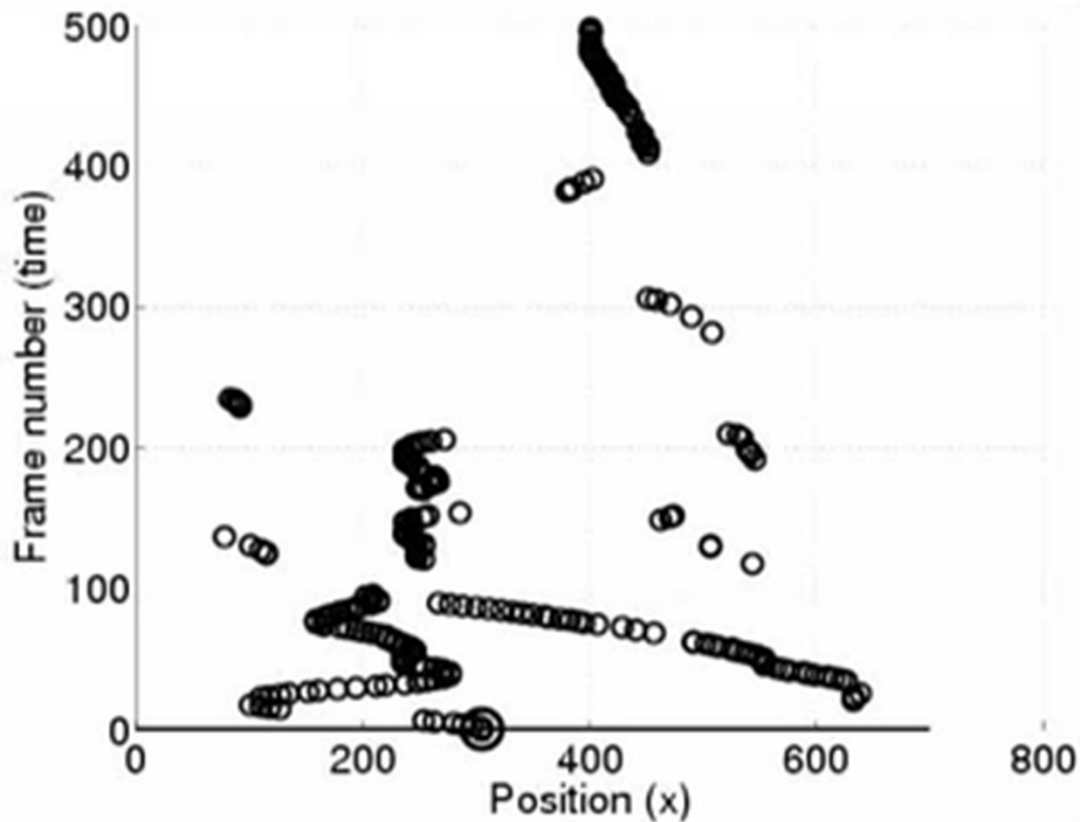
Faces: Region tubes for tracking faces



[Sivic, Everingham and Zisserman, 2005]

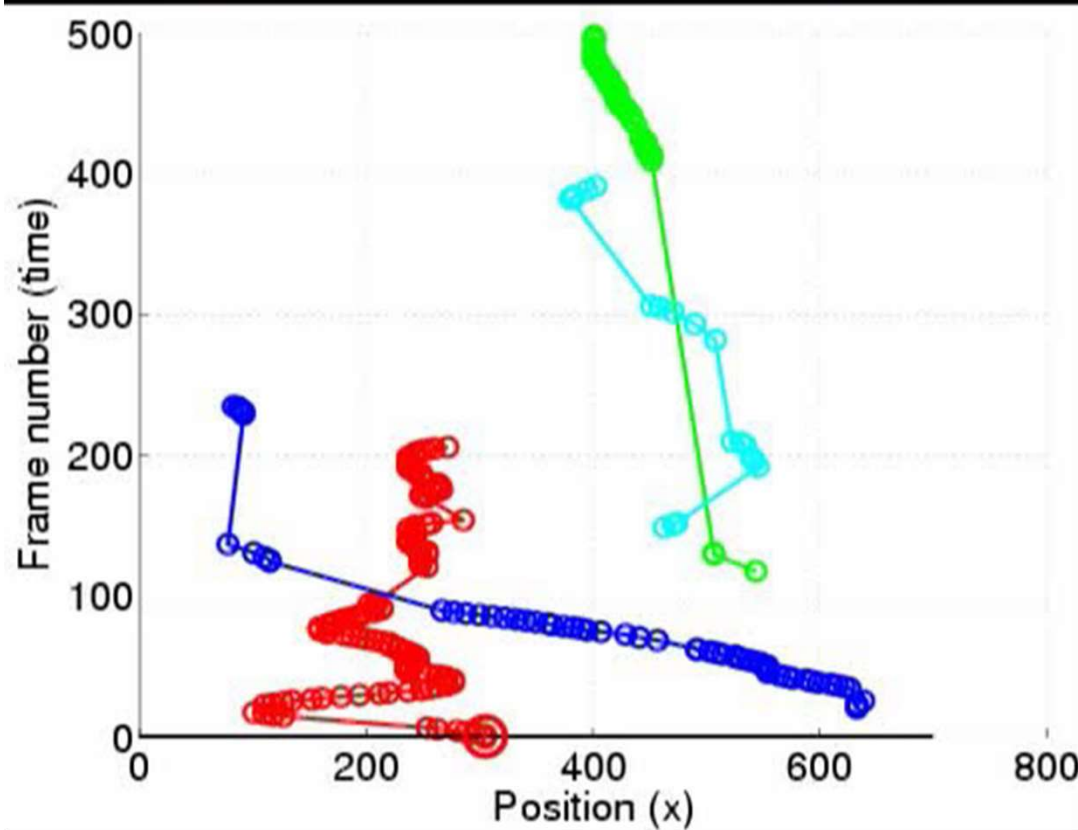


Raw face
detections



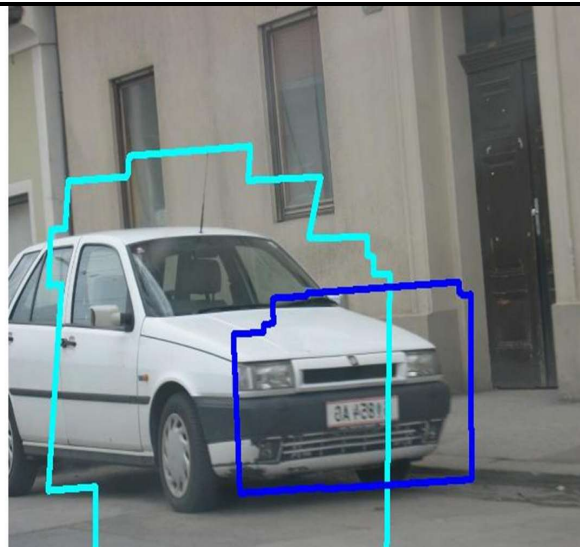
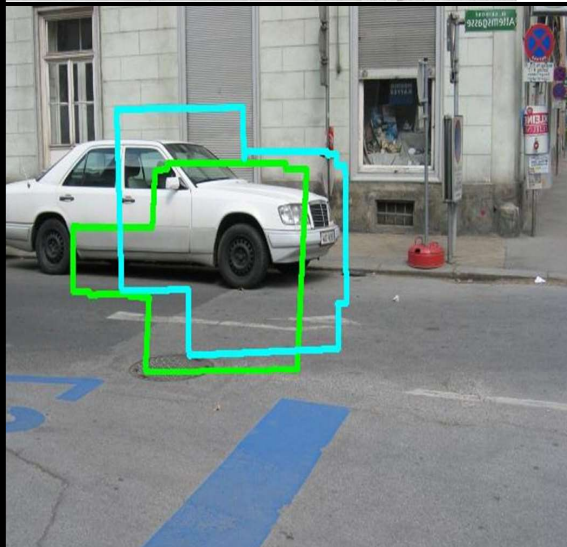
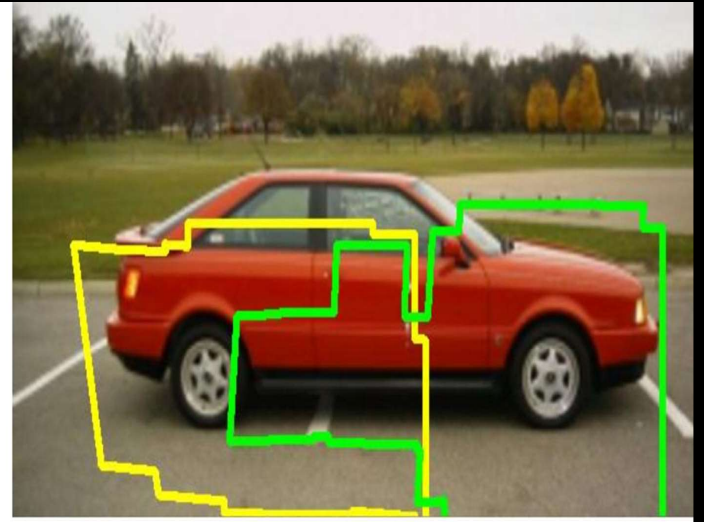
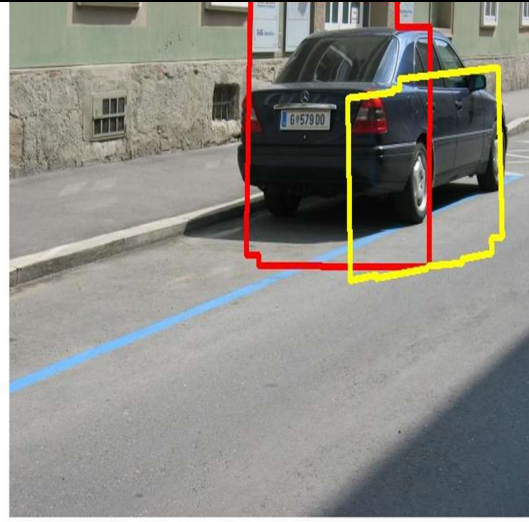


Tracking by
detection and
recognition



Connected face
tracks

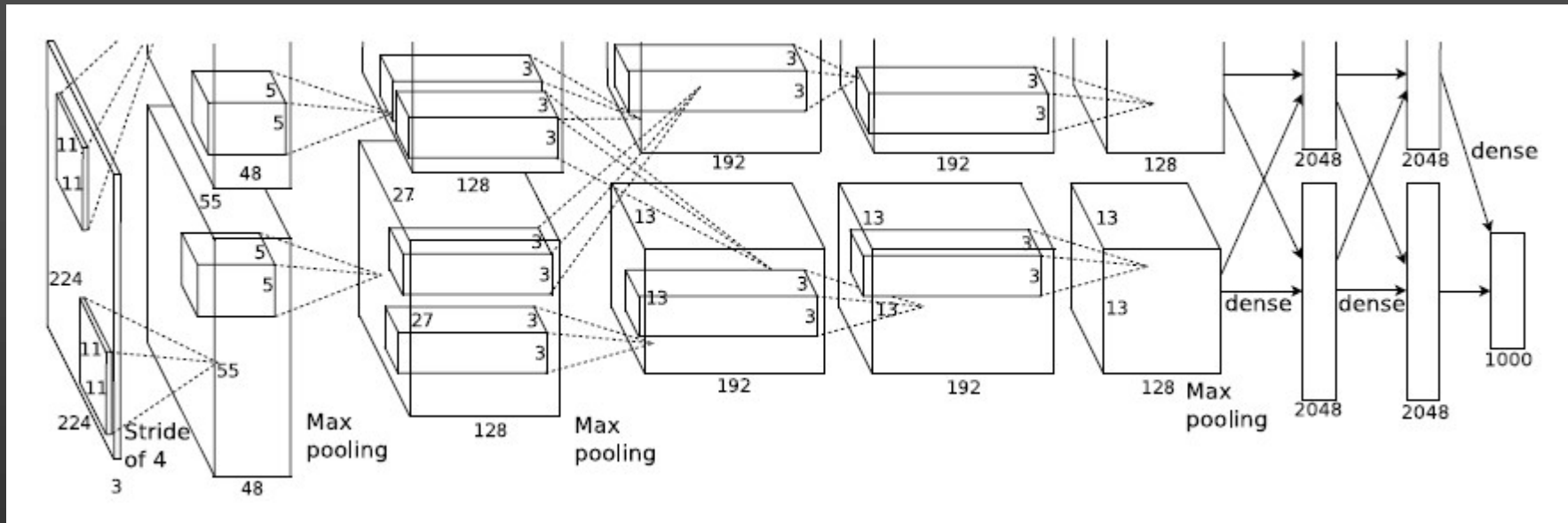
Recognition



(Kushal et al., 2007)

Convolutional neural networks

[Krizhevsky et al. NIPS'12]



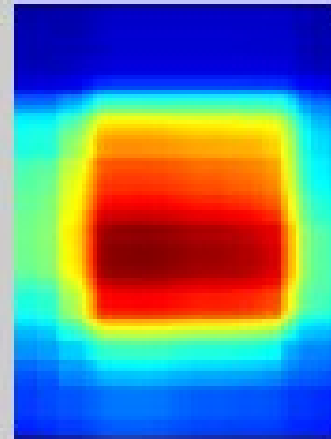
Convolutional Neural Networks:

- 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.**

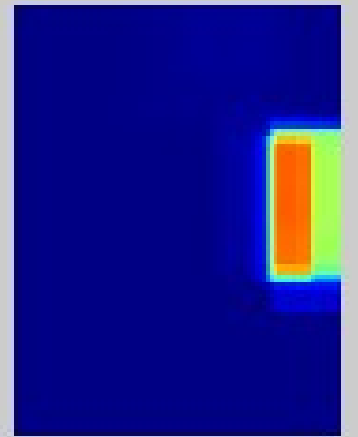
Some results



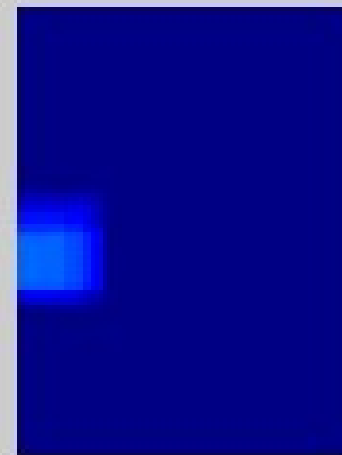
bus 203.2477



person 7.8236



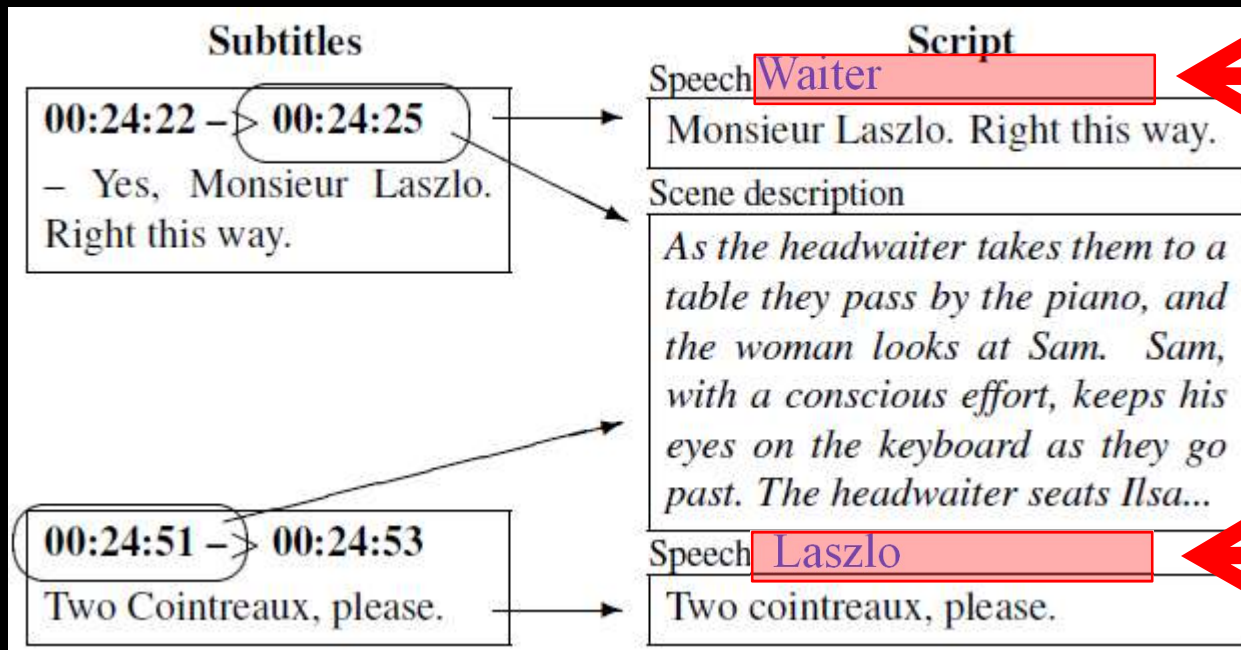
car 2.2312



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

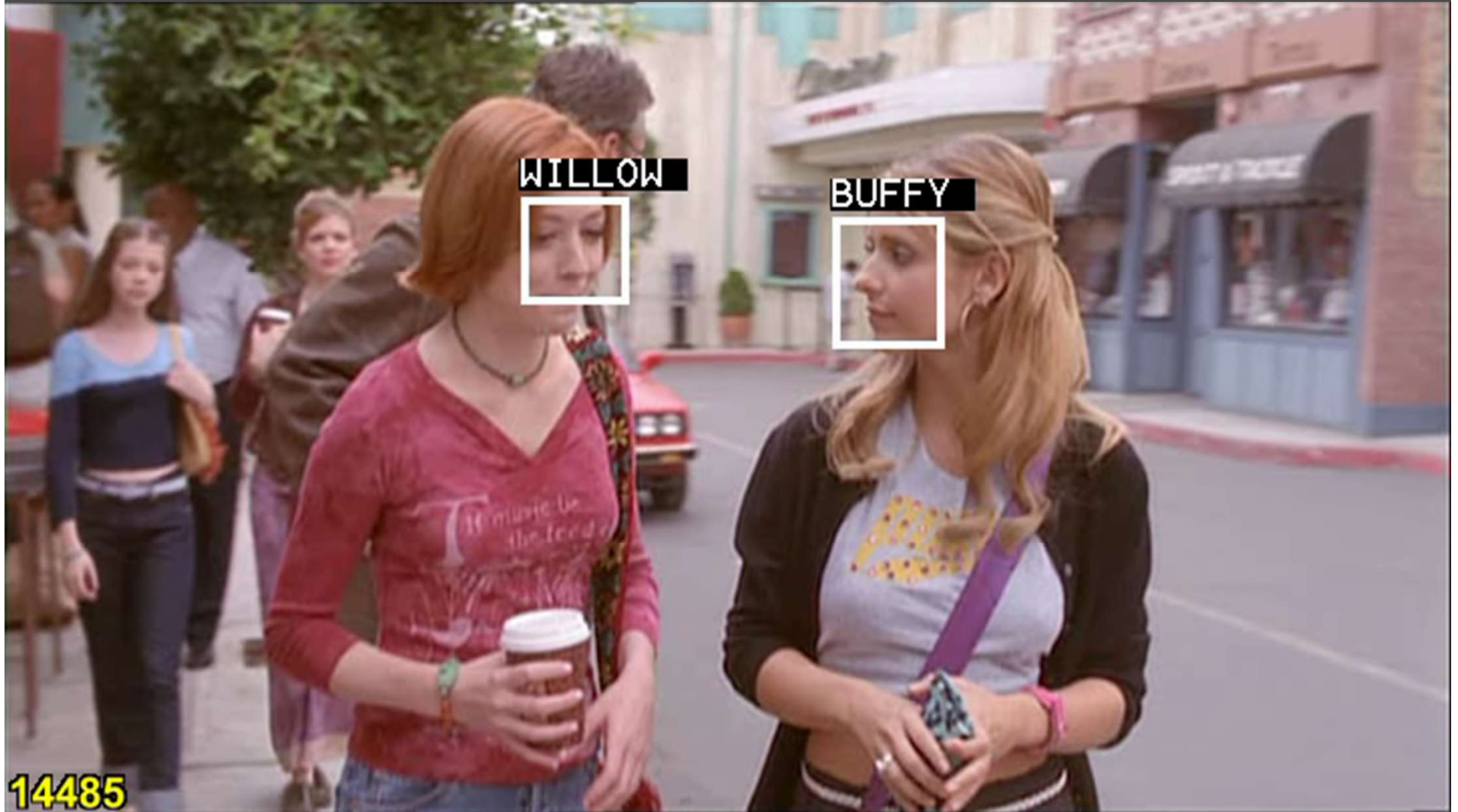
Automatic learning from video scripts

Input: Videos with aligned shooting scripts.



Output: Recognizer for each character in the video

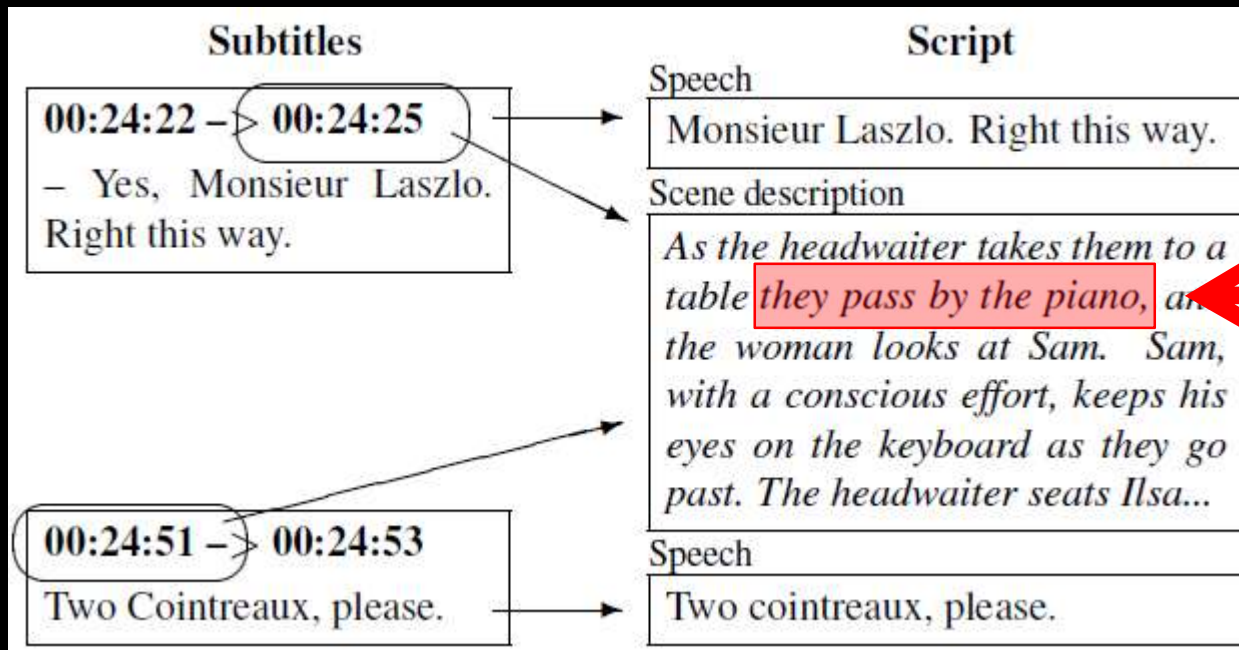
Recognizing people



(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

Clip number 0101

(Bojanowski et al., 2014)

Unsupervised object discovery

aeroplane-0004-029

- Object colocalization per class
- Unsupervised object discovery



(Suha et al., 2015)

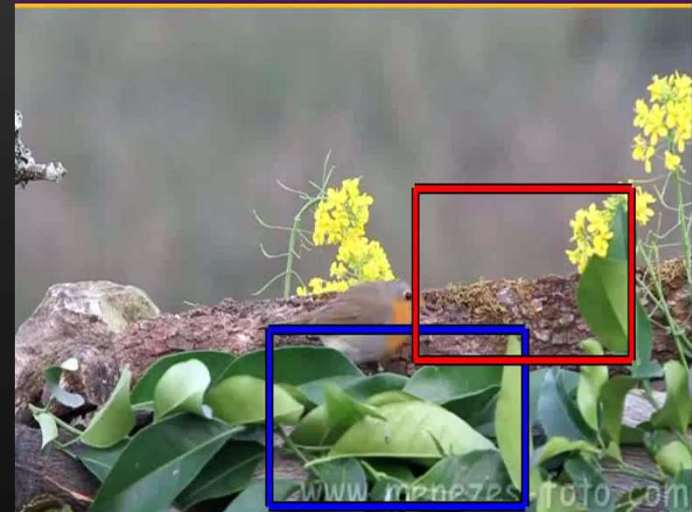
aeroplane-0013-140

- Object colocalization per class
- Unsupervised object discovery



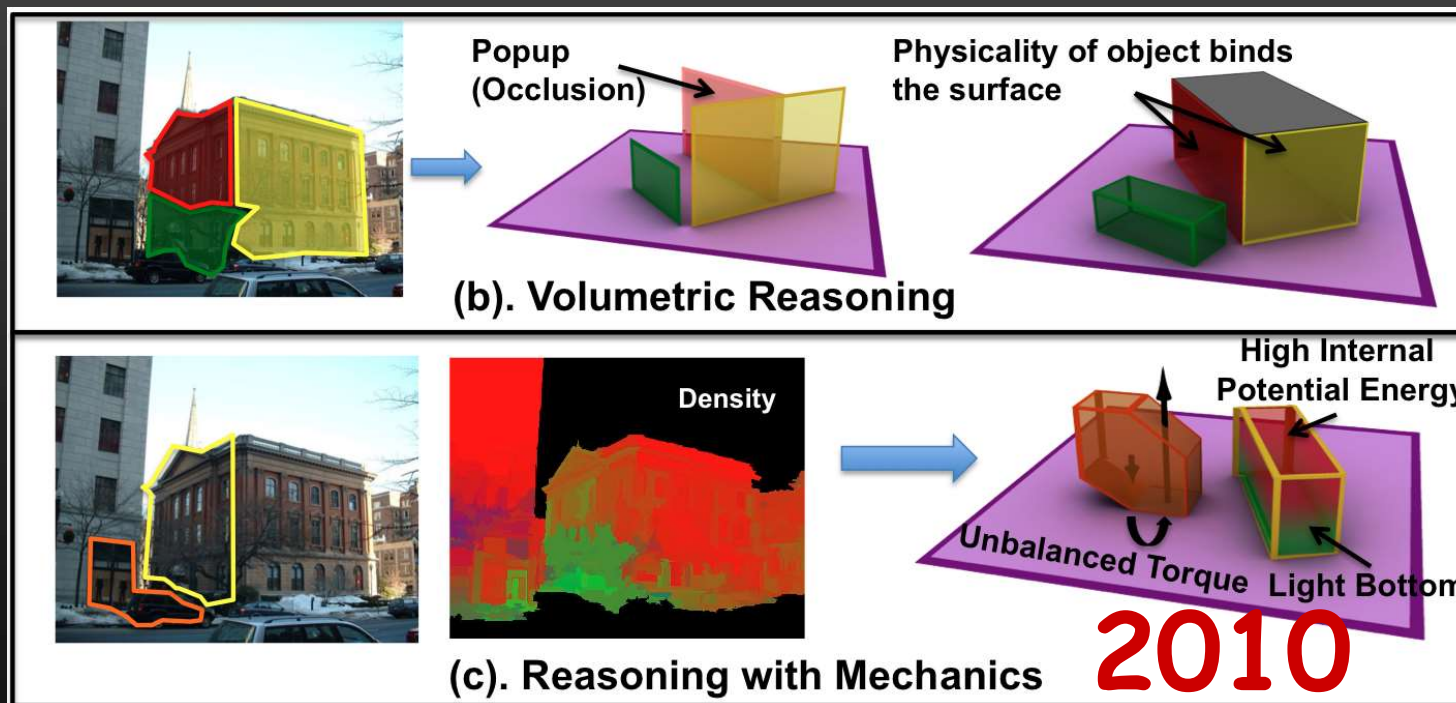
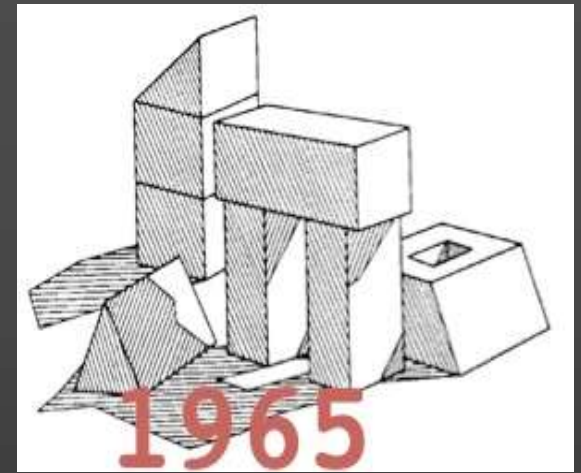
bird-0004-016

- Object colocalization per class
- Unsupervised object discovery



What about scene understanding?

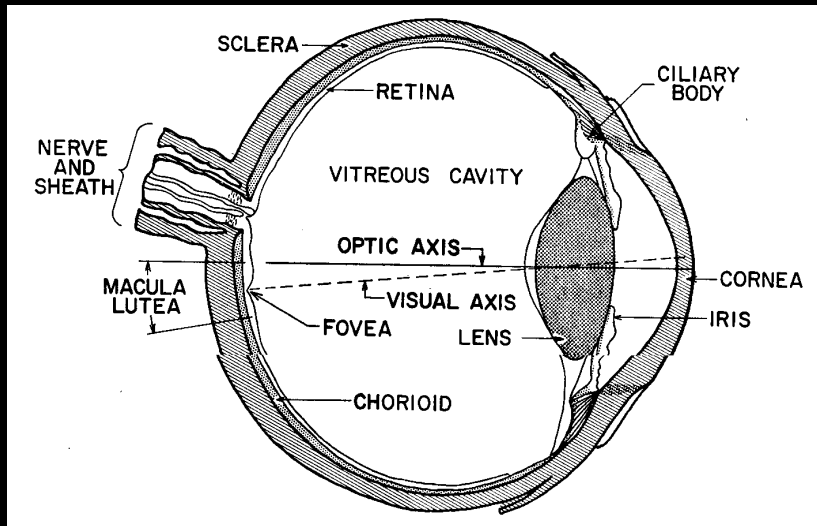
The blocks world revisited



(Gupta, Efros, Hebert, ECCV'10)

Camera geometry and calibration I

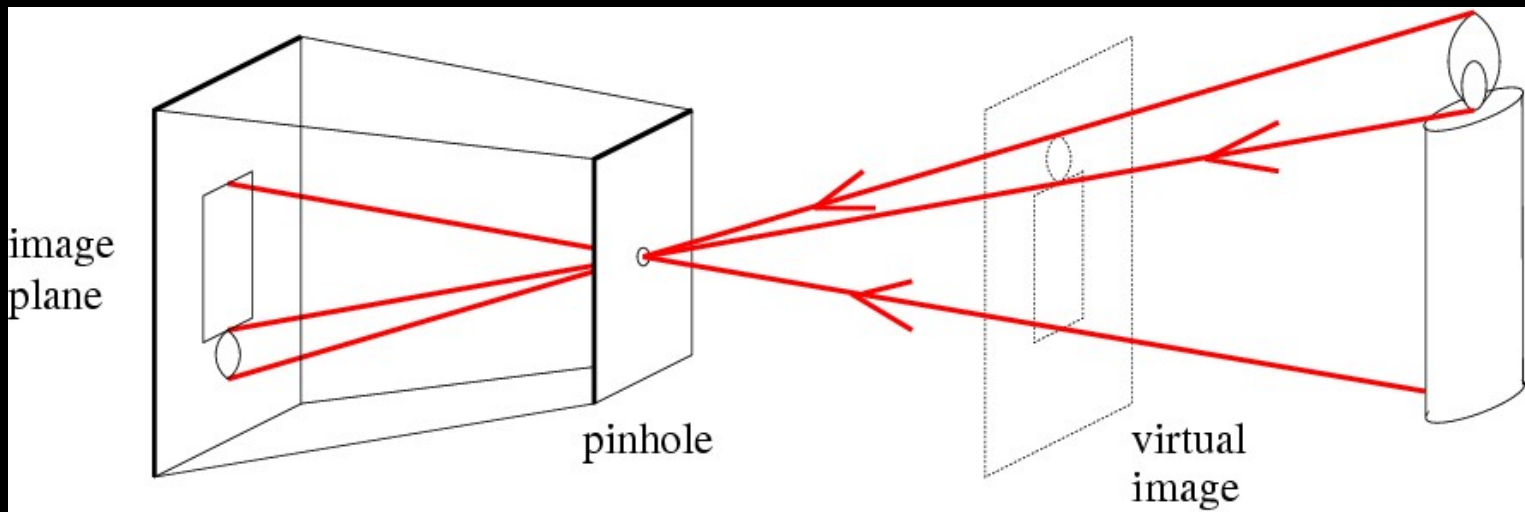
- Pinhole perspective projection
- Orthographic and weak-perspective models
- Non-standard models
- A detour through sensing country
- Intrinsic and extrinsic parameters



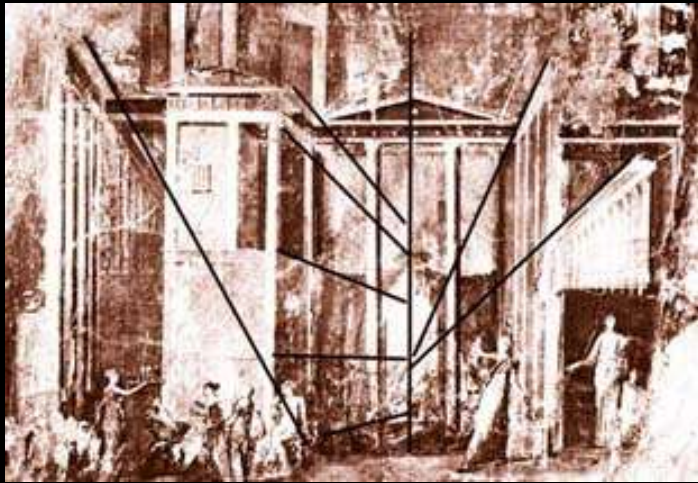
Animal eye: a looonng time ago.



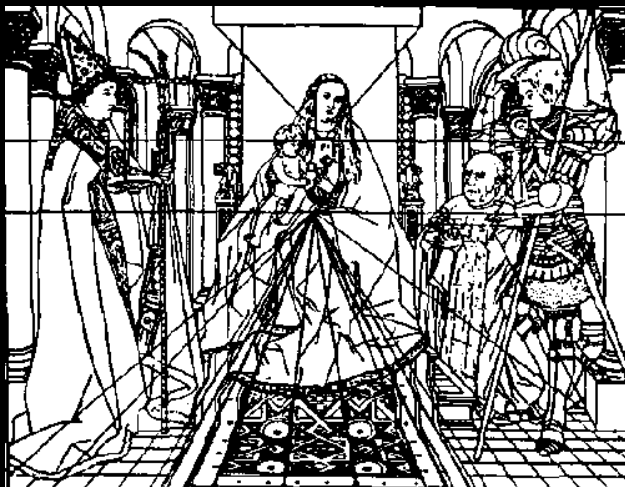
Photographic camera: Niepce, 1816.



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

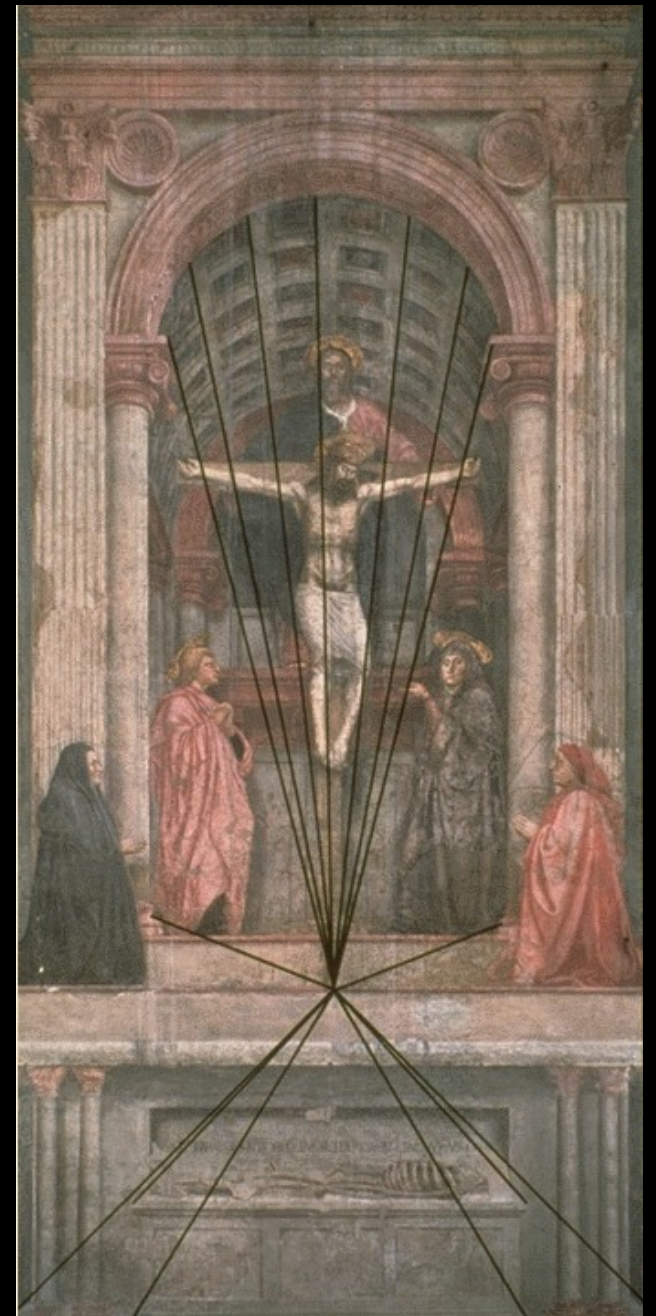


Pompei painting, 2000 years ago

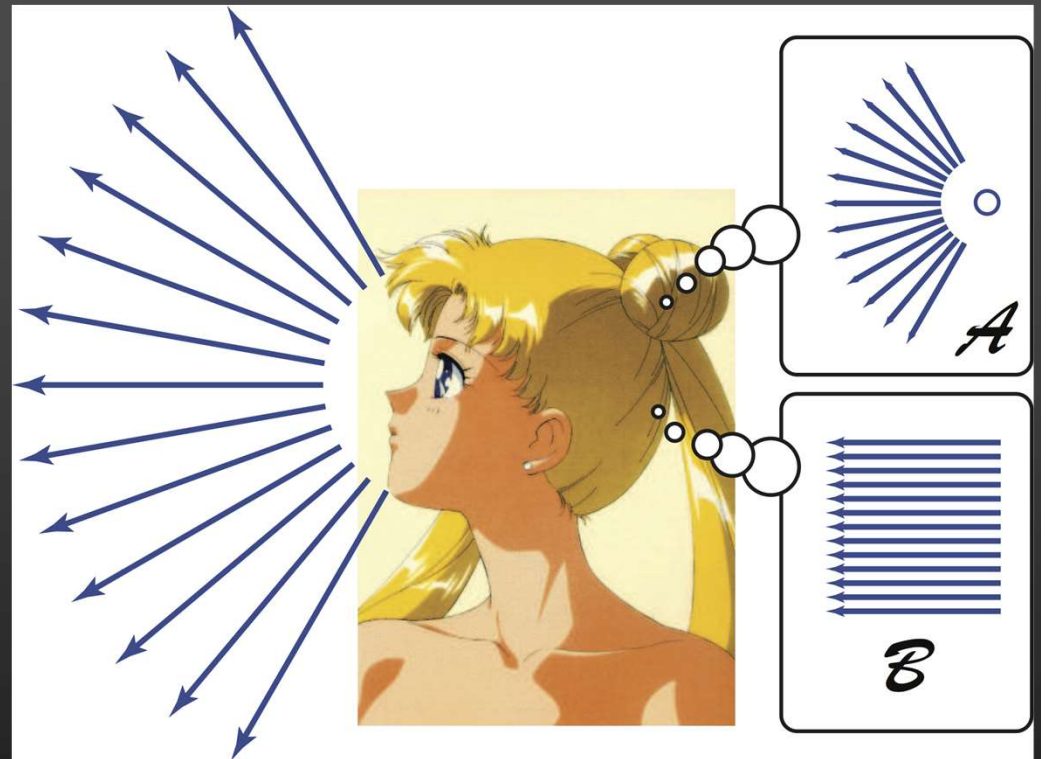
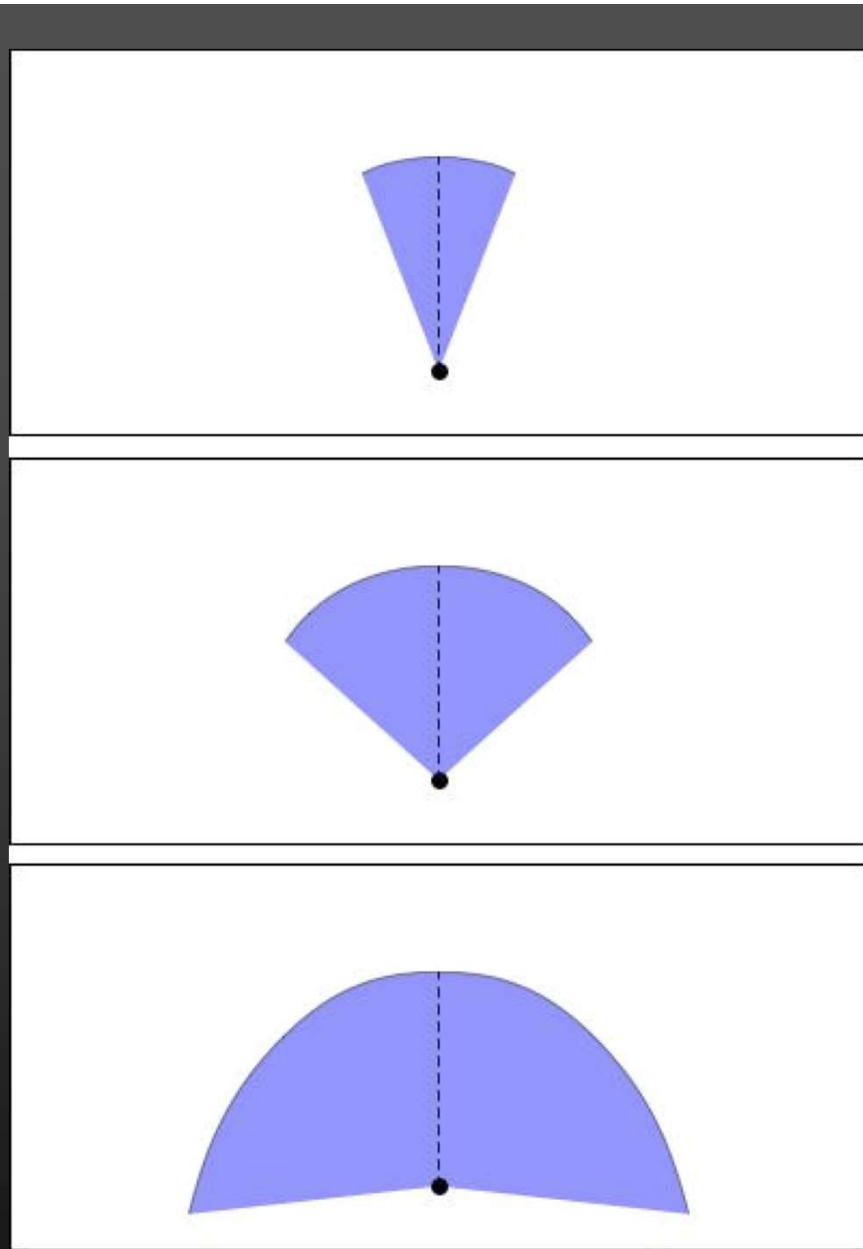


Van Eyk, XIVth Century

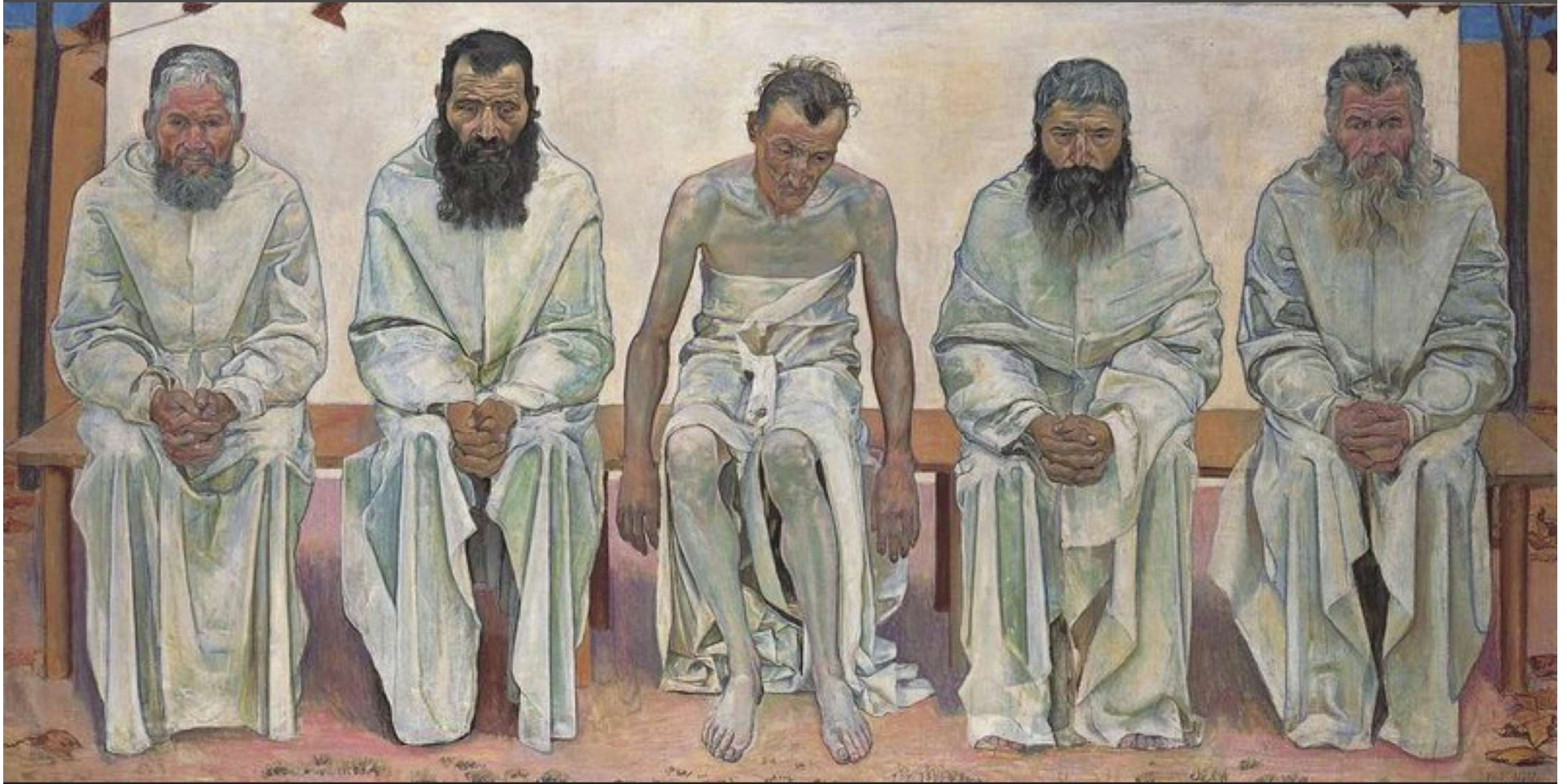
Brunelleschi, 1415



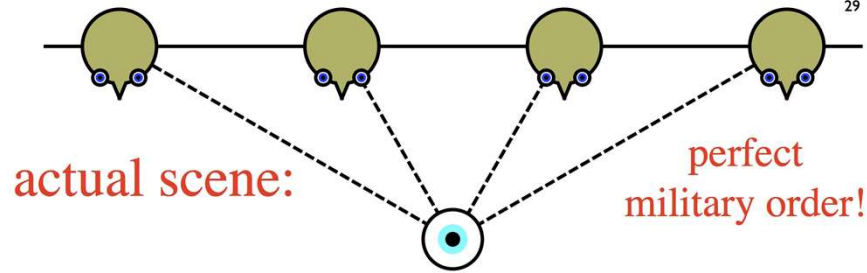
Massaccio's Trinity, 1425



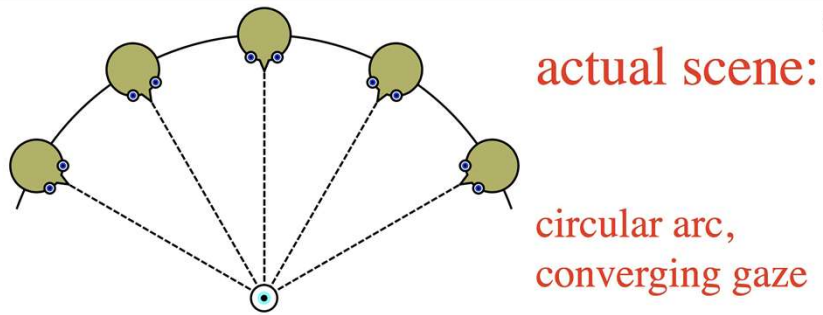
Most people don't experience the divergence of visual rays in a veridical manner. This is fine. [Koenderink]



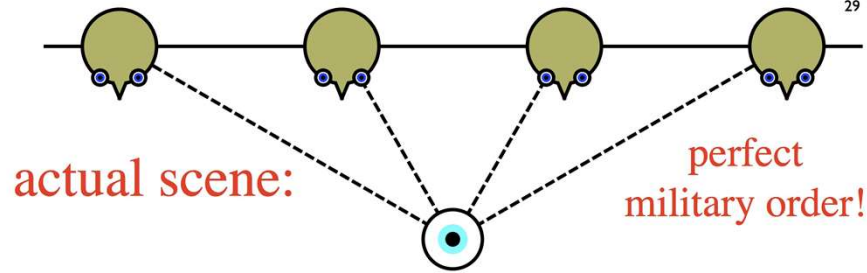
Ferdinand Hodler



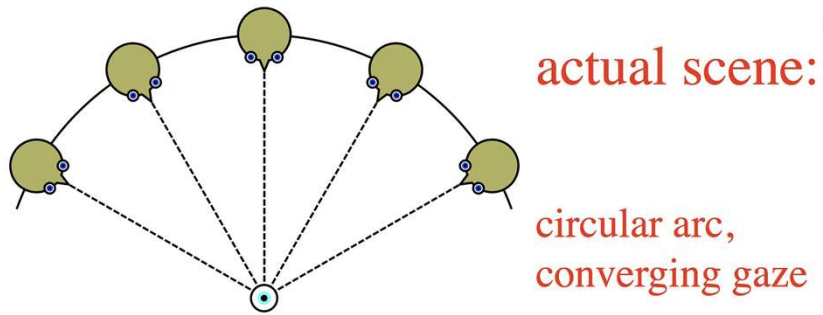
impression: gazes diverge, persons arranged on a curve



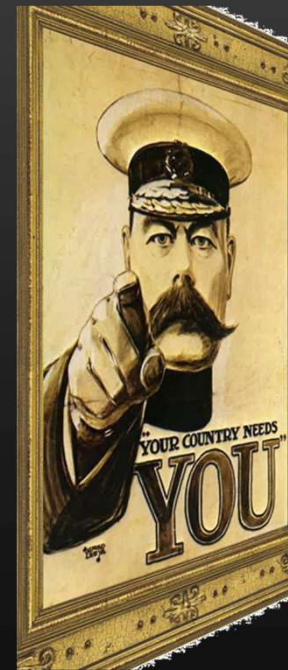
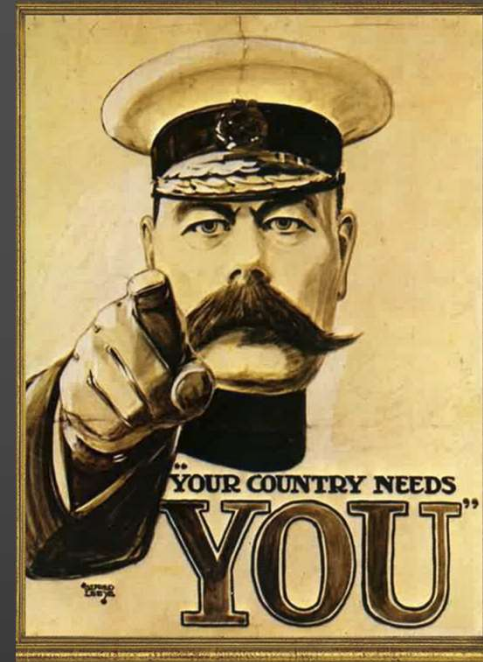
impression: perfect military lineup!

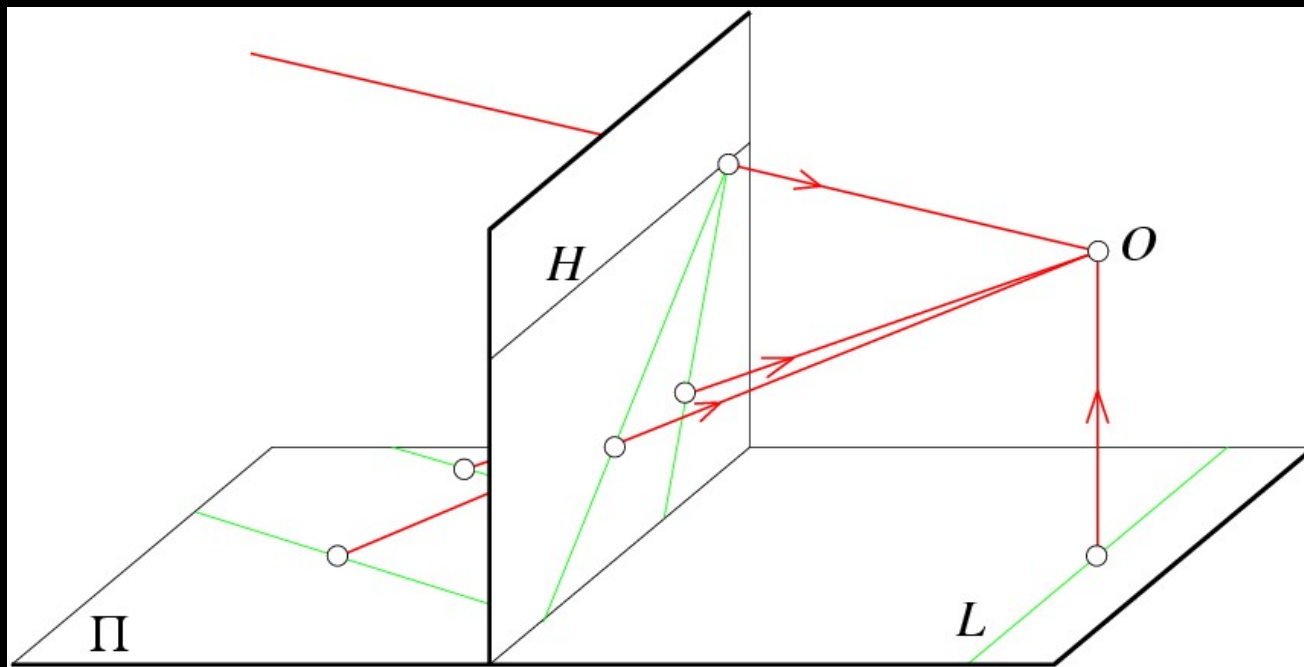
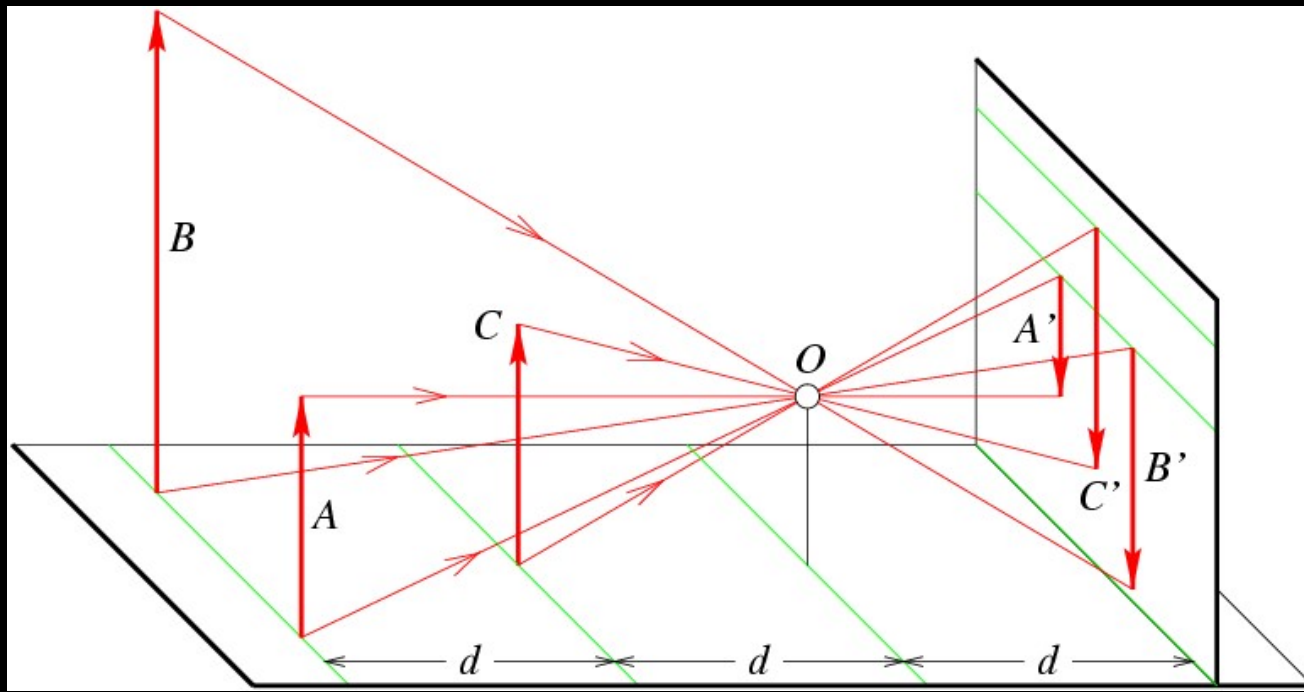


impression: gazes diverge, persons arranged on a curve

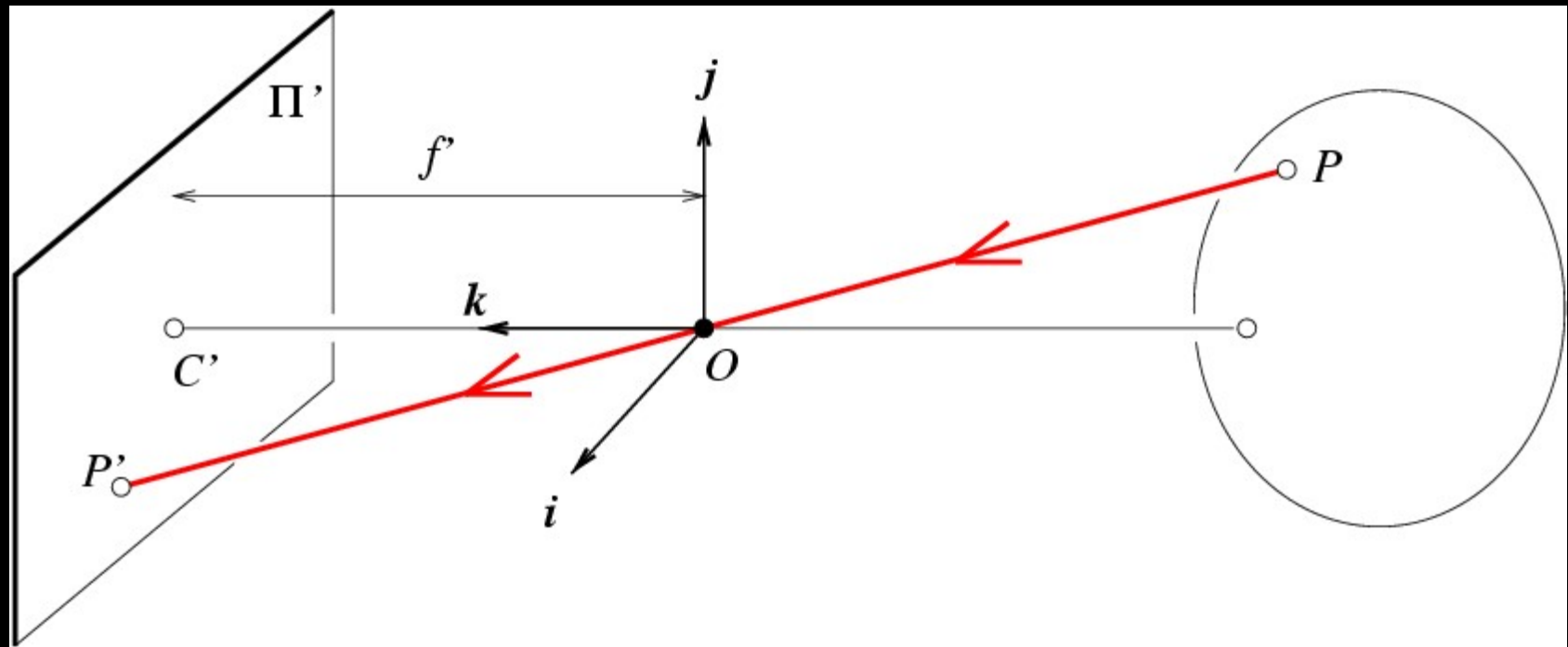


impression: perfect military lineup!





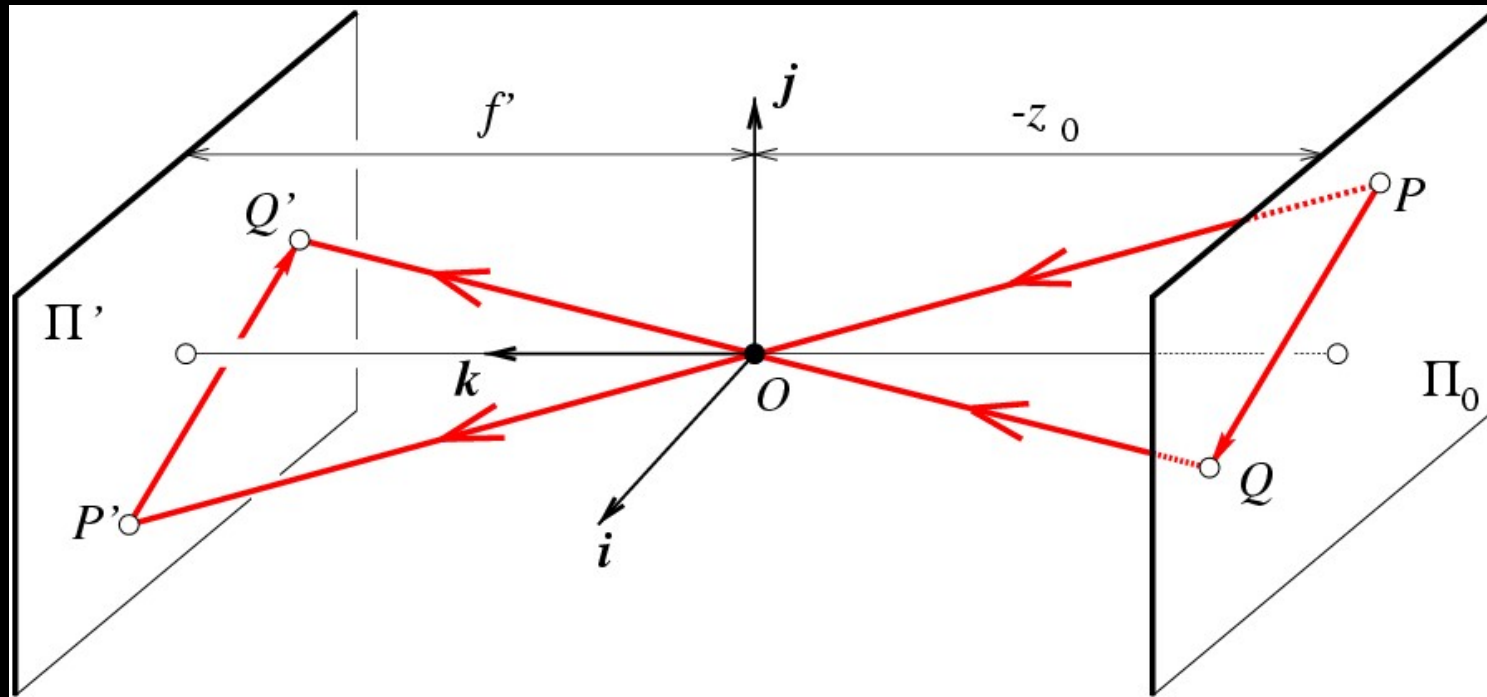
Pinhole Perspective Equation



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

NOTE: z is always negative..

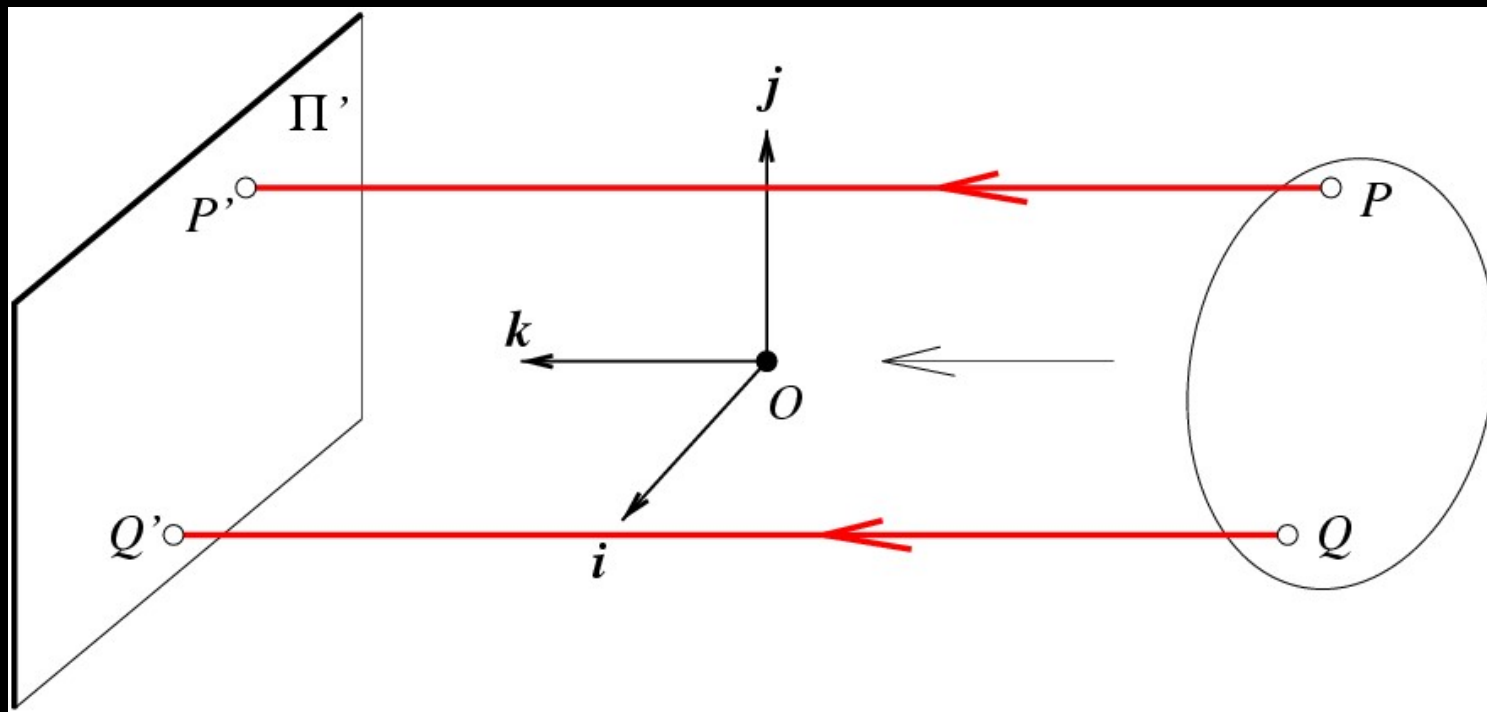
Affine projection models: Weak perspective projection



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

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



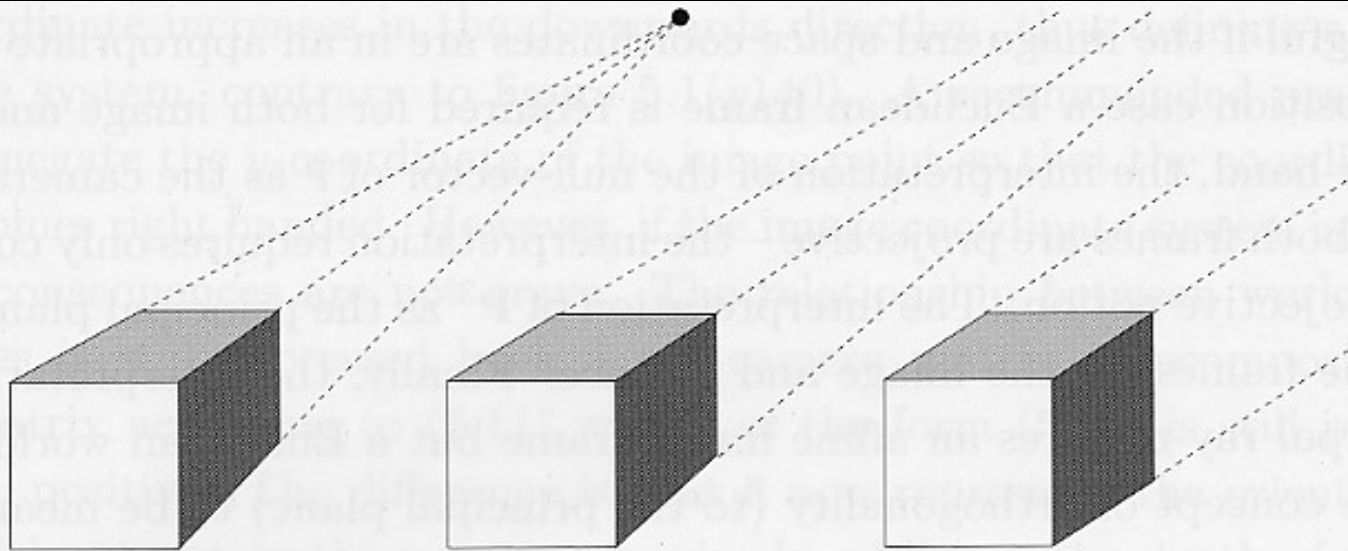
$$\begin{cases} x' = x \\ y' = y \end{cases}$$

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

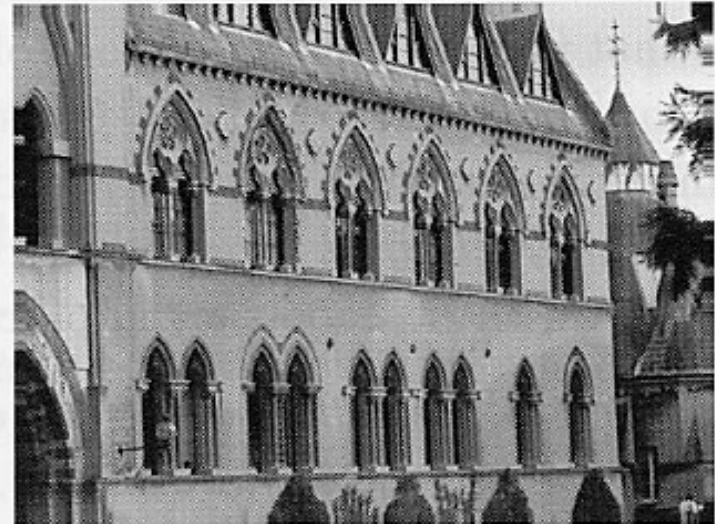
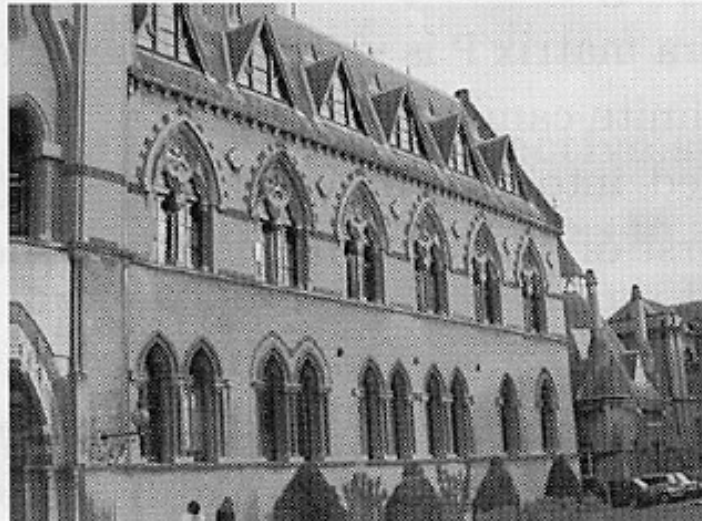


perspective

weak perspective

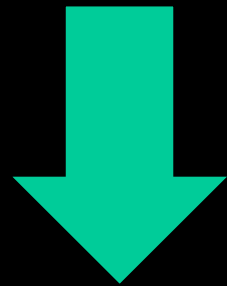
————— increasing focal length —————>

————— increasing distance from camera —————>



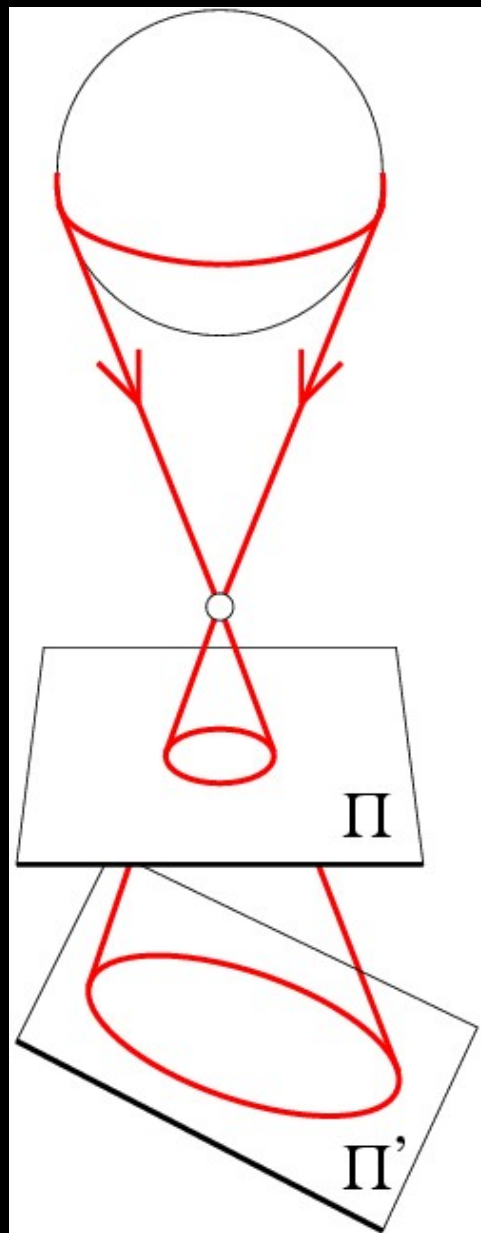
From Zisserman & Hartley

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

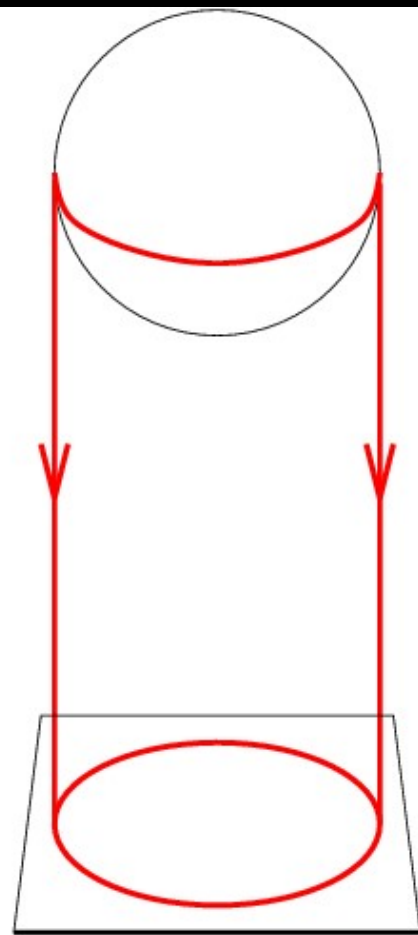


Weak perspective:
Angles are better preserved
The projections of parallel
lines are (almost) parallel

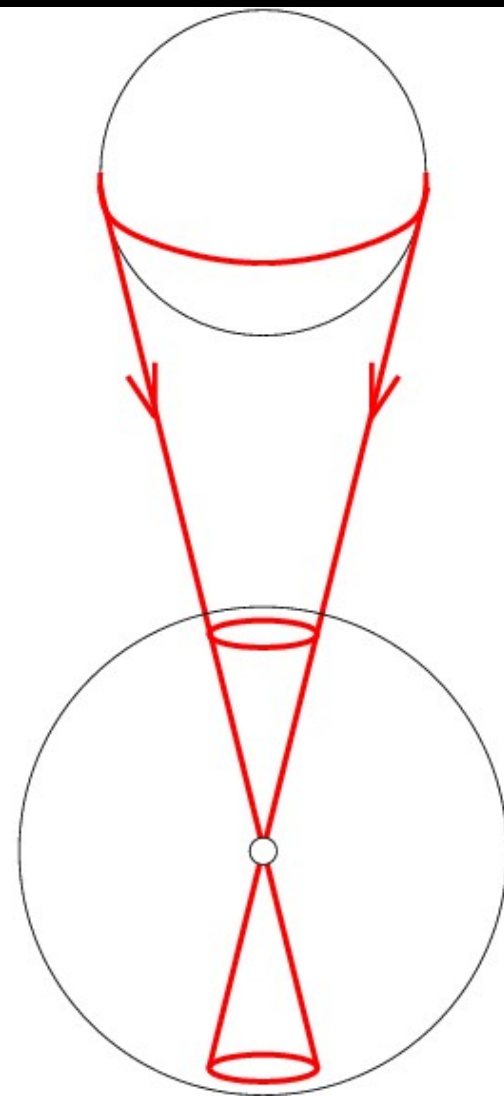




Planar pinhole
perspective



Orthographic
projection



Spherical pinhole
perspective