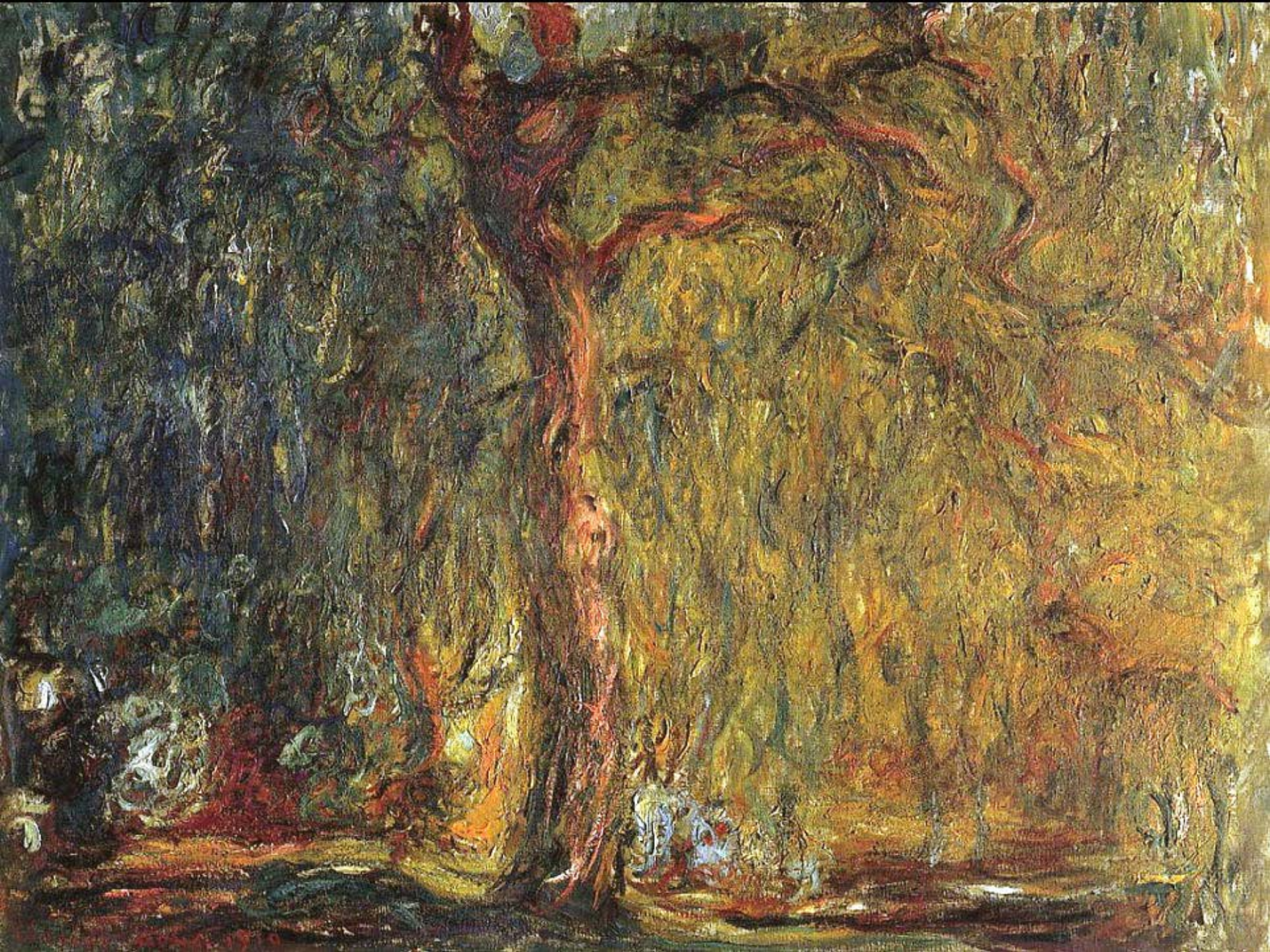


Reconnaissance d'objets et vision artificielle

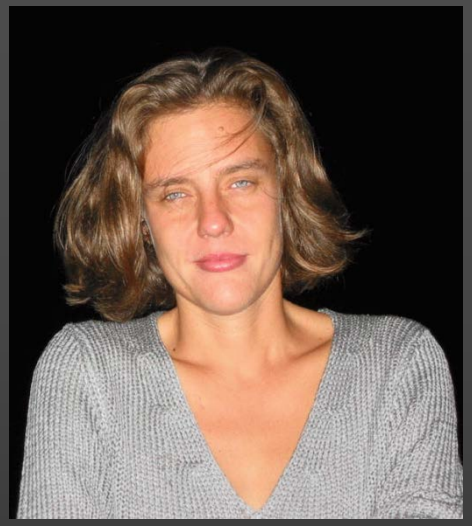
http://www.di.ens.fr/willow/teaching/recvis12/

Jean Ponce (ponce@di.ens.fr)
http://www.di.ens.fr/~ponce
Equipe-projet WILLOW
ENS/INRIA/CNRS UMR 8548
Département d'Informatique
Ecole Normale Supérieure, Paris

Jean Ponce

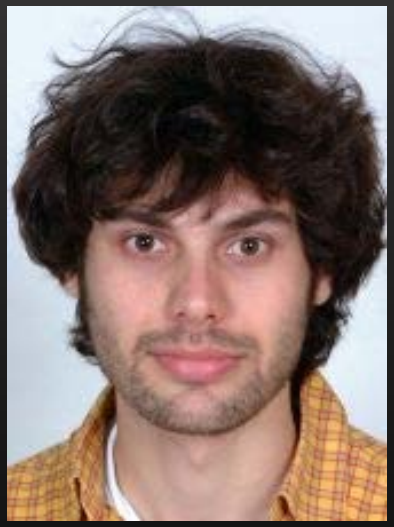
http://www.di.ens.fr/~ponce/

Cordelia Schmid



http://lear.inrialpes.fr/~schmid/

Josef Sivic



http://www.di.ens.fr/~josef/

Ivan Laptev

http://www.irisa.fr/vista/Equipe/People/Ivan.Laptev.html

Nous cherchons
toujours
des stagiaires
à la fin du cours

Initiation à la vision artificielle

Jean Ponce (ponce@di.ens.fr)

Jeudis, salle R, 9-12h

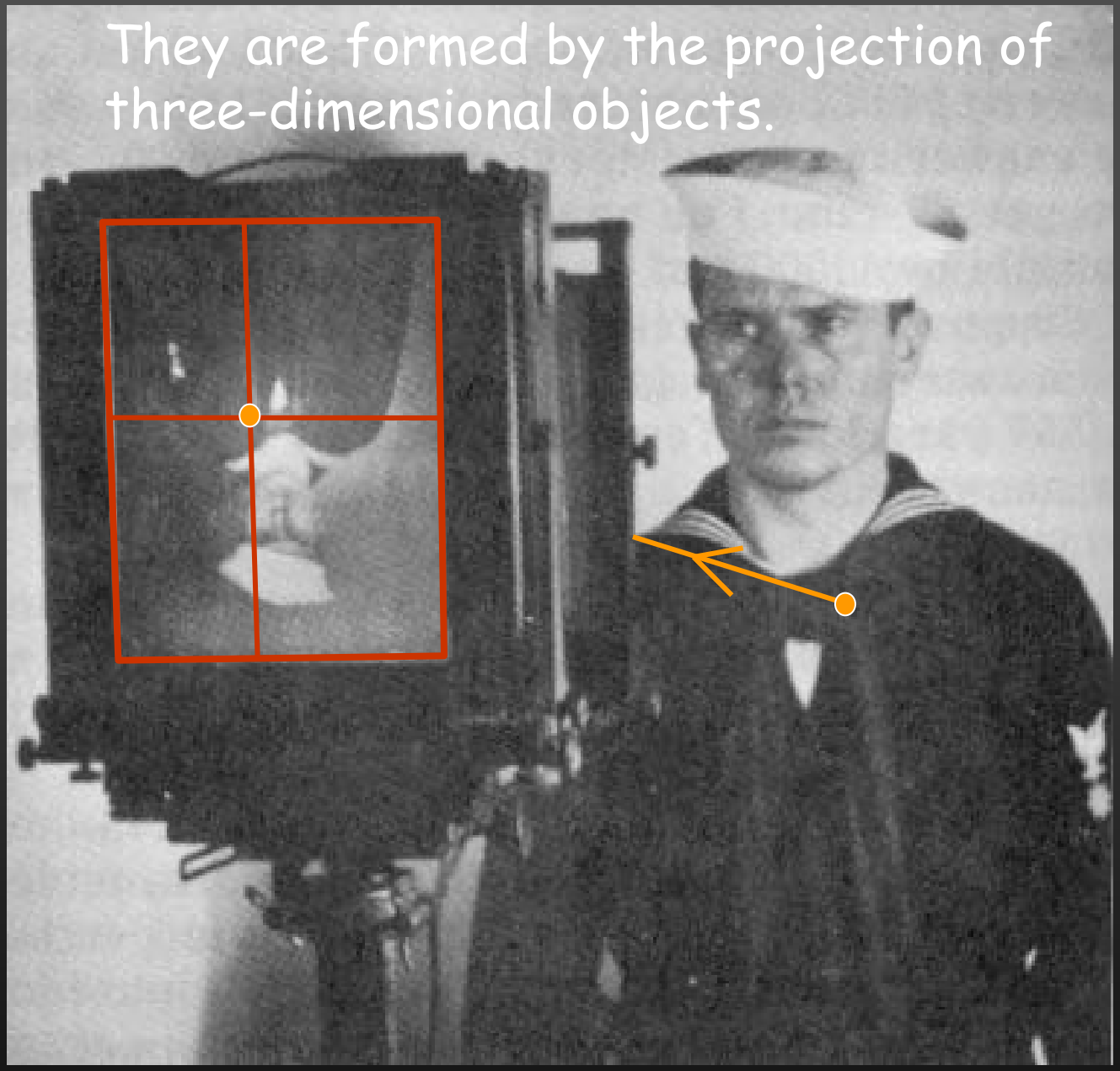
Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry

Why?

NAO (Aldebaran Robotics)

(Mairal, Bach, Ponce, PAMI'12)



Images are brightness/color patterns drawn in a plane.

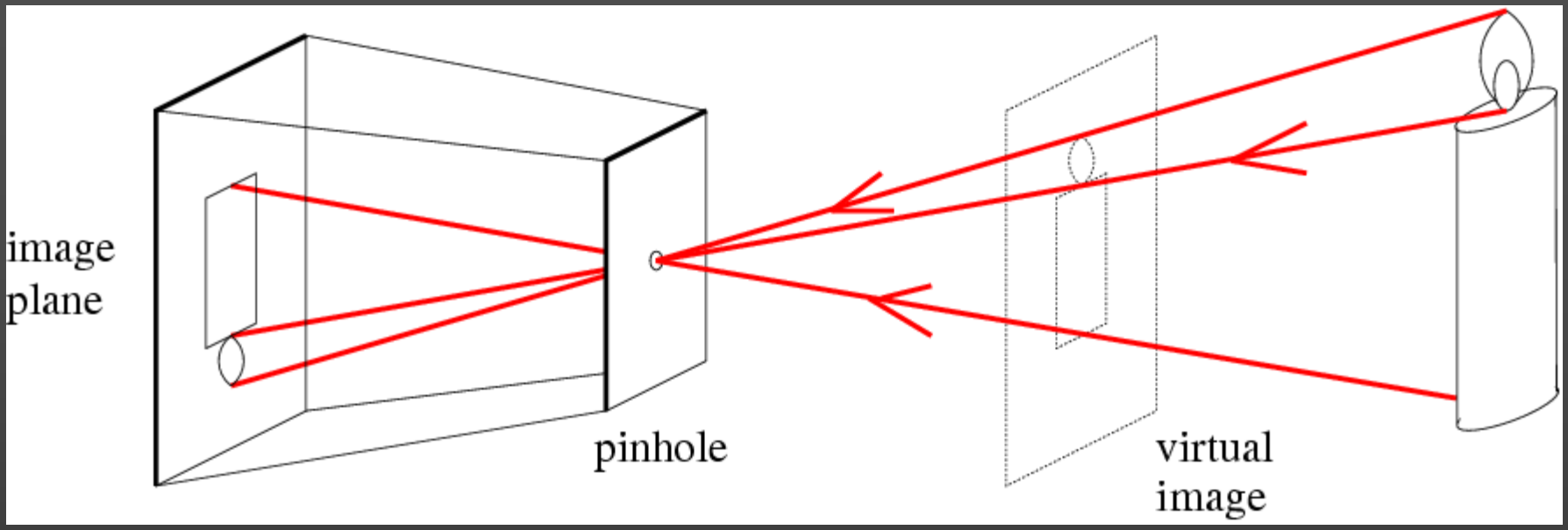
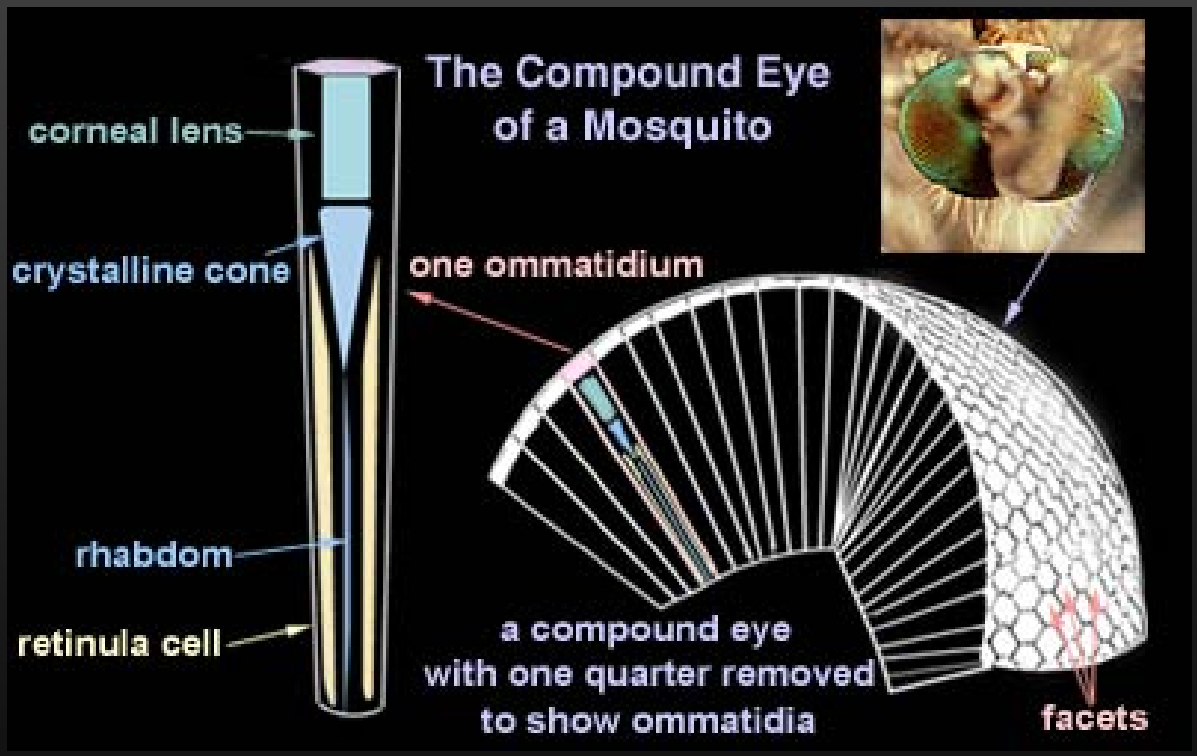
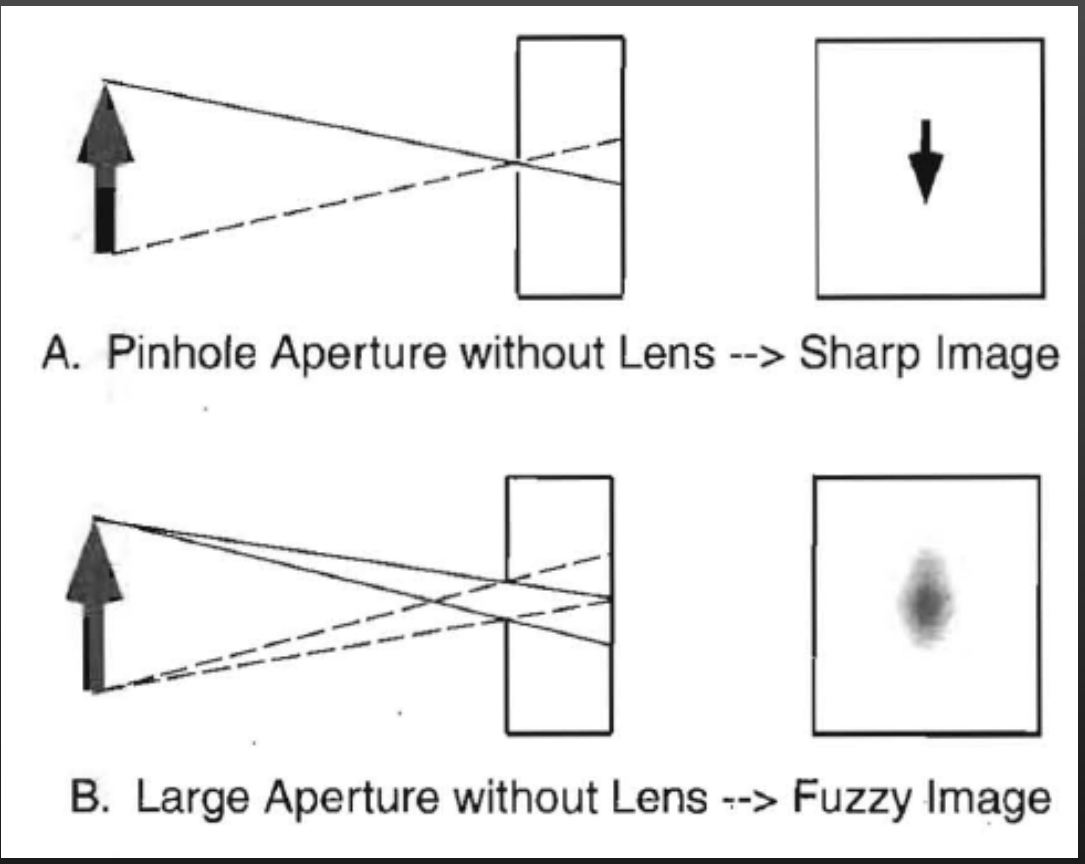


image
plane
pinhole
virtual
image


The Compound Eye
of a Mosquito
corneal lens
crystalline cone
rhabdom
retinula cell
one ommatidium
a compound eye
with one quarter removed
to show ommatidia
facets

Pinhole camera: trade-off between sharpness and light transmission

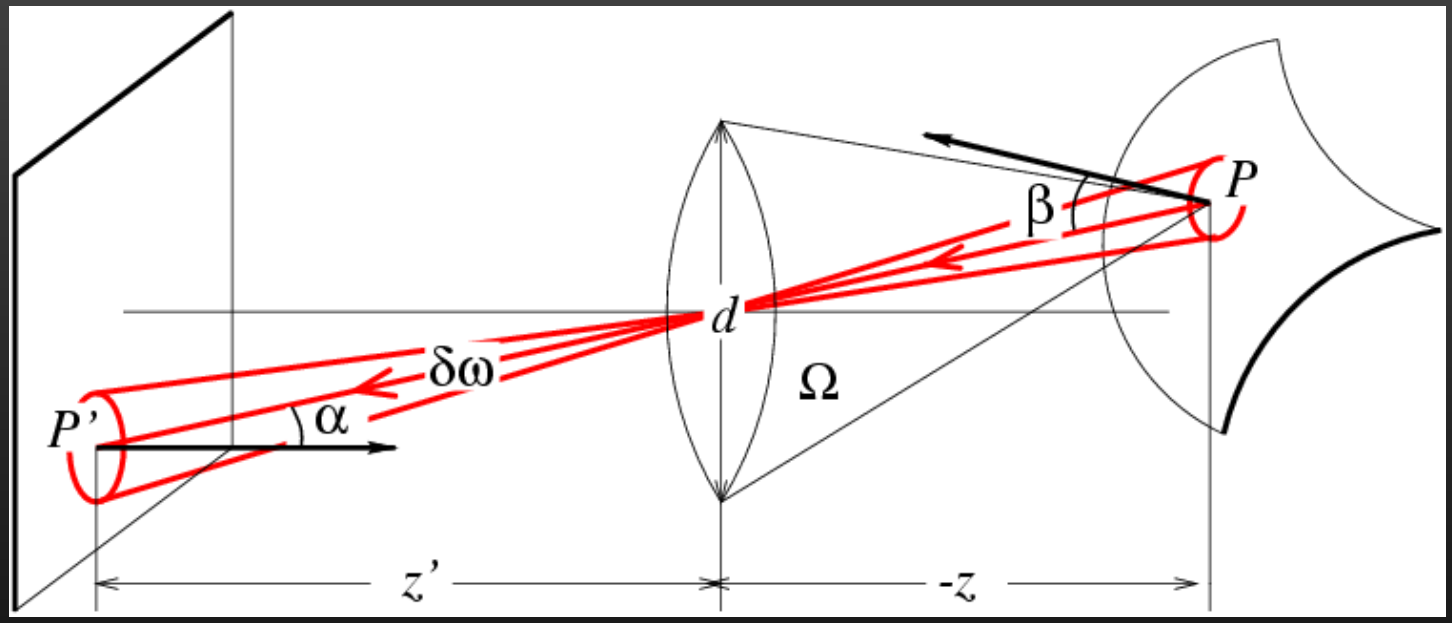


Camera Obscura in Edinburgh

Advantages of lens systems

Lenses

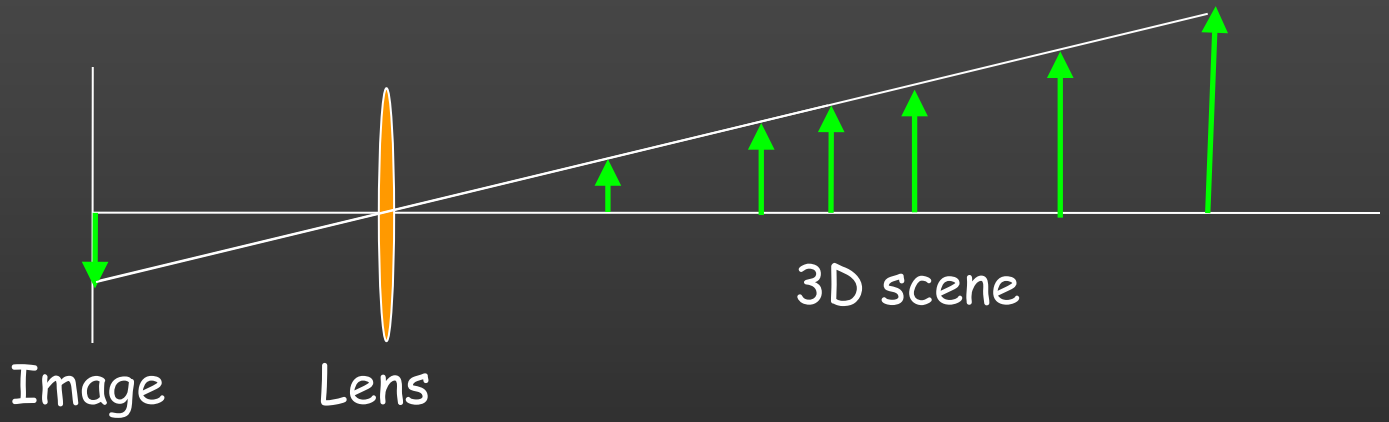
- can focus sharply on close and distant objects
- transmit more light than a pinhole camera

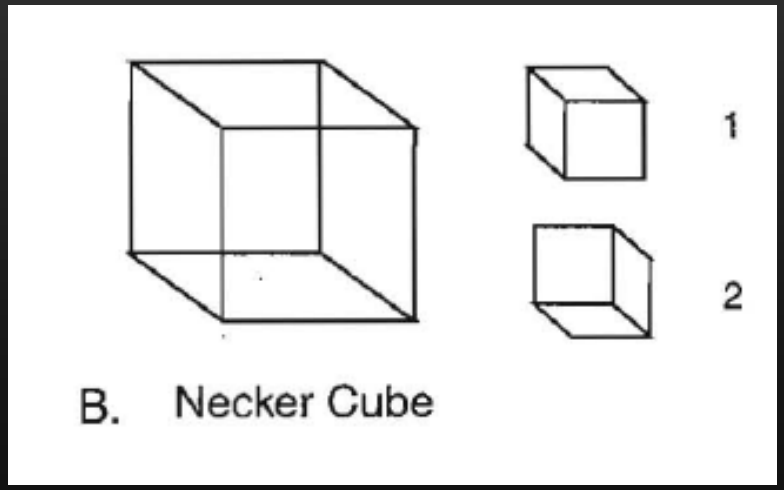


$$E=(\Pi/4)\ [\ (d/z')^2 \cos^4\alpha\ ]\ L$$

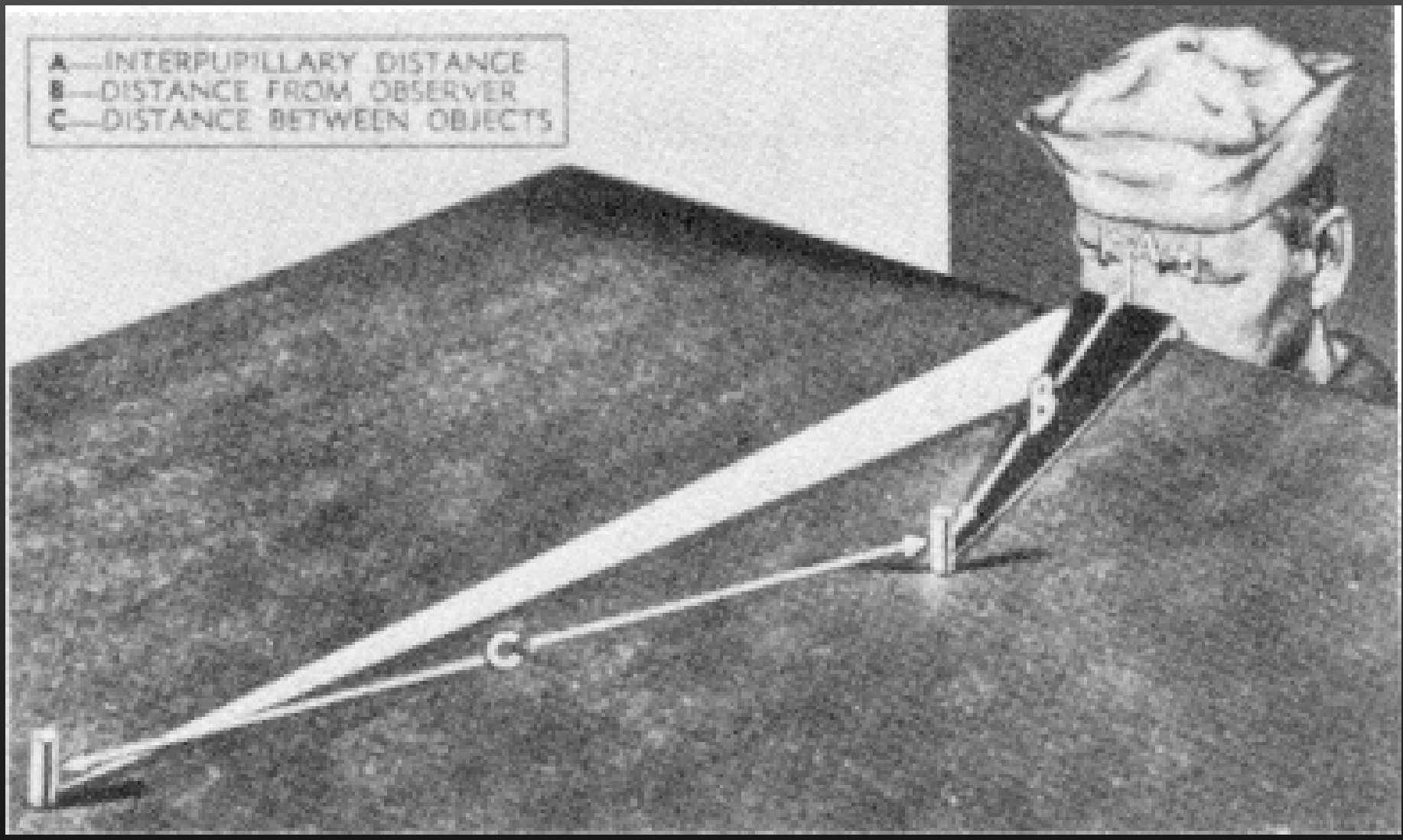
Fundamental problem I:
3D world is "flattened" to 2D images

➡ Loss of information





Question : how do we see "in 3D" ?

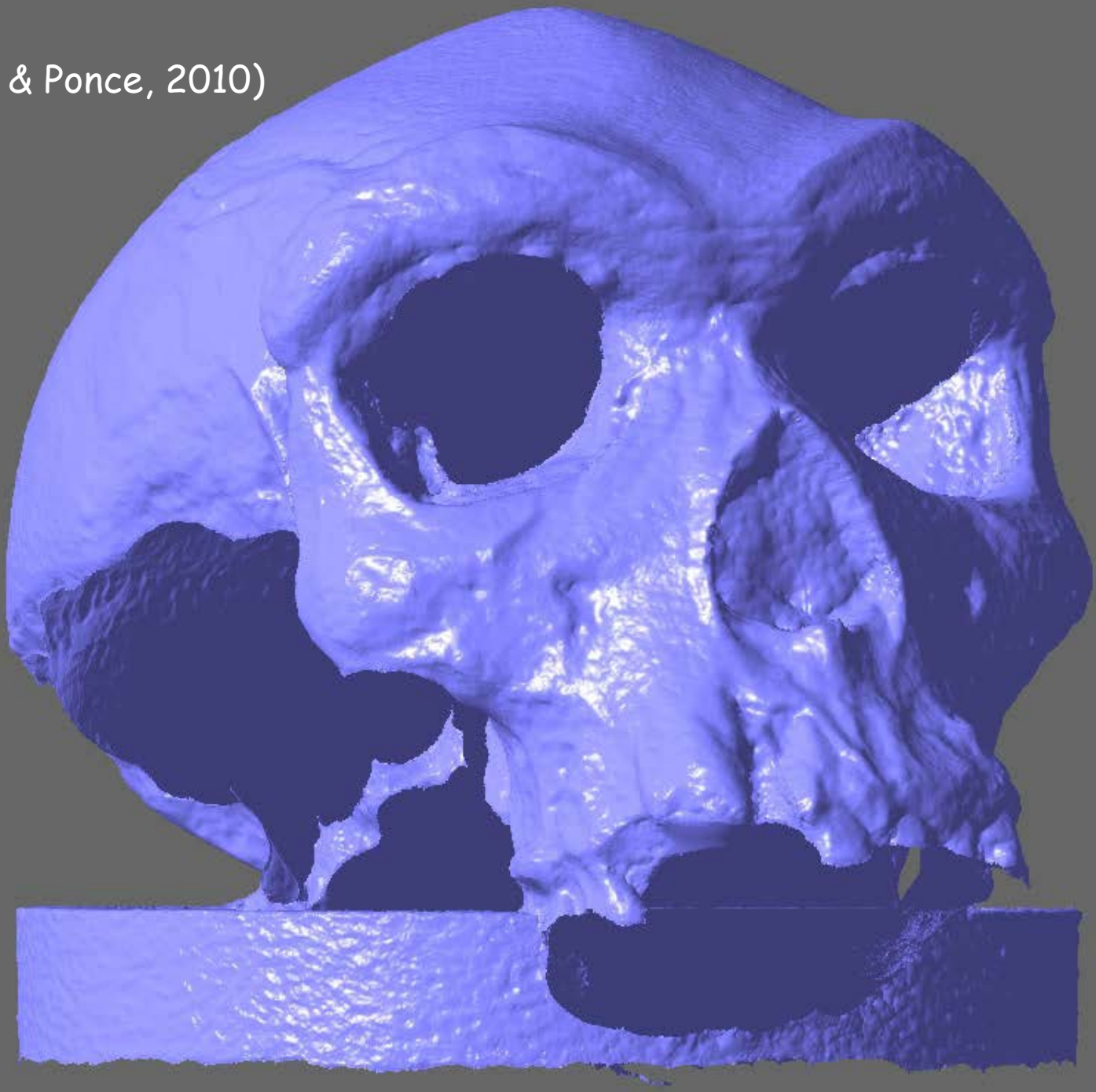


(First-order) answer: with our two eyes.

Simulated 3D perception

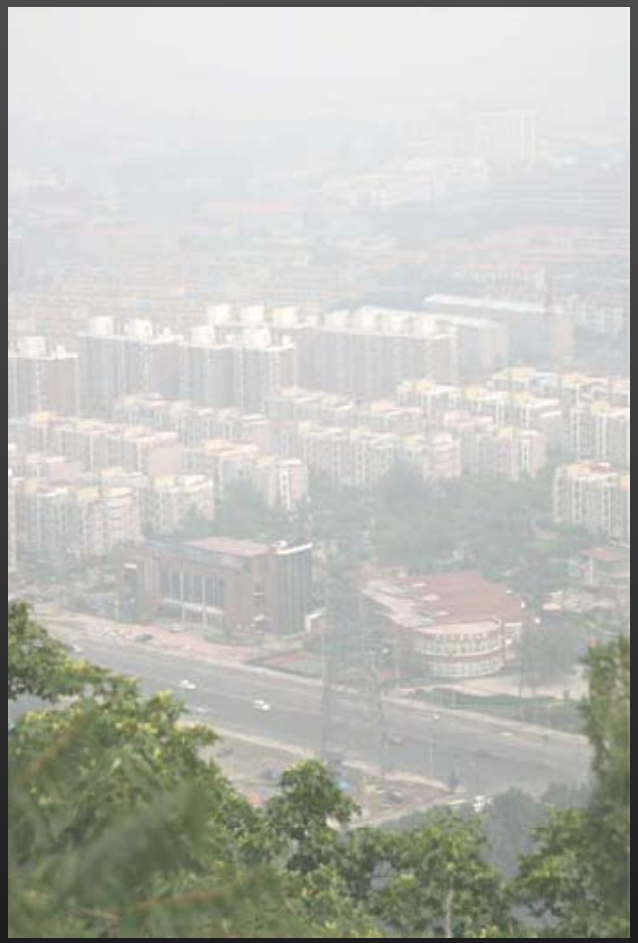
PMVS

(Furukawa & Ponce, 2010)



But there are other cues..

Depth from haze

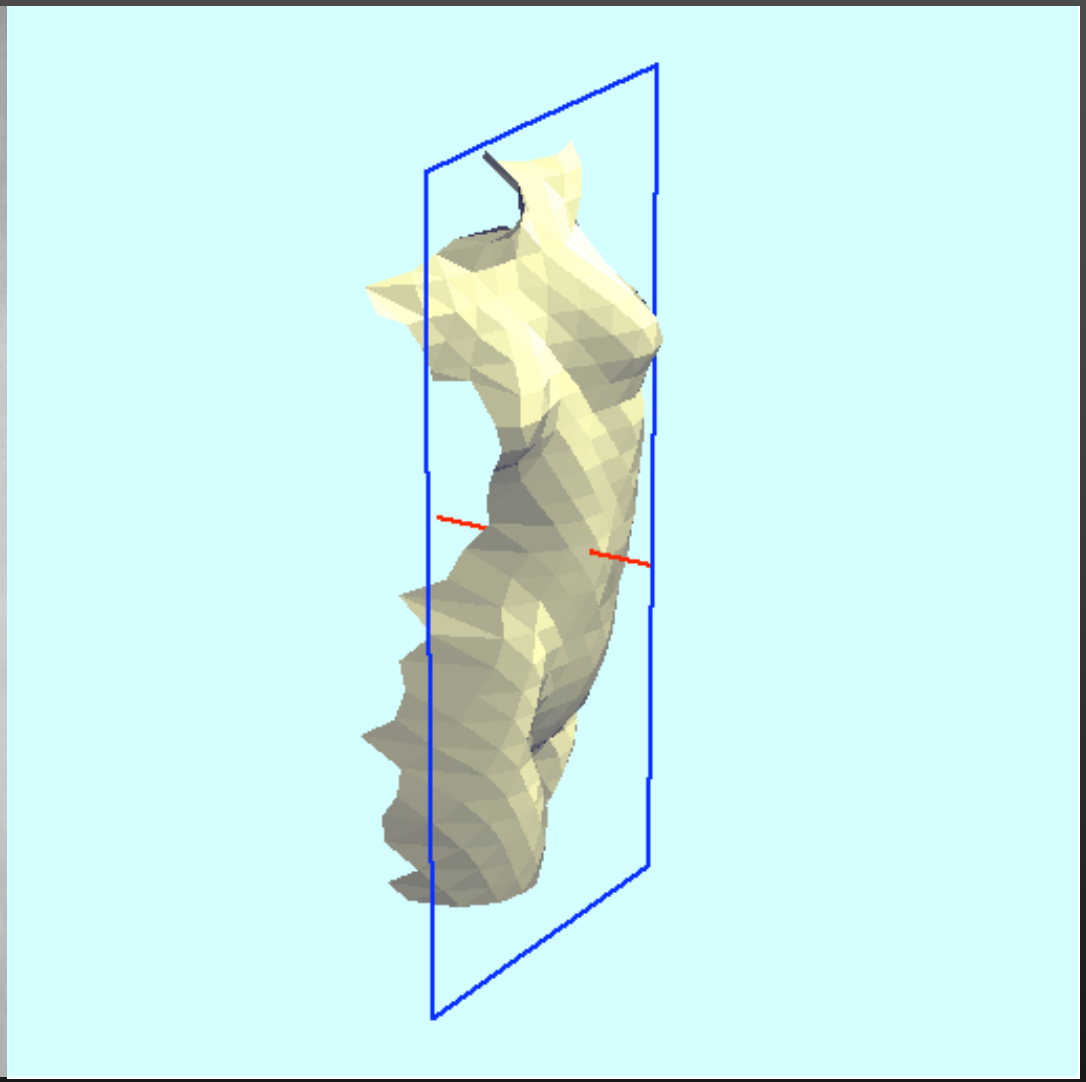


Input haze image

Reconstructed images

Recovered depth map

[K. HE, J. Sun and X. Tang, CVPR 2009]



Source: J. Koenderink

Source: J. Koenderink

What is happening with the shadows?

Image source: F. Durand

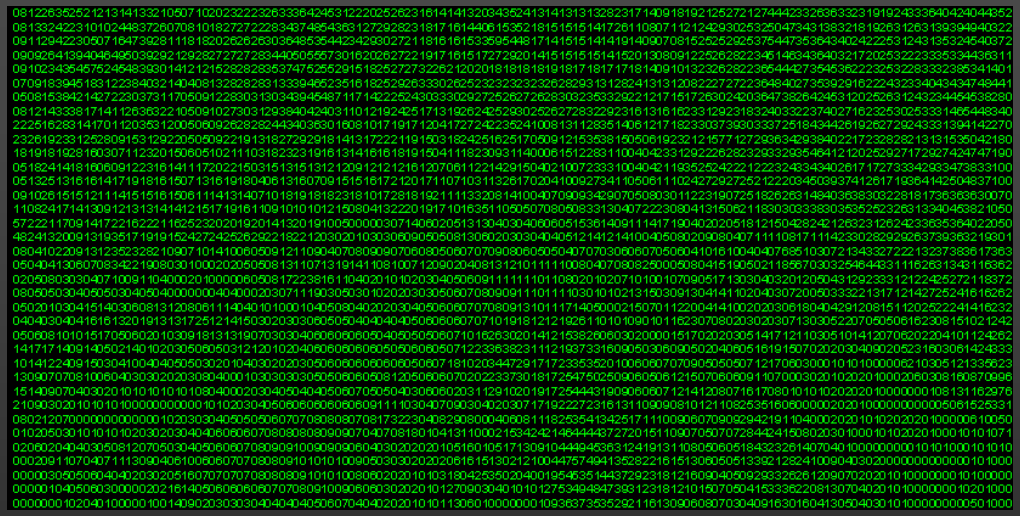
Challenges or opportunities?

Image source: J. Koenderink

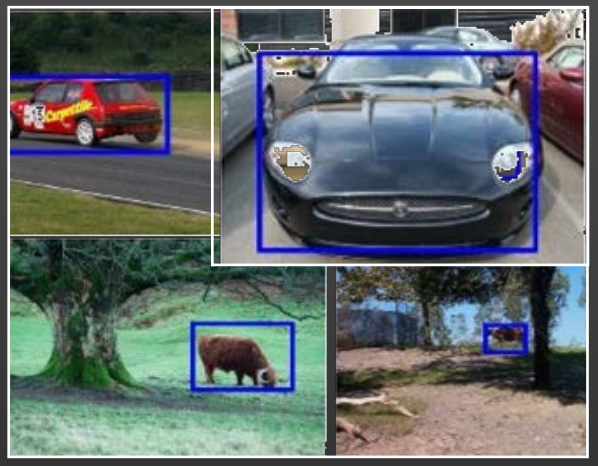
- Images are confusing, but they also reveal the structure of the world through numerous cues.
- Our job is to interpret the cues!

But we want much more than 3D: ex: Visual scene analysis

How to make sense of “pixel chaos”?



Object class recognition



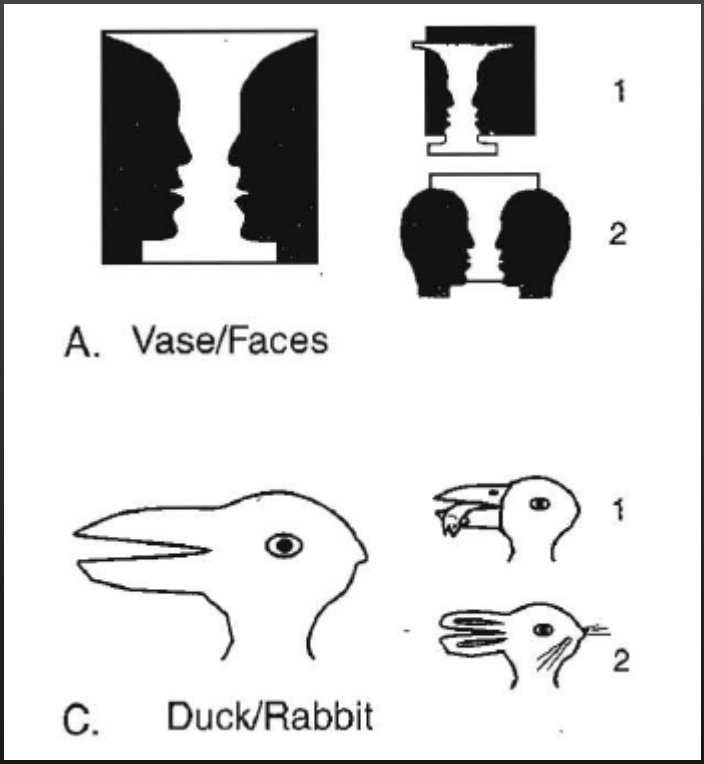
Face recognition

Action recognition

3D Scene reconstruction

Fundamental problem II: Cameras do not measure semantics

- We need lots of prior knowledge to make meaningful interpretations of an image



Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry

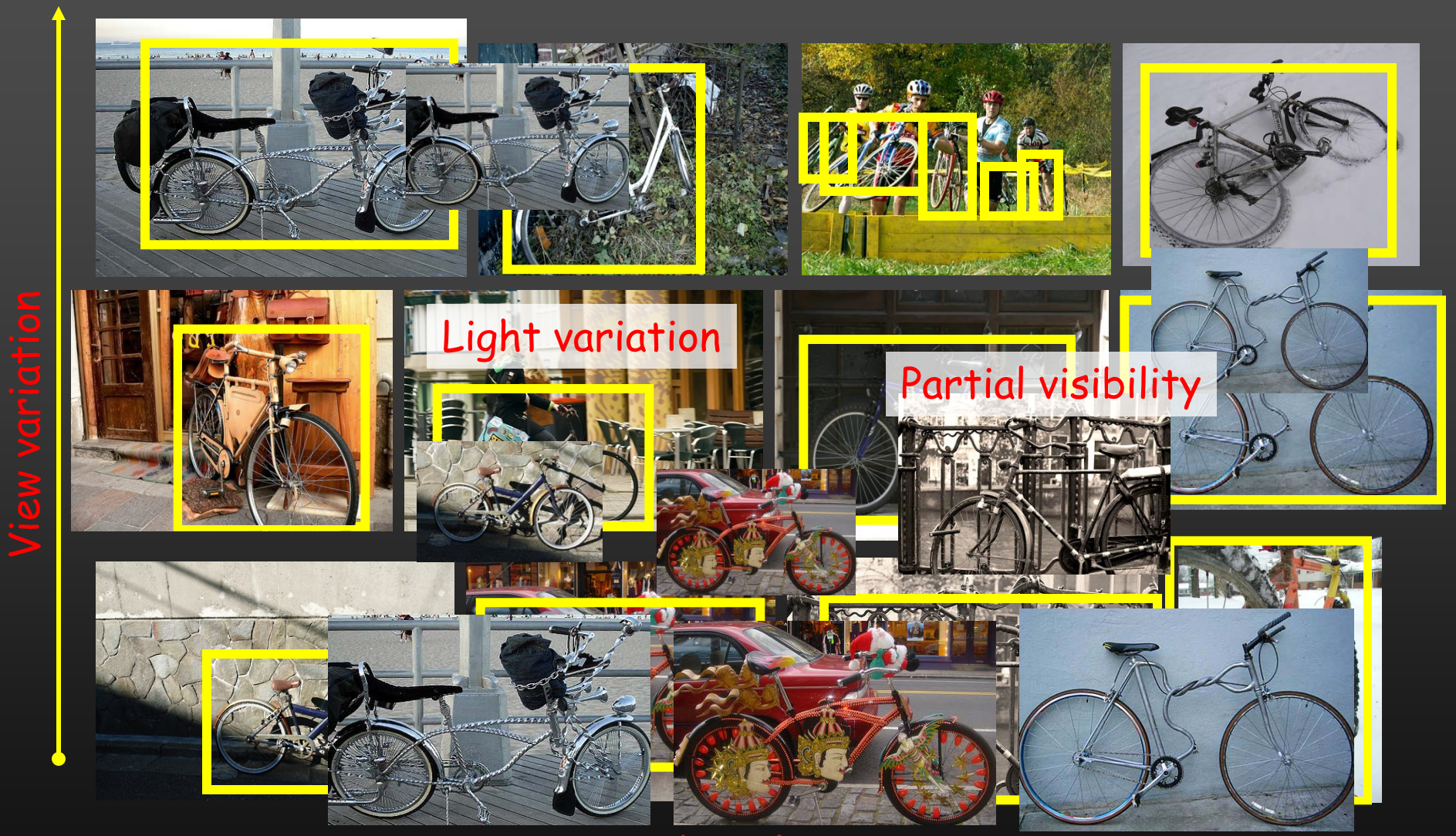
Specific object detection

(Lowe, 2004)

Image classification

Caltech 101 : http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Object category detection



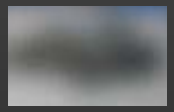
Example: part-based models

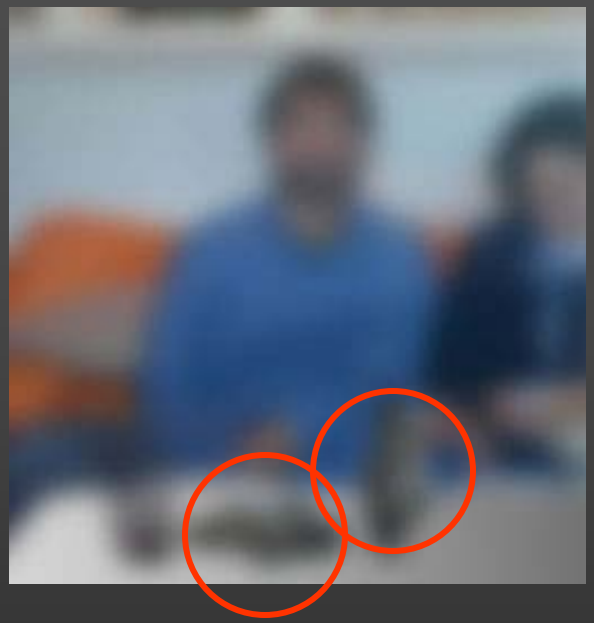
Qualitative experiments on Pascal VOC'07 (Kushal, Schmid, Ponce, 2008)

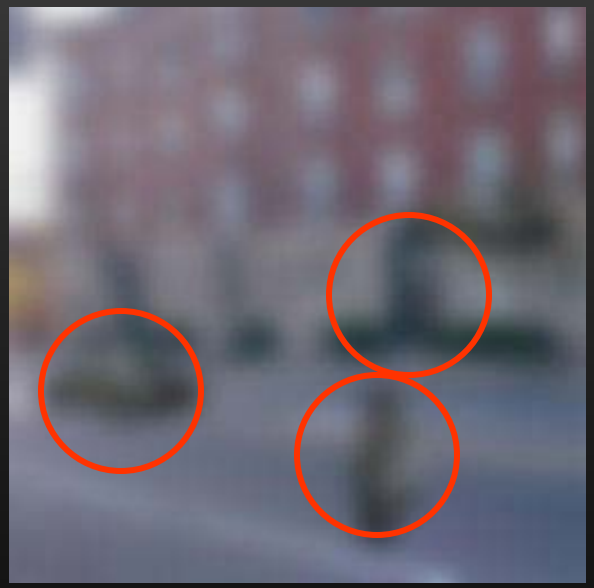
Scene understanding

Photo courtesy A. Efros.

Local ambiguity and global scene interpretation







slide credit: Fei-Fei, Fergus & Torralba

This class

1. Introduction plus recap on geometry (J. Ponce)
2. Instance-level recognition I. - Local invariant features (C. Schmid)
3. Instance-level recognition II. - Correspondence, efficient visual search (I. Laptev)
4. Very large scale image indexing; bag-of-feature models for category-level recognition (C. Schmid)
5. Sparse coding (J. Ponce); category-level localization I (J. Sivic)
6. Neural networks; optimization
7. Category-level localization II; pictorial structures; human pose (J. Sivic)
8. Motion and human action (I. Laptev)
9. Face detection and recognition; segmentation (C. Schmid)
10. Scenes and objects (J. Sivic)
11. Final project presentations (J. Sivic, I. Laptev)

Computer vision books

- D.A. Forsyth and J. Ponce, "Computer Vision: A Modern Approach, Prentice-Hall, 2nd edition, 2011.

- J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, "Toward category-level object recognition", Springer LNCS, 2007.
- 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.
- J. Koenderink, "Solid Shape", MIT Press, 1990.

Class web-page

http://www.di.ens.fr/willow/teaching/recvis12/

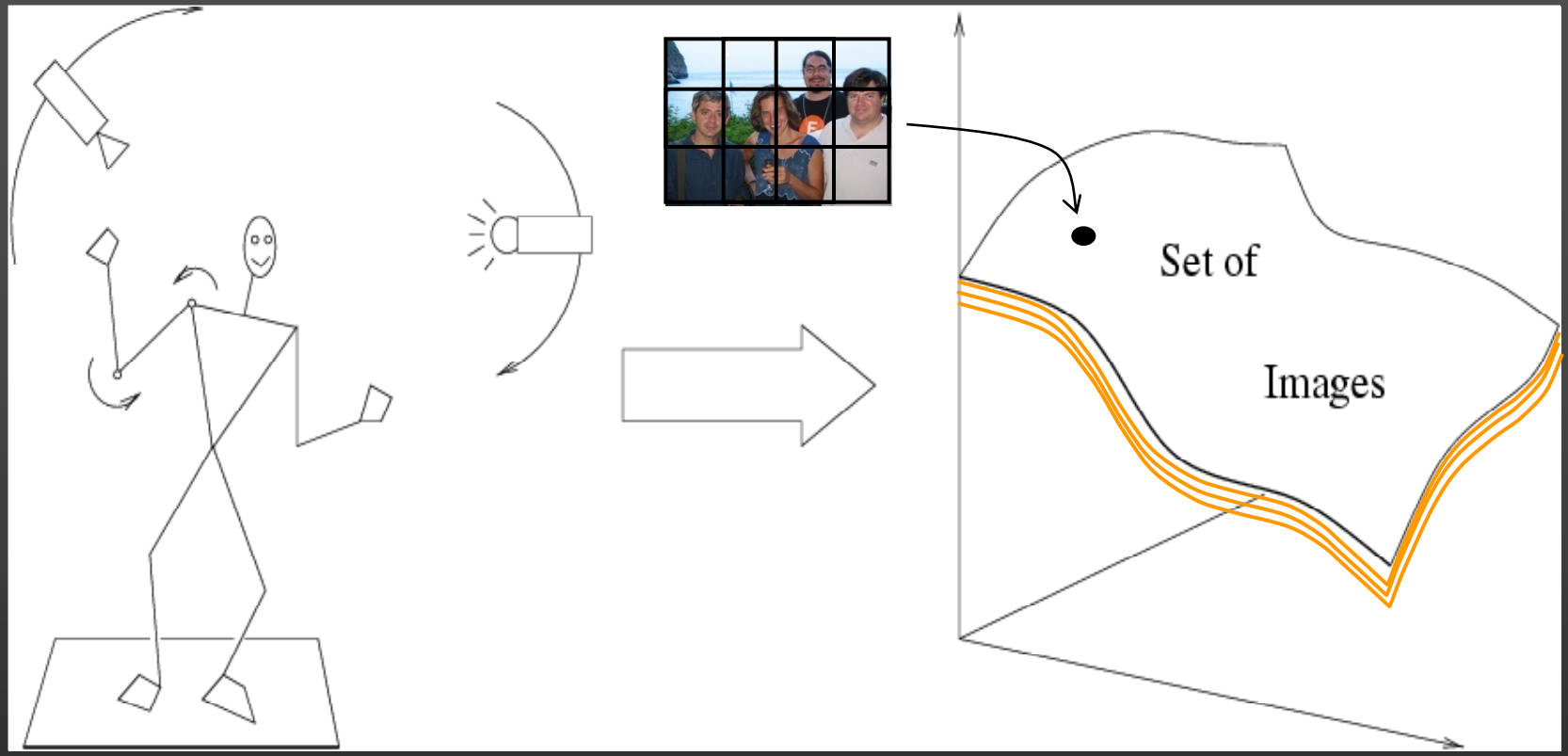
Slides available after classes:

http://www.di.ens.fr/willow/teaching/recvis12/lecture1.pptx
http://www.di.ens.fr/willow/teaching/recvis12/lecture1.pdf

Note: Much of the material used in this lecture is courtesy of Svetlana Lazebnik:, http://www.cs.illinois.edu/homes/slazebni/

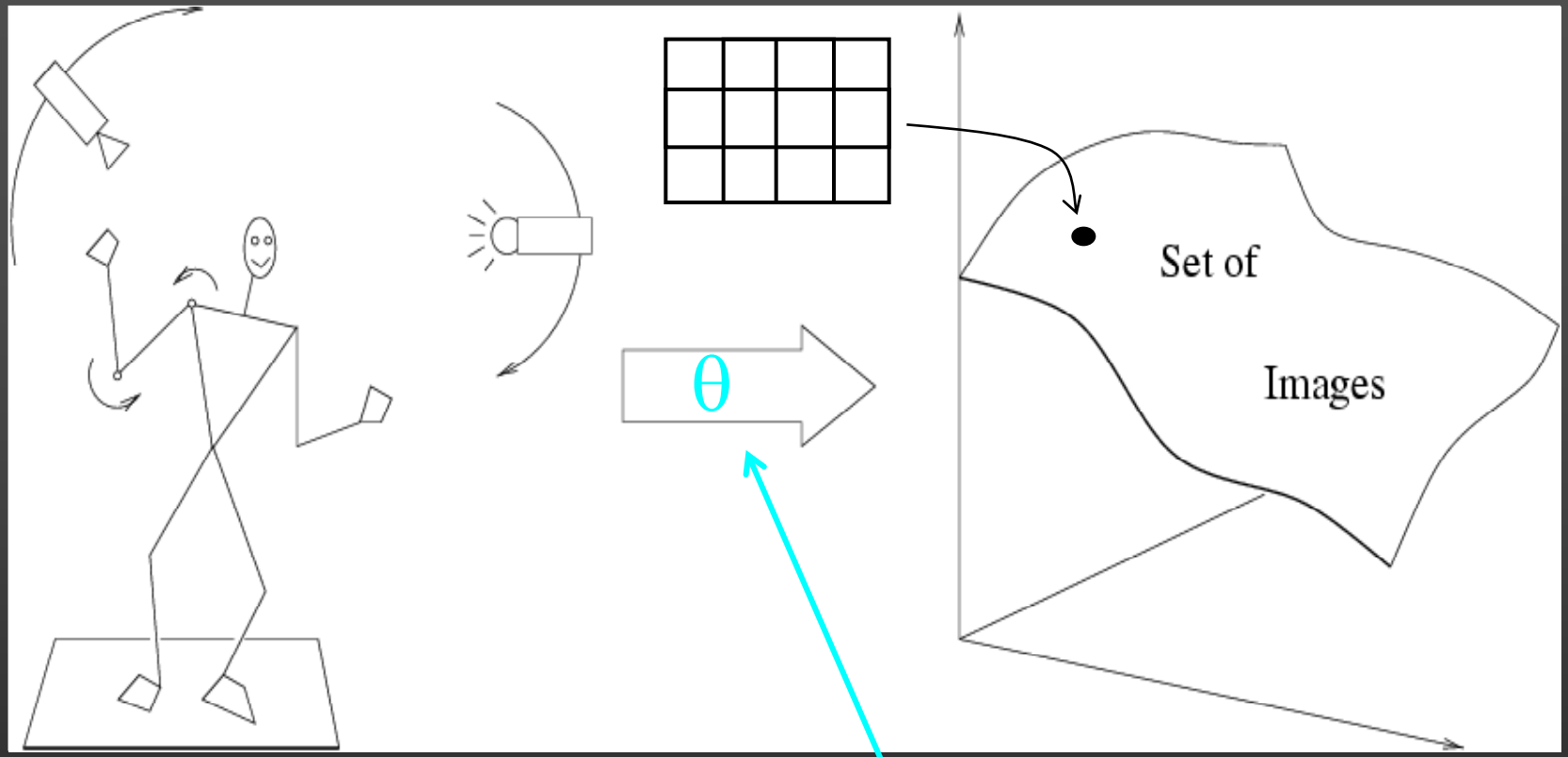
Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry



Variability:

Camera position
Illumination
Internal parameters
Within-class variations

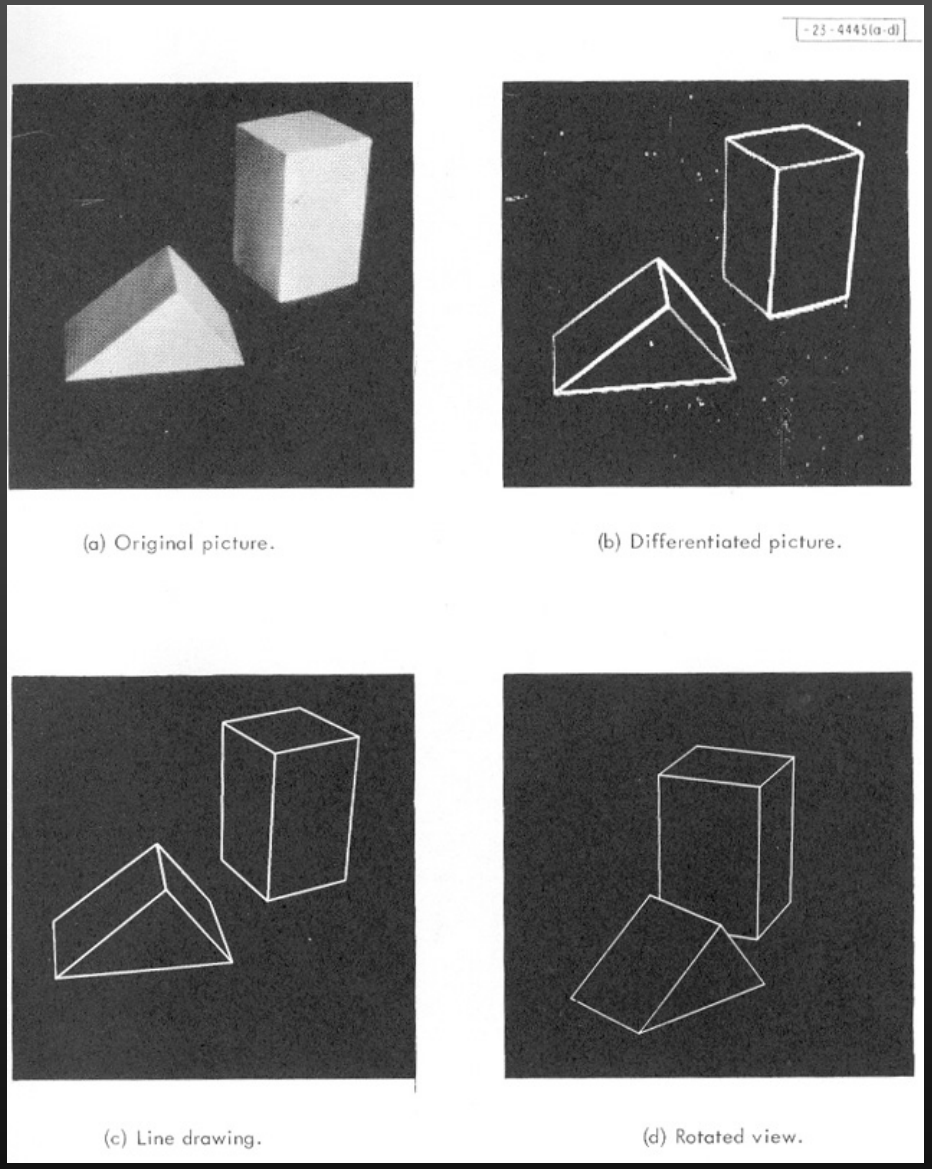


Variability:

Camera position
Illumination
Internal parameters

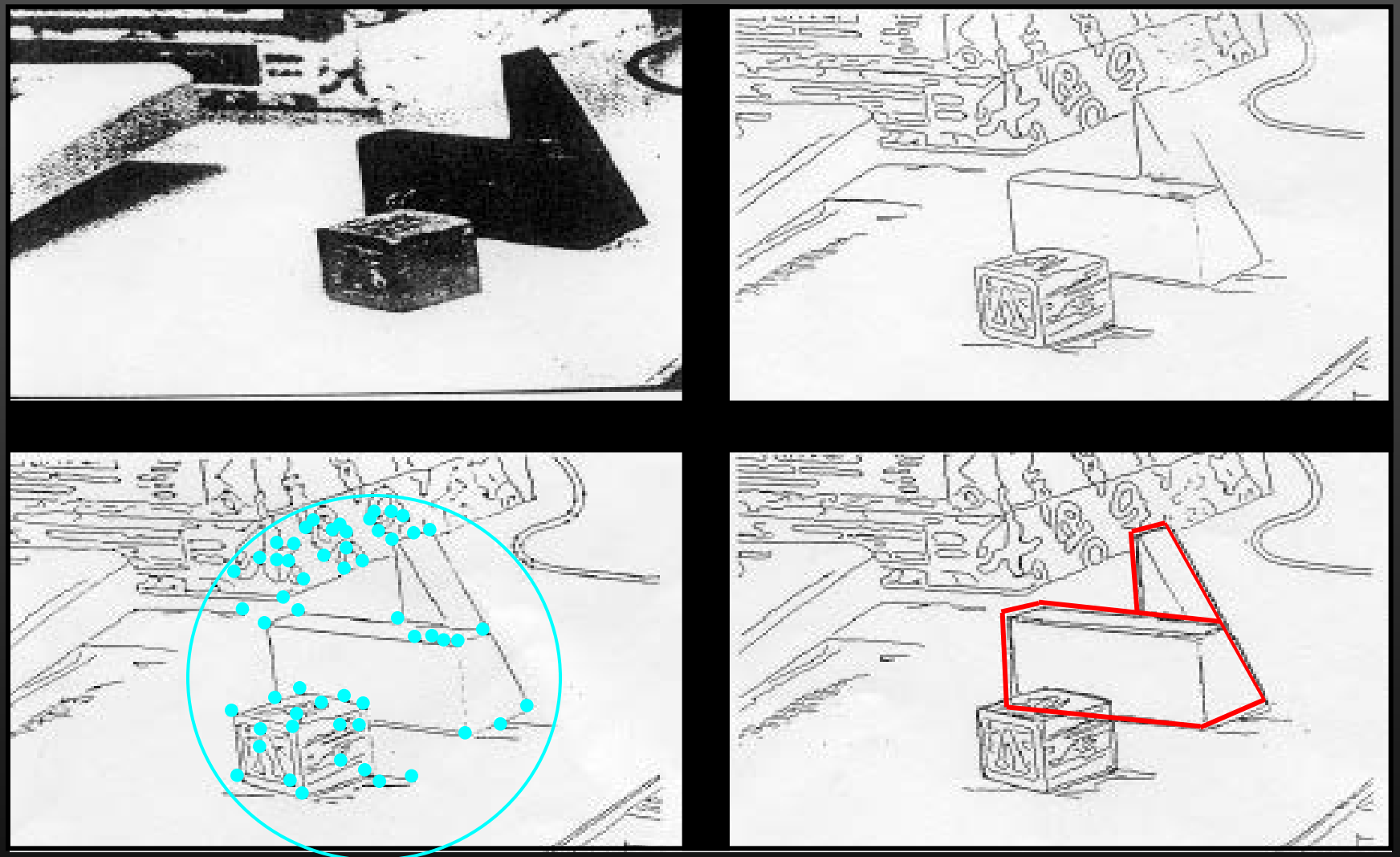
Roberts (1963); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

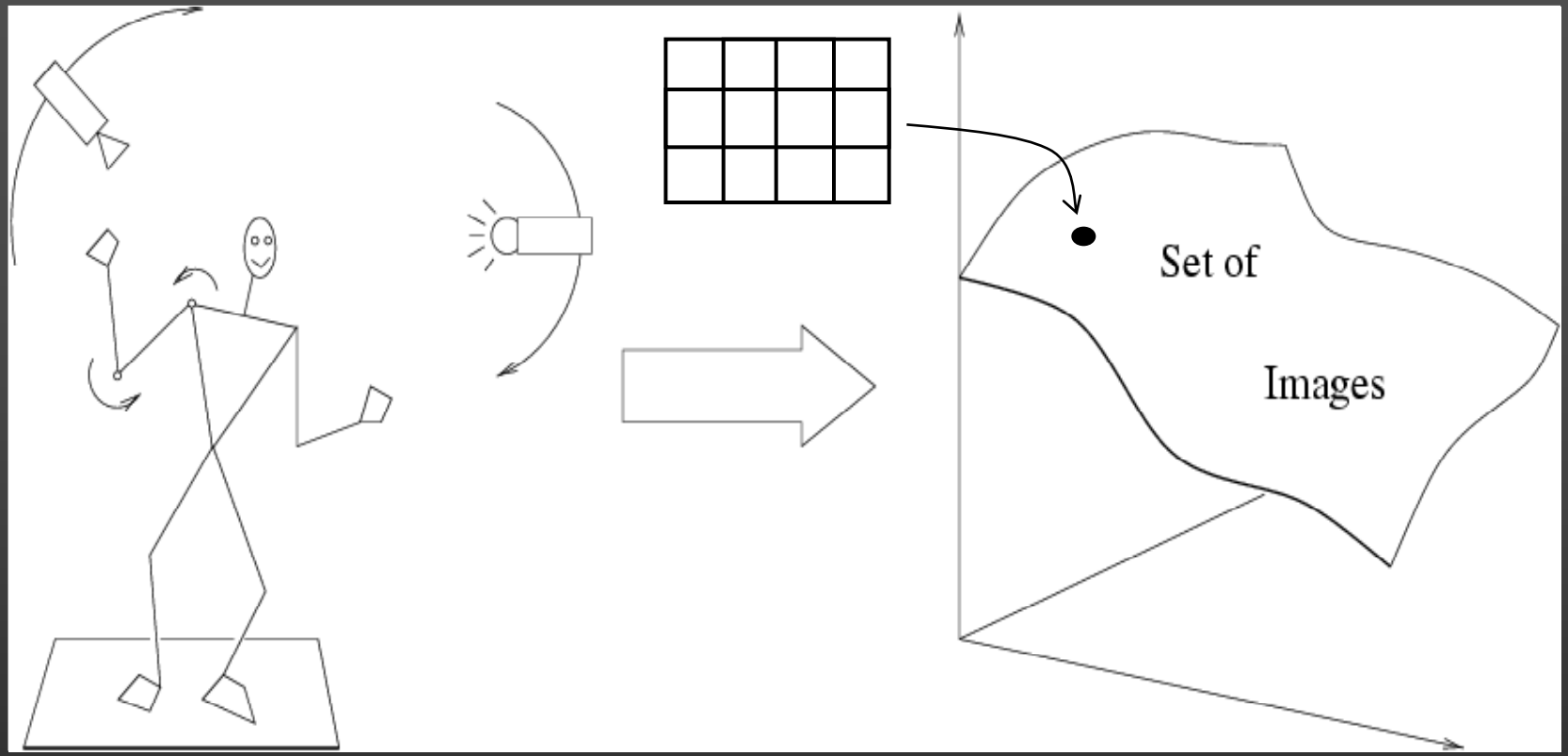
Origins of computer vision



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

Huttenlocher & Ullman (1987)



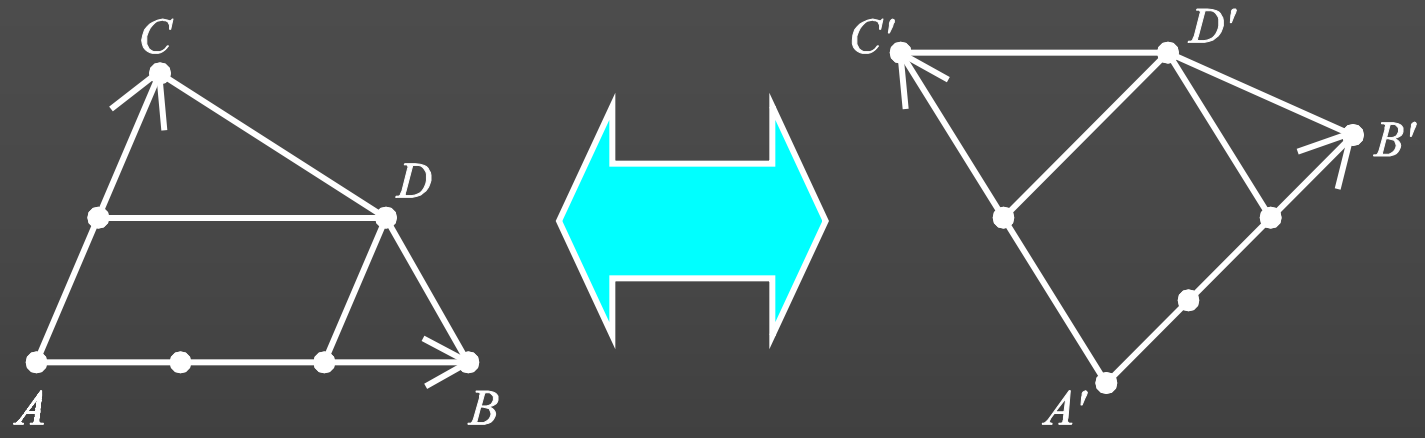


~~Variability~~ Invariance to:

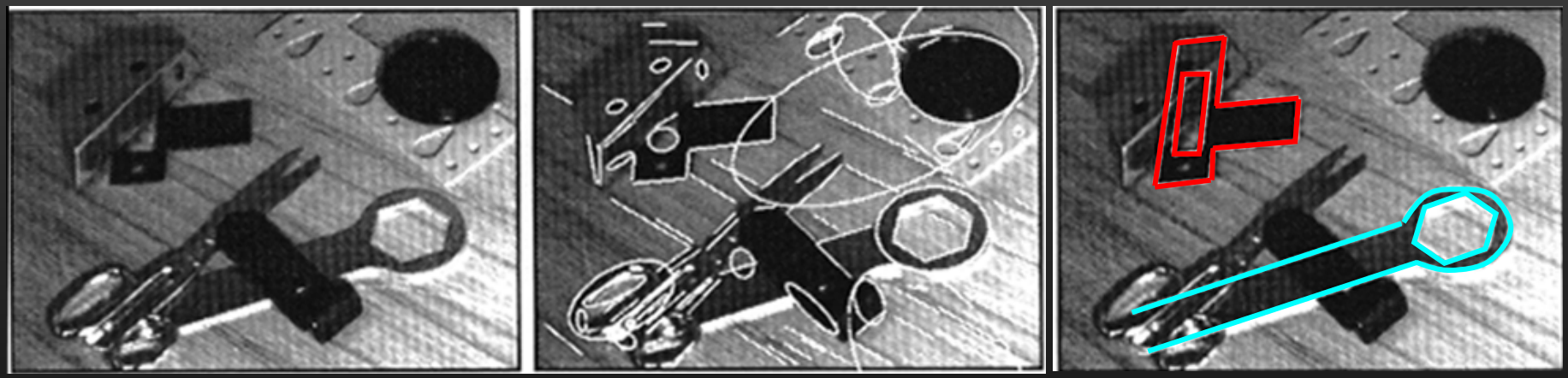
- Camera position
- Illumination
- Internal parameters

Duda & Hart (1972); Weiss (1987); Mundy et al. (1992-94); Rothwell et al. (1992); Burns et al. (1993)

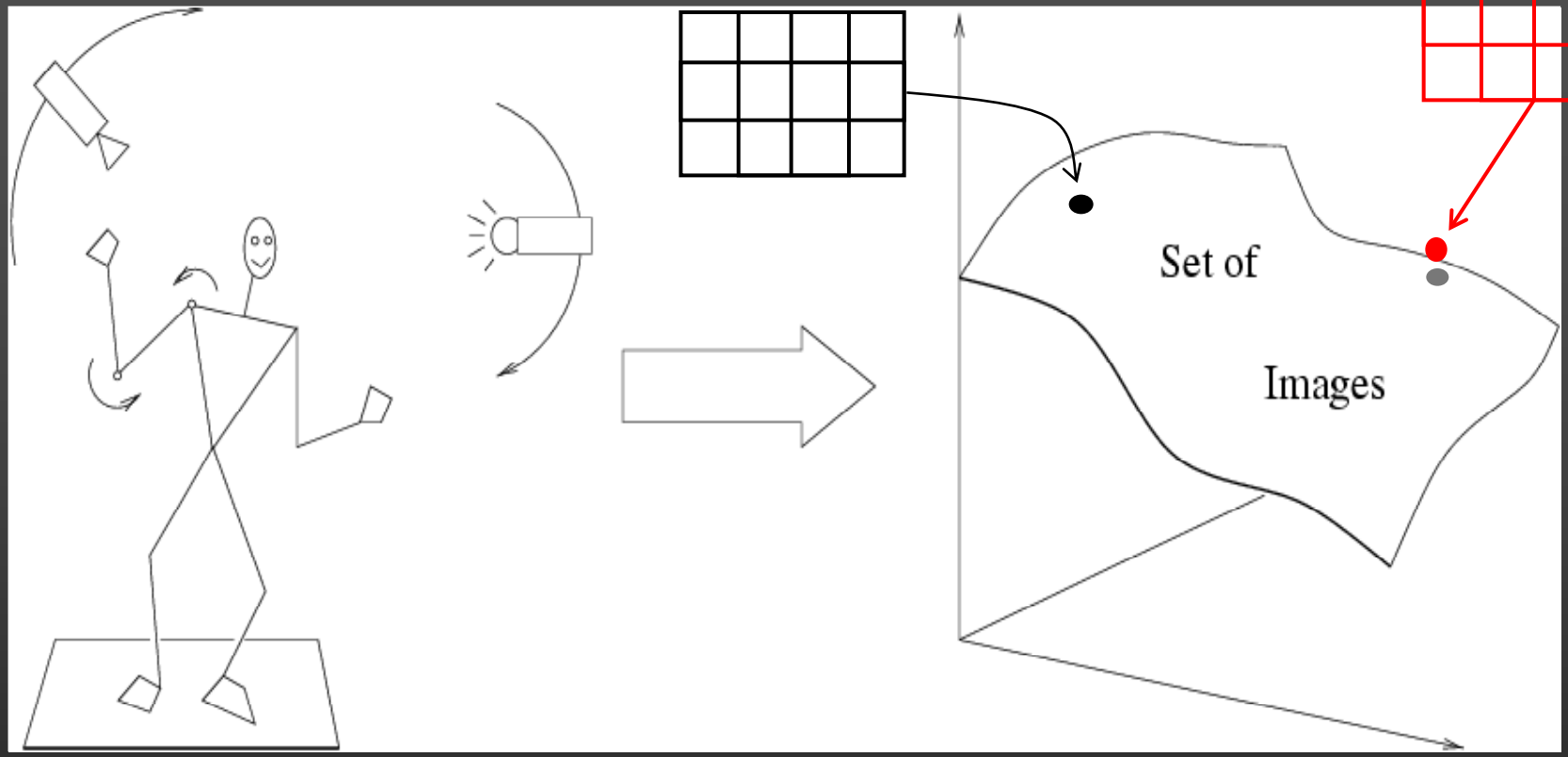
Example: affine invariants of coplanar points



Projective invariants (Rothwell et al., 1992):



BUT: True 3D objects do not admit monocular viewpoint invariants (Burns et al., 1993) !!



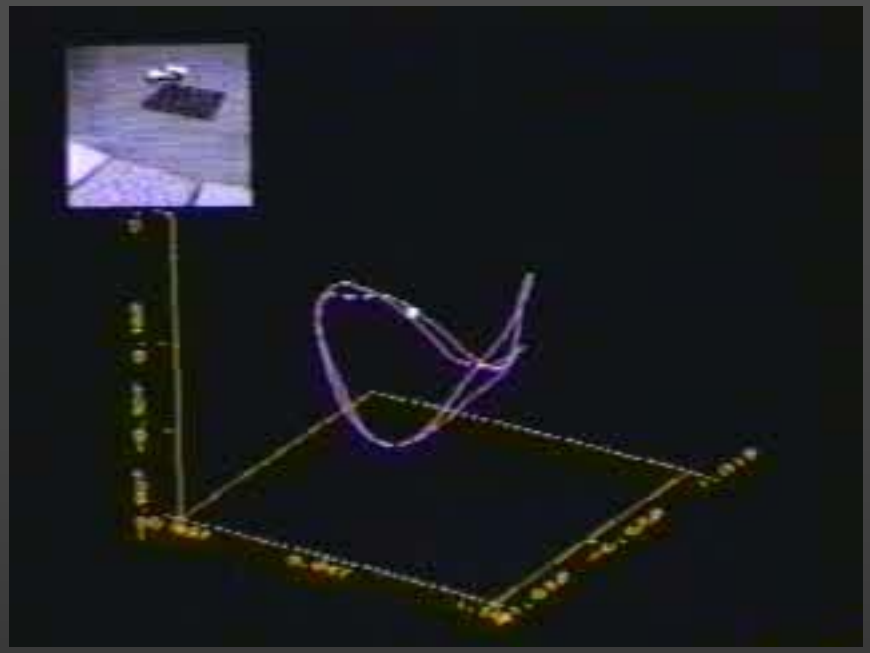
Empirical models of image variability:

Appearance-based techniques

Turk & Pentland (1991); Murase & Nayar (1995); etc.

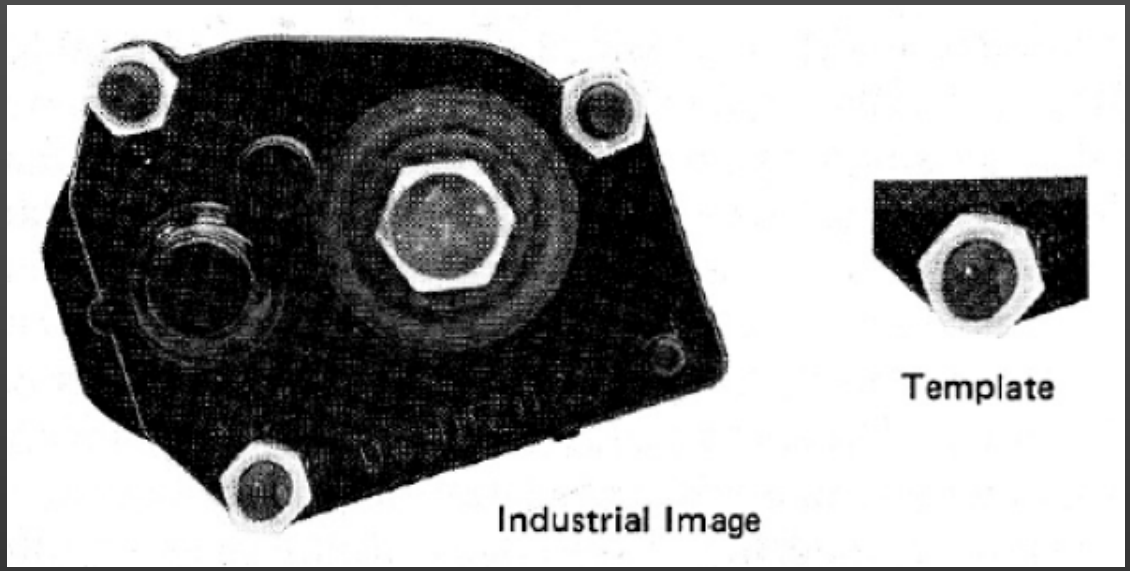
Eigenfaces (Turk & Pentland, 1991)

Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20



Appearance manifolds
(Murase & Nayar, 1995)

Correlation-based template matching (60s)



Ballard & Brown (1980, Fig. 3.3). Courtesy Bob Fisher and Ballard & Brown on-line.

- Automated target recognition
- Industrial inspection
- Optical character recognition
- Stereo matching
- Pattern recognition

In the late 1990s, a new approach emerges:
Combining *local* appearance, spatial constraints, invariants, and classification techniques from machine learning.

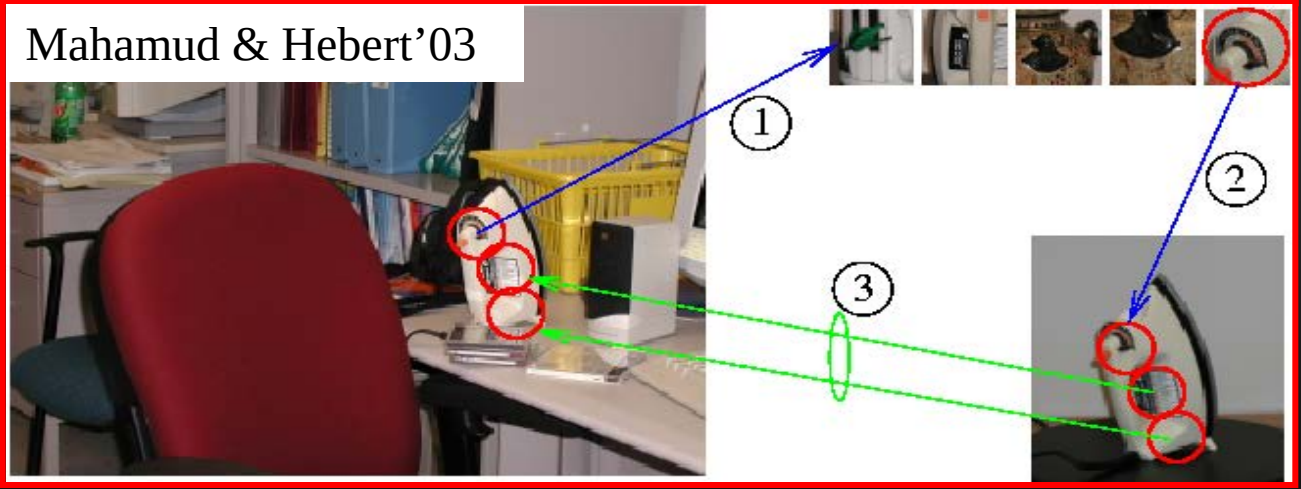
Query

Retrieved (10° off)

Schmid & Mohr'97

Lowe'02

Mahamud & Hebert'03



Late 1990s: Local appearance models

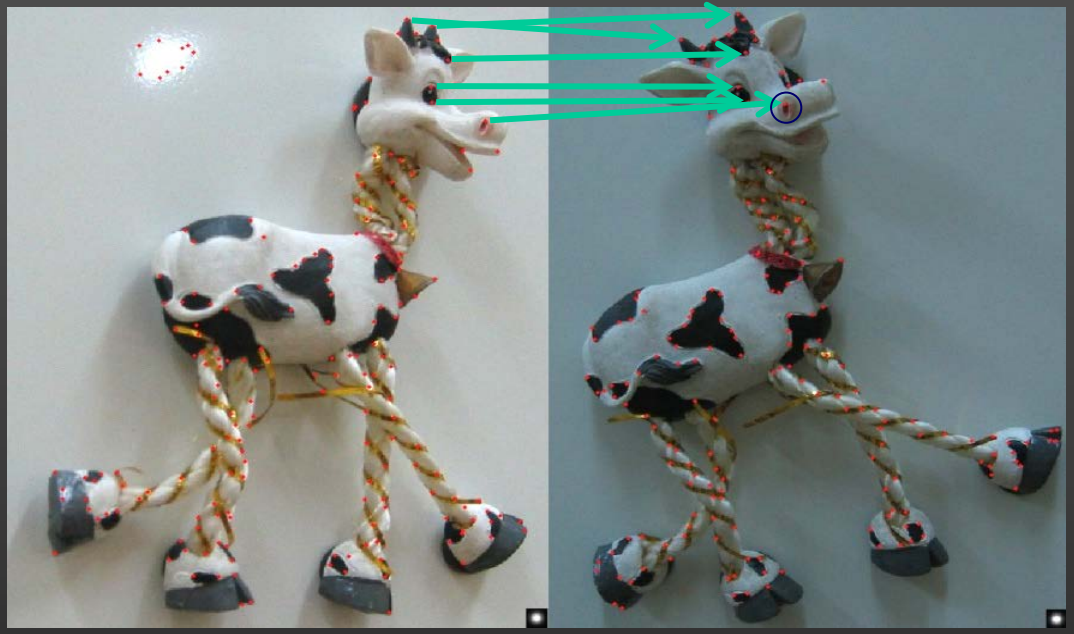
(Image courtesy of C. Schmid)

Late 1990s: Local appearance models

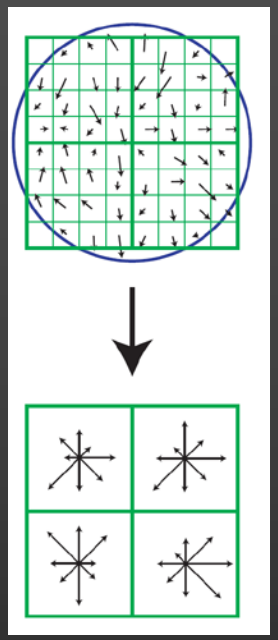
(Image courtesy of C. Schmid)

- Find features (interest points).

Late 1990s: Local appearance models



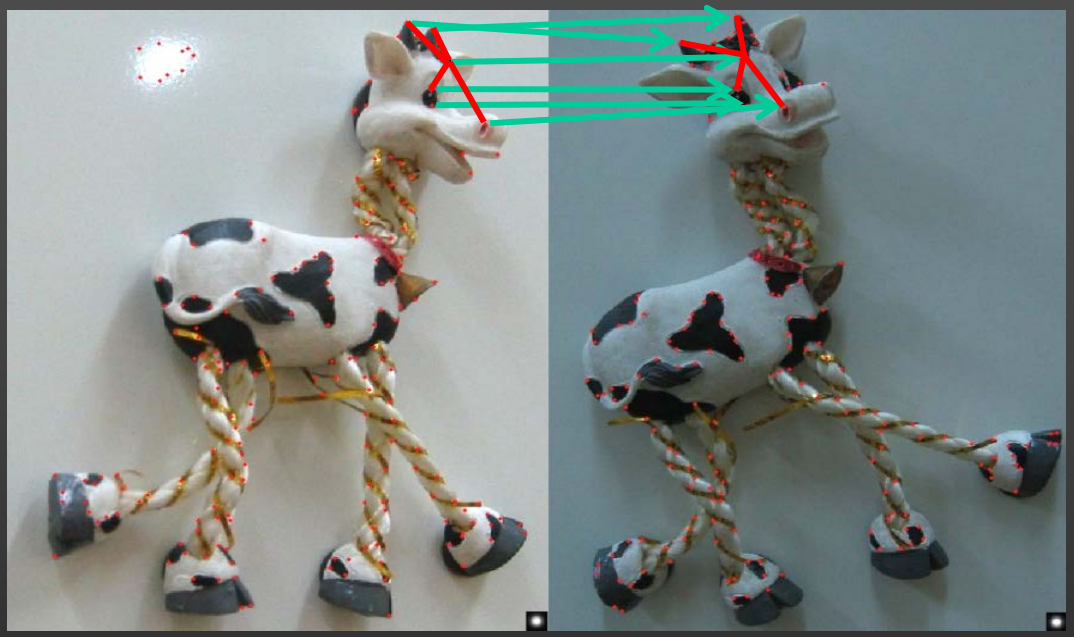
(Image courtesy of C. Schmid)



(Lowe 2004)

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

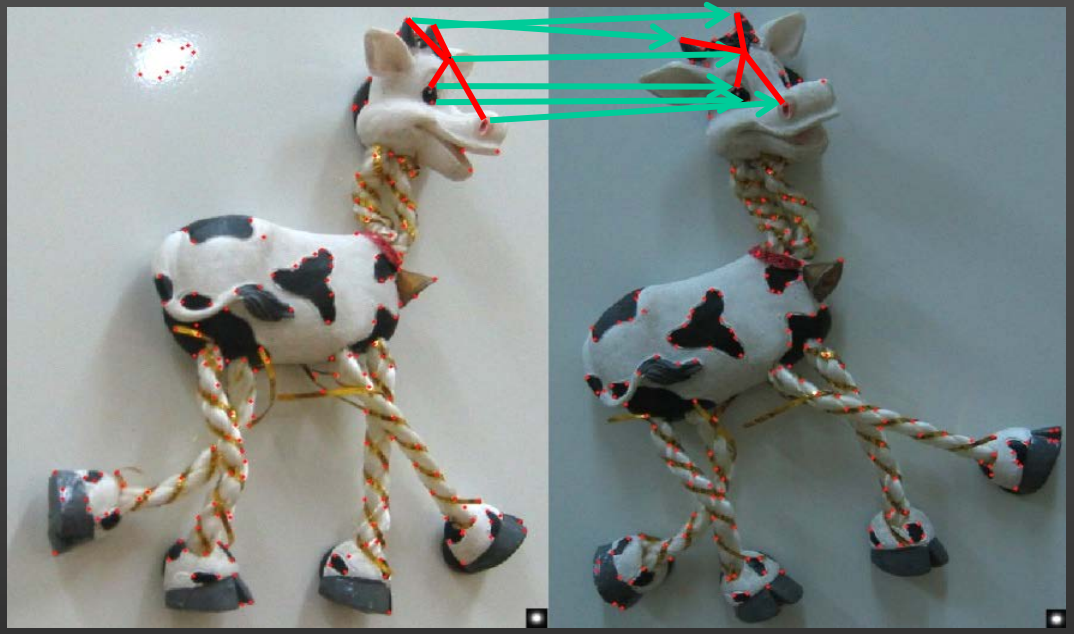
Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).
- Optional: Filter out outliers using geometric consistency.

Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).
- Optional: Filter out outliers using geometric consistency.
- Vote.

See, for example, Schmid & Mohr (1996); Lowe (1999);Tuytelaars & Van Gool, (2002); Rothganger et al. (2003); Ferrari et al., (2004).

Bags of words: Visual "Google"

(Sivic & Zisserman, ICCV' 03)

"Visual word" clusters

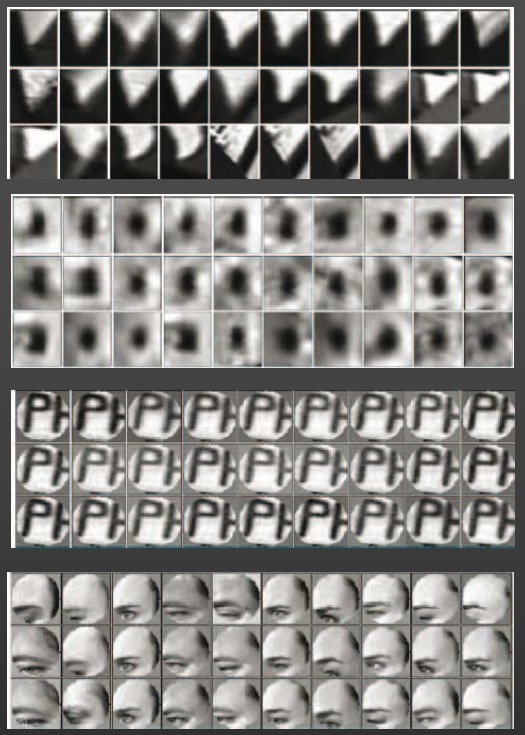
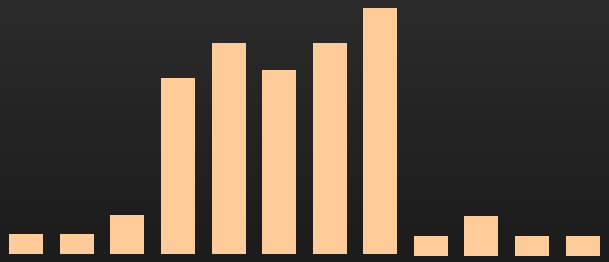


Image retrieval in videos



Vector quantization into histogram (the "bag of words")

Bags of words: Visual "Google"
(Sivic & Zisserman, ICCV' 03)

Retrieved shots

Select a region

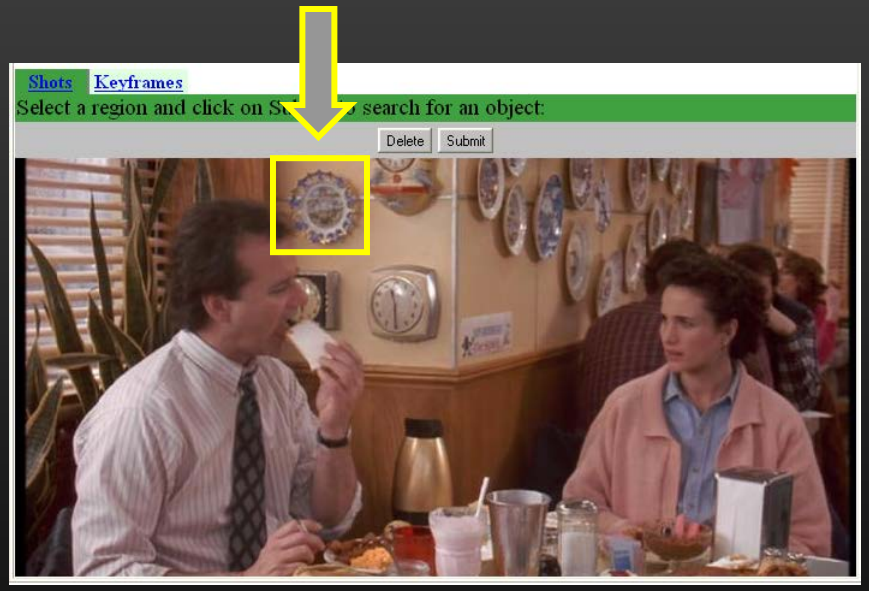
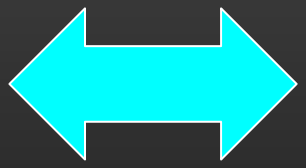
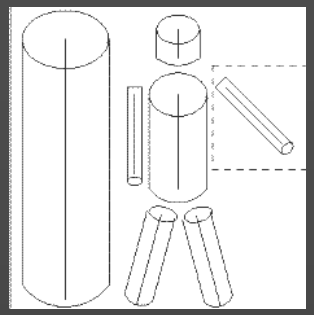


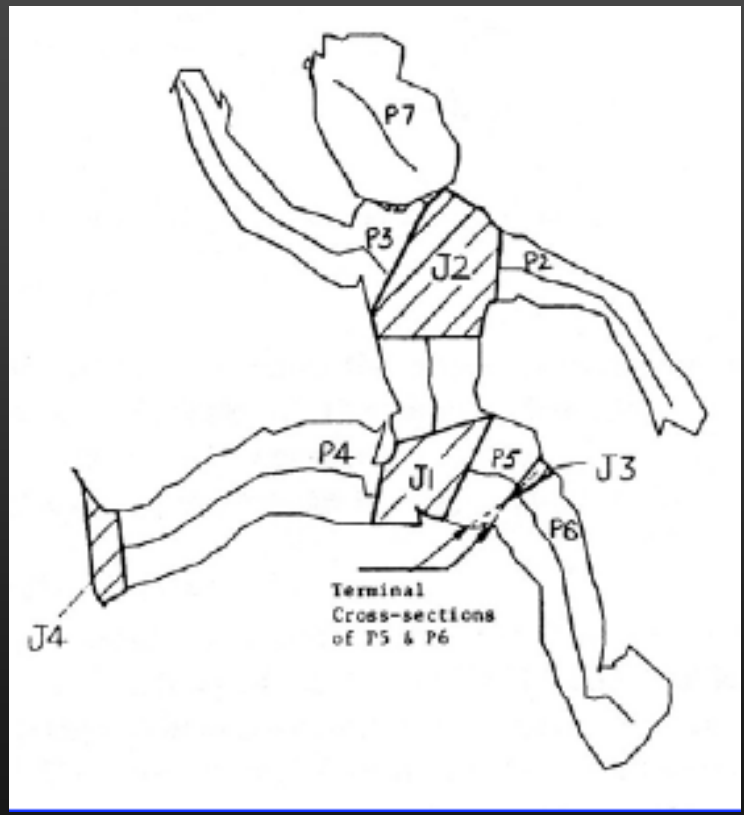
Image categorization is harder





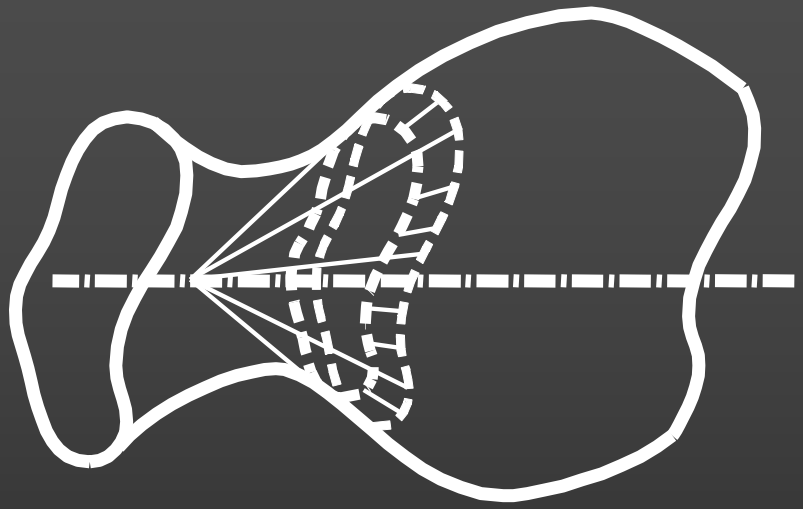
Structural part-based models

(Binford, 1971; Marr & Nishihara, 1978)

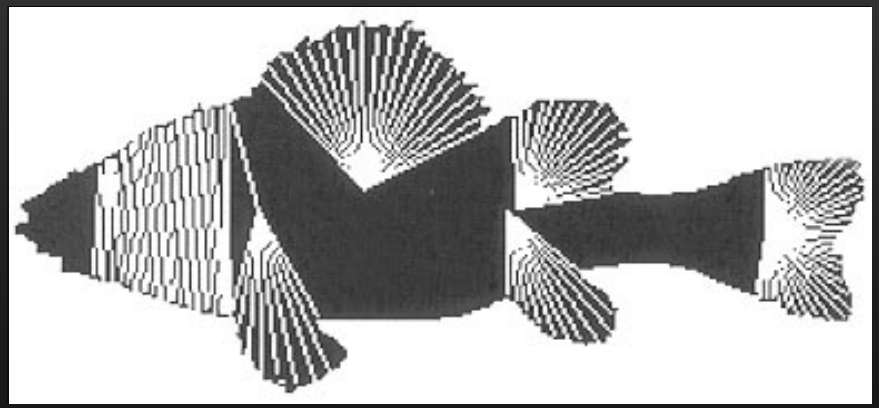


(Nevatia & Binford, 1972)

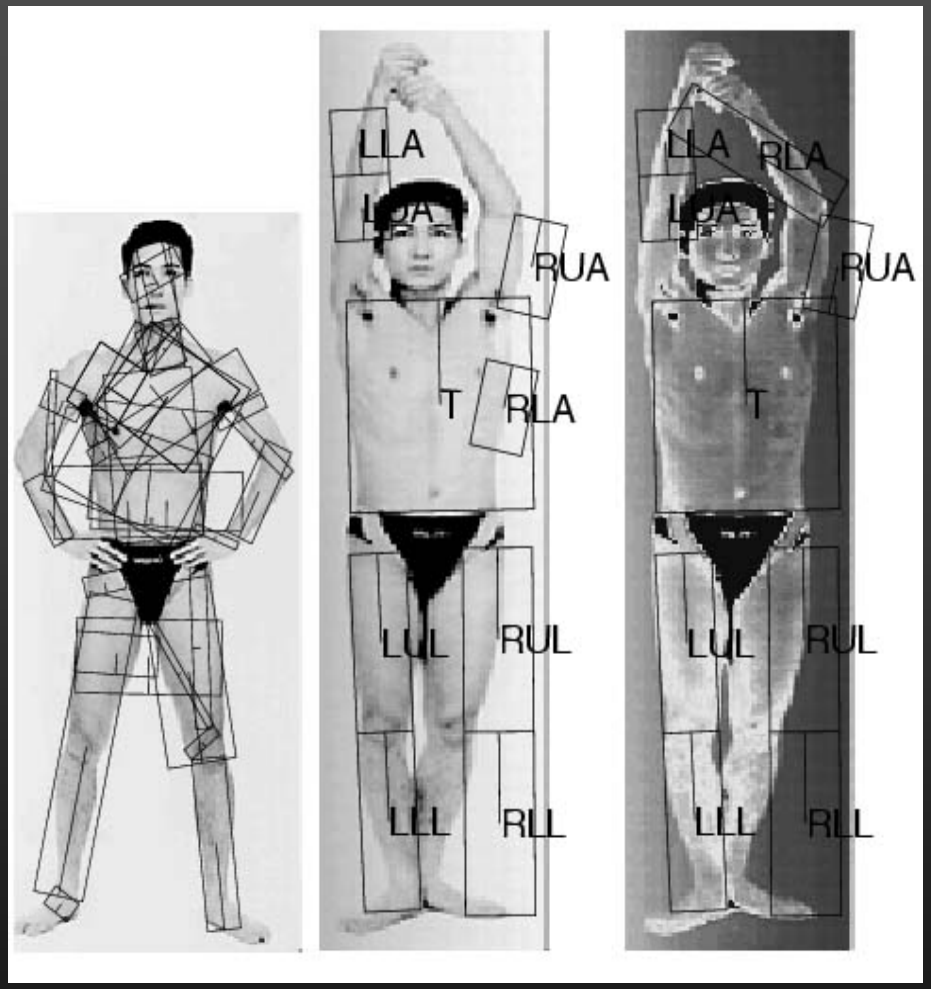
Helas, this is hard to operationalize



Ponce et al. (1989)

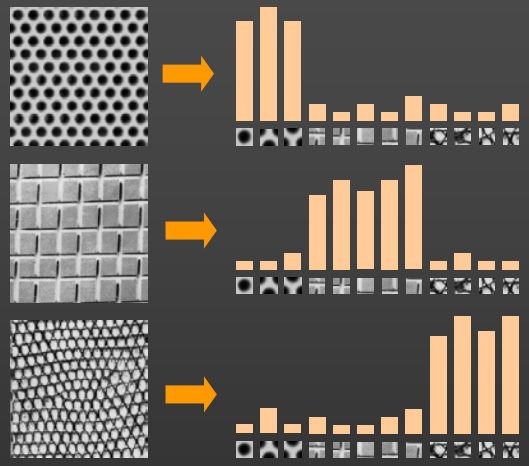


Zhu and Yuille (1996)



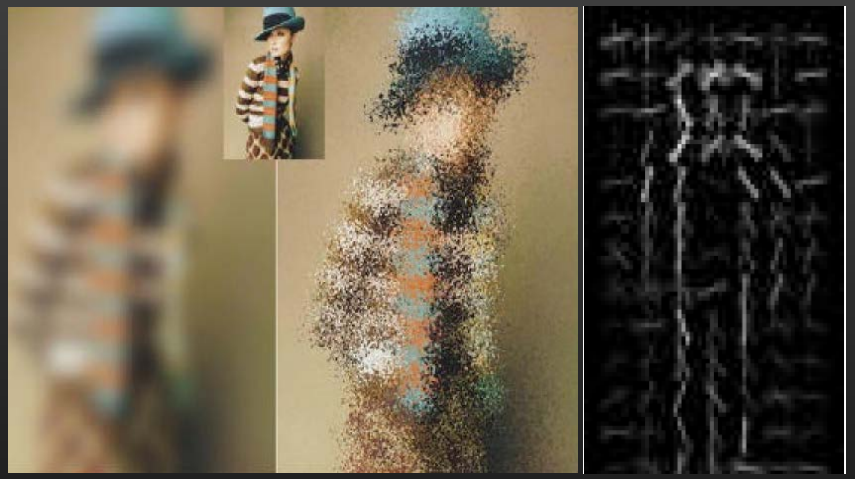
Ioffe and Forsyth (2000)

Bags of words and their variants have become the dominant model for image categorization



(Swain & Ballard'91; Lazebnik, Schmid, Ponce'03; Sivic & Zisserman,'03; Csurka et al.'04; Zhang et al.'06)

Locally orderless image models



(Koenderink & Van Doorn'99; Dalal & Triggs'05; Lazebnik, Schmid, Ponce'06; Chum & Zisserman'07)

Image categorization as supervised classification

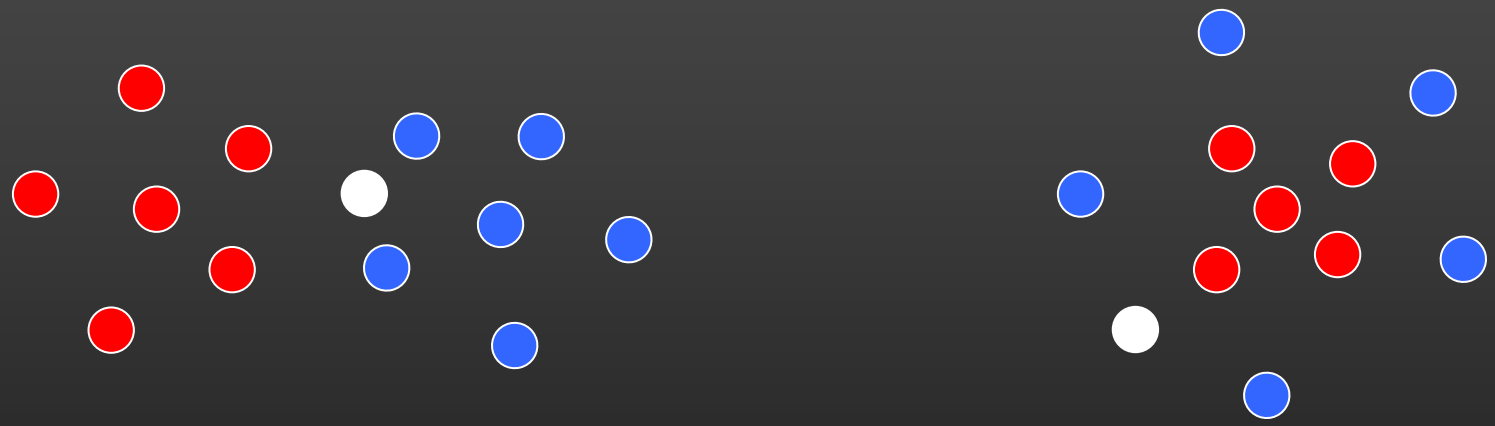
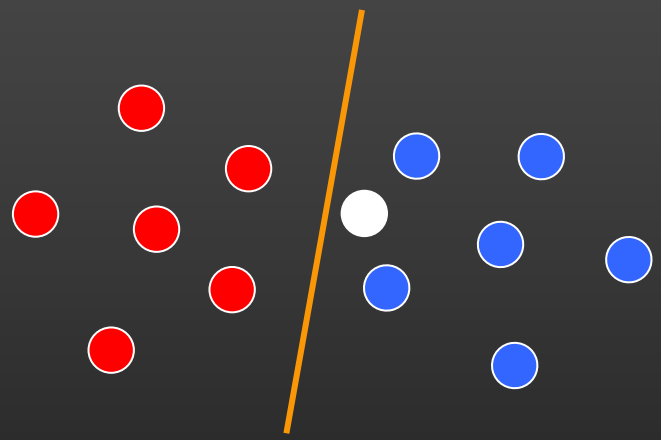


Image categorization as supervised classification



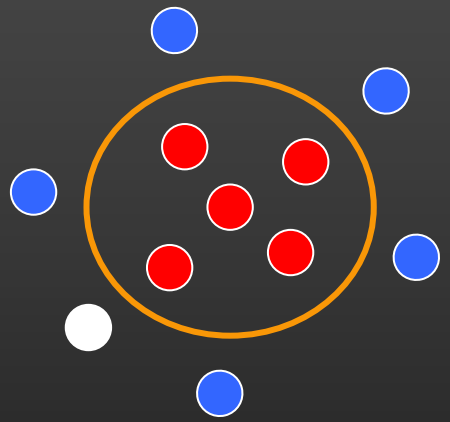


Image categorization as supervised classification

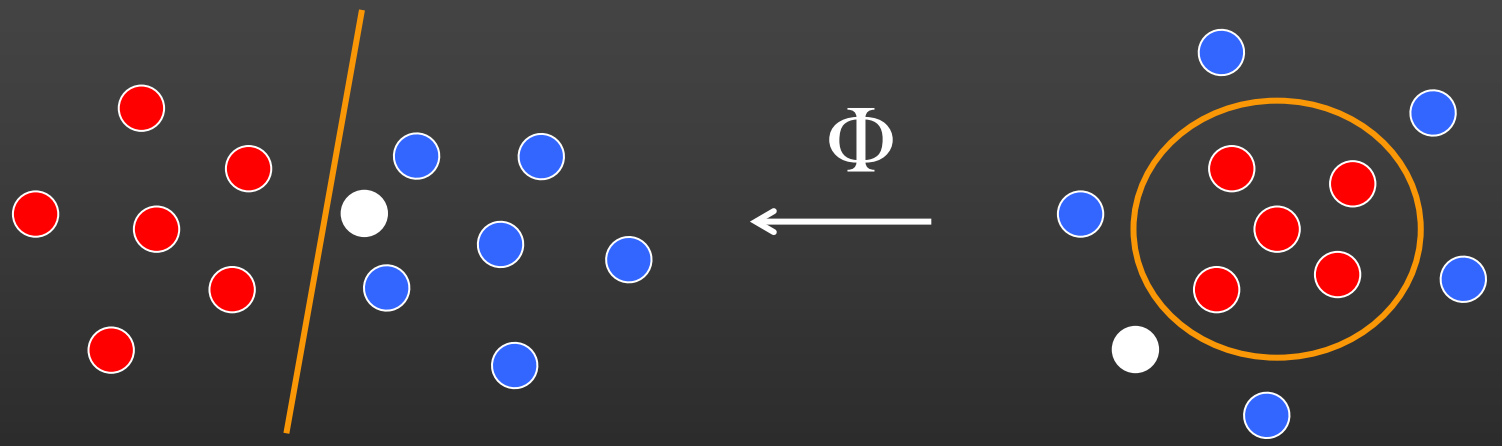
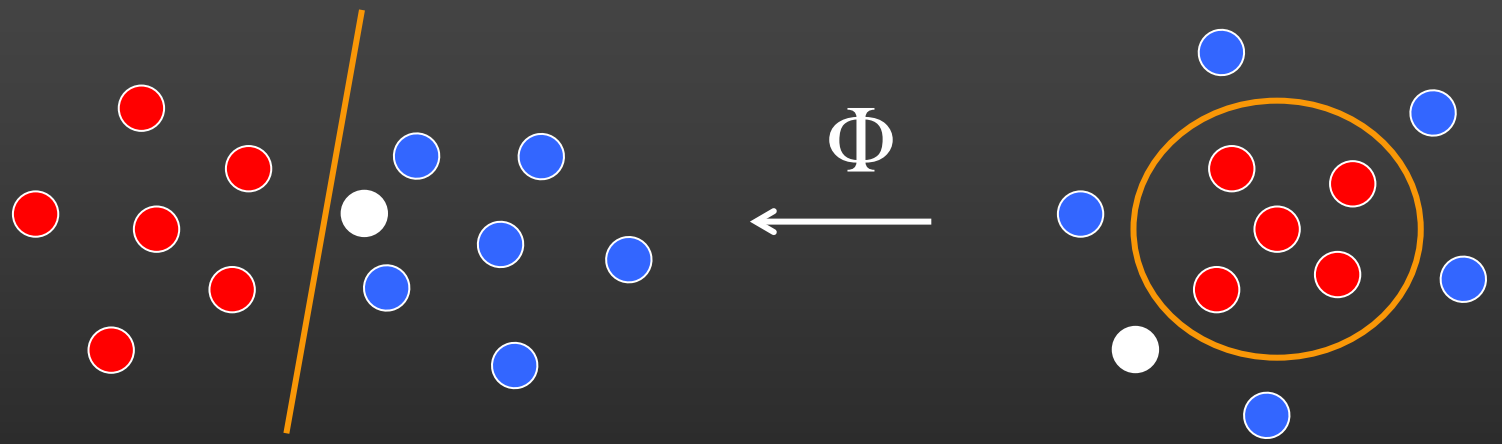


Image categorization as supervised classification

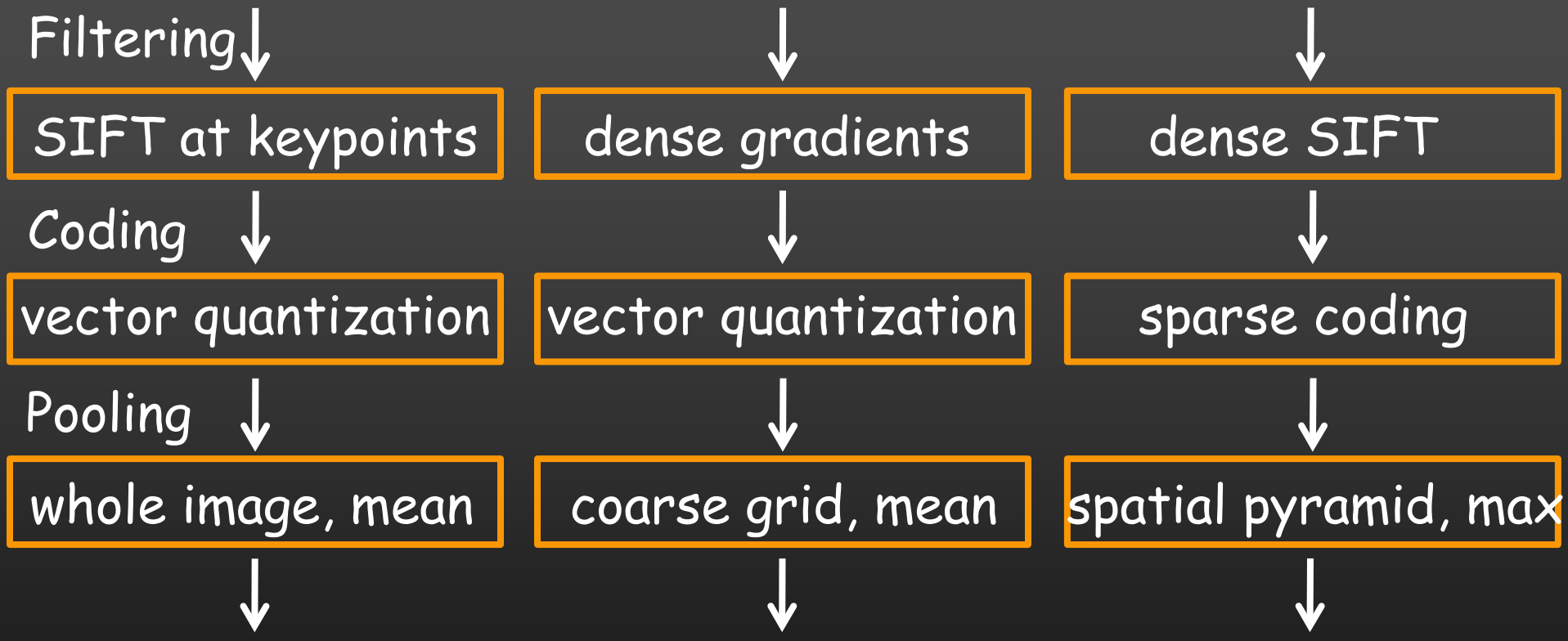


$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(y_i, f^\top \Phi(x_i)) + \frac{\lambda}{2}\|f\|^2$$

$$k(x, y) = \Phi(x) \cdot \Phi(y)$$

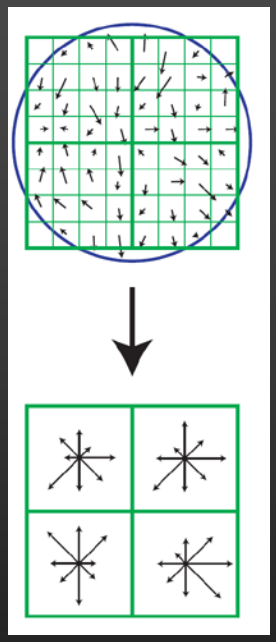
(Schölkopf & Smola, 2001; Shawe-Taylor & Cristianini, 2004; Wahba, 1990)

A common architecture for image classification



(Lowe'04, Csurka et al.'04, Dalal & Triggs'05)
(Yang et al.'09-10, Boureau et al.'10, Mallat'11)

A common architecture for image classification



SIFT

dense gradients

vector quantization

coarse grid, mean

HOG

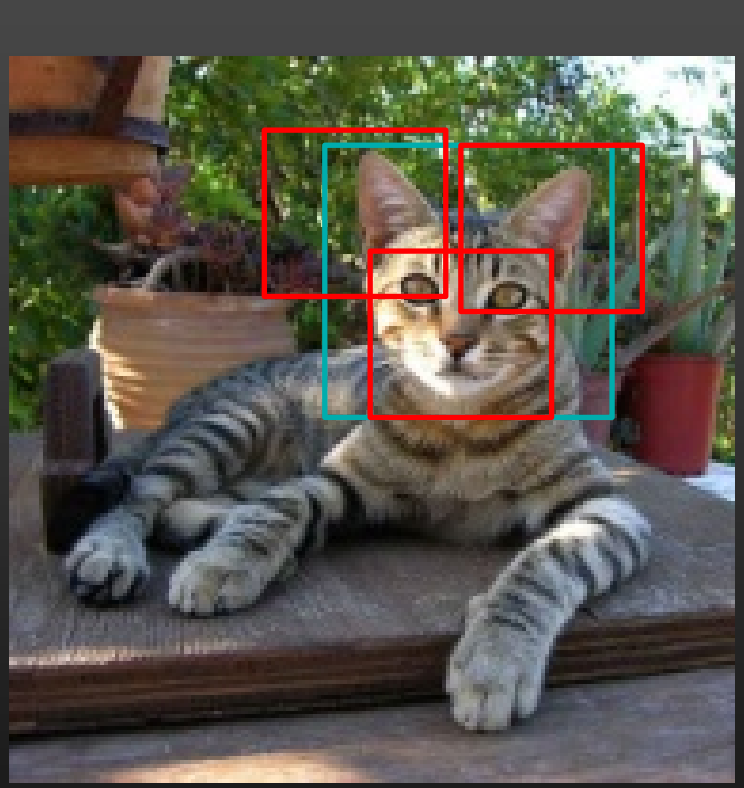
dense gradients

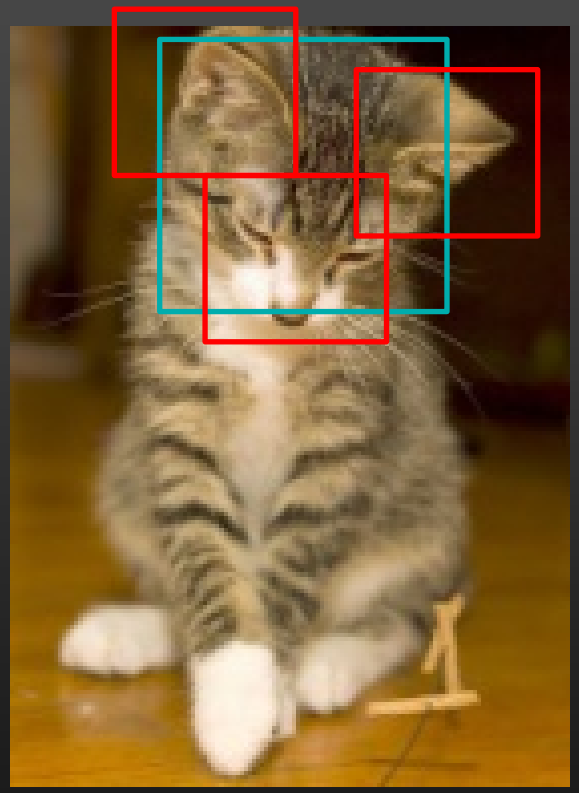
vector quantization

coarse grid, mean

(Lowe'04, Csurka et al.'04, Dalal & Triggs'05)
(Yang et al.'09-10, Boureau et al.'10, Mallat'11)

Object detection
(vue "d'artiste")



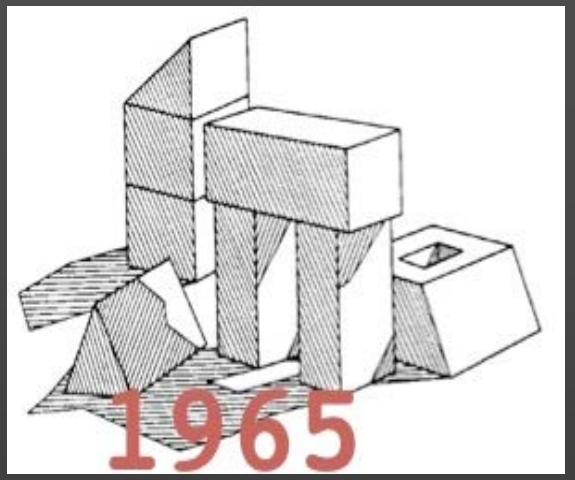


(A la Felzenszwalb, McAllester, Ramanan, 2008)

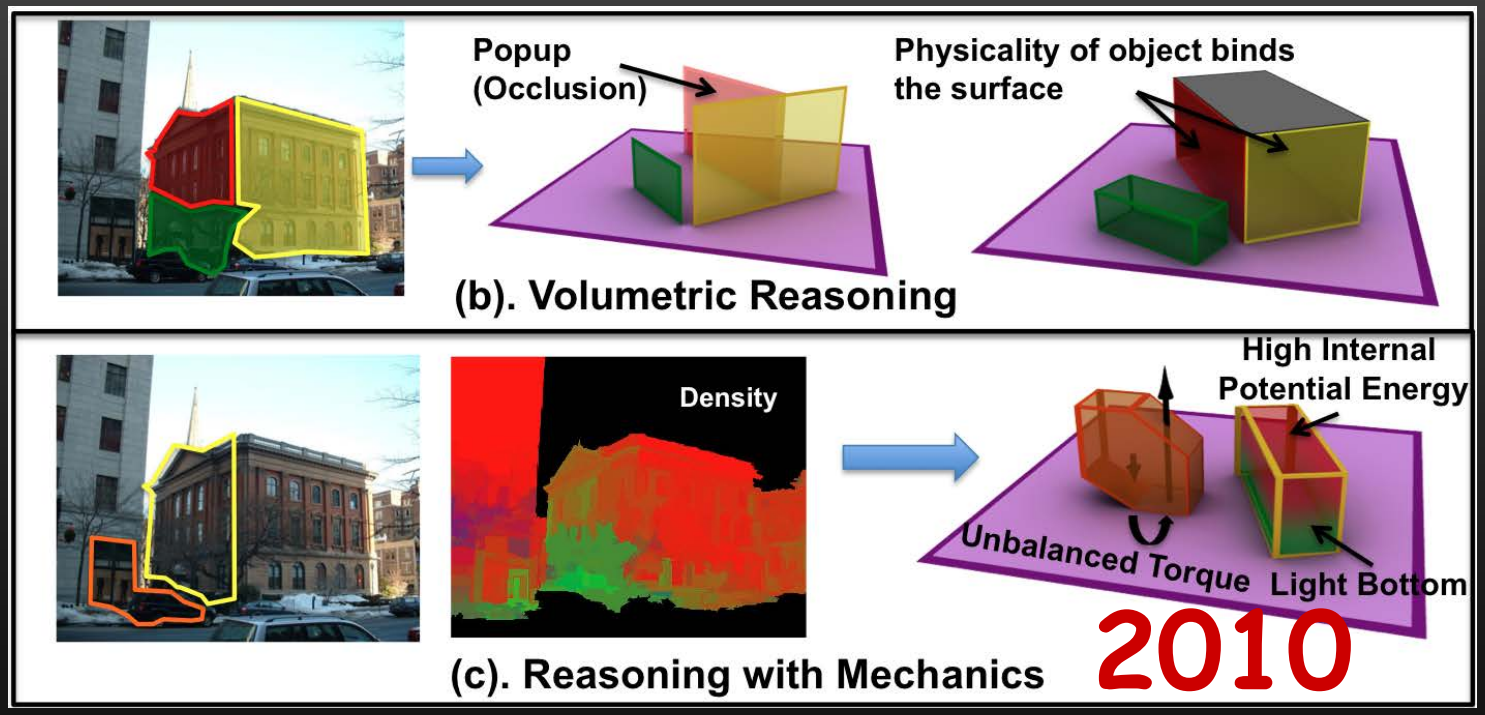
Sharing parts among aspects

(Kushal, Schmid, Ponce, CVPR'07)

What about scene understanding?



The blocks world revisited

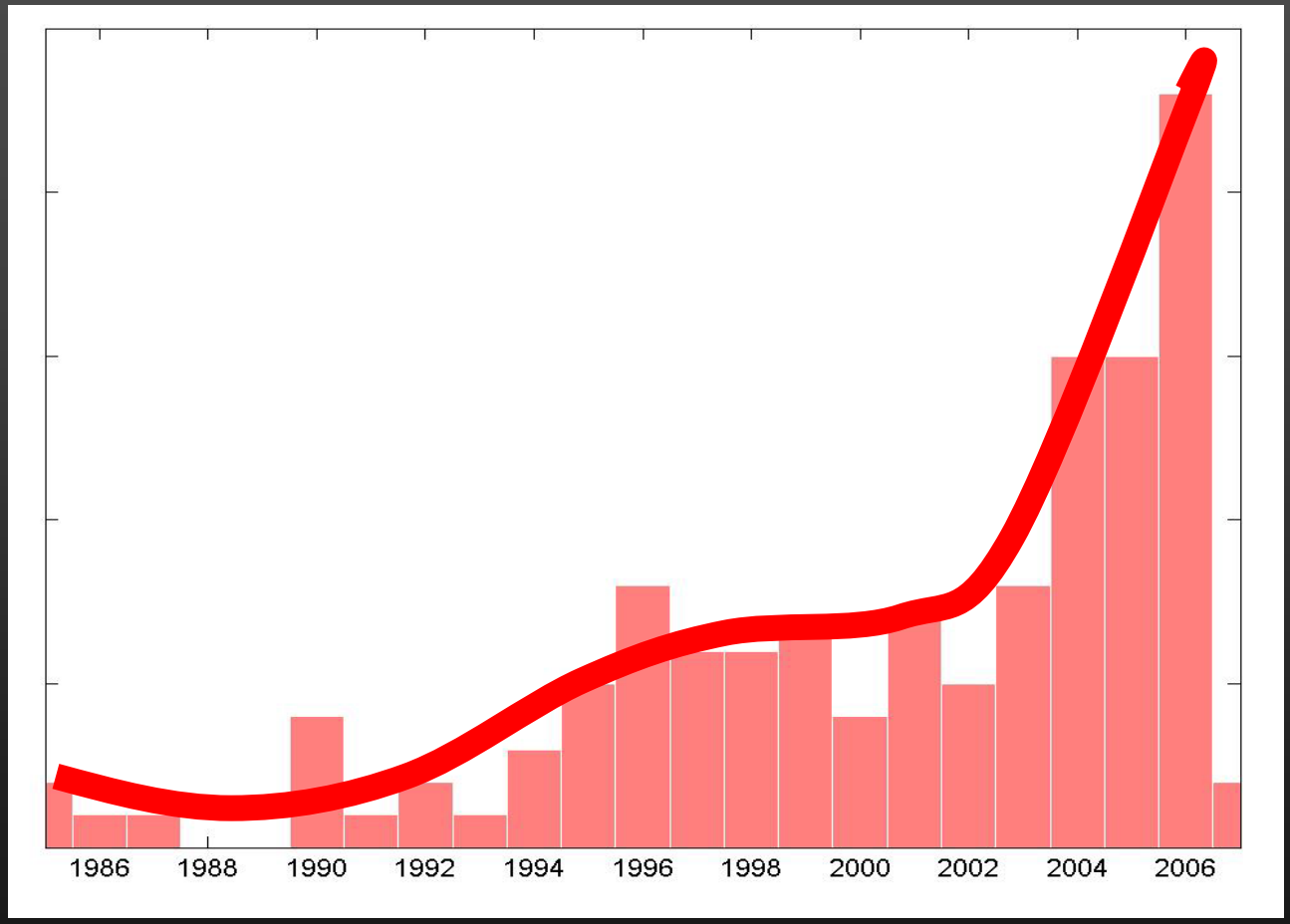


(Gupta, Efros, Hebert, ECCV'10)

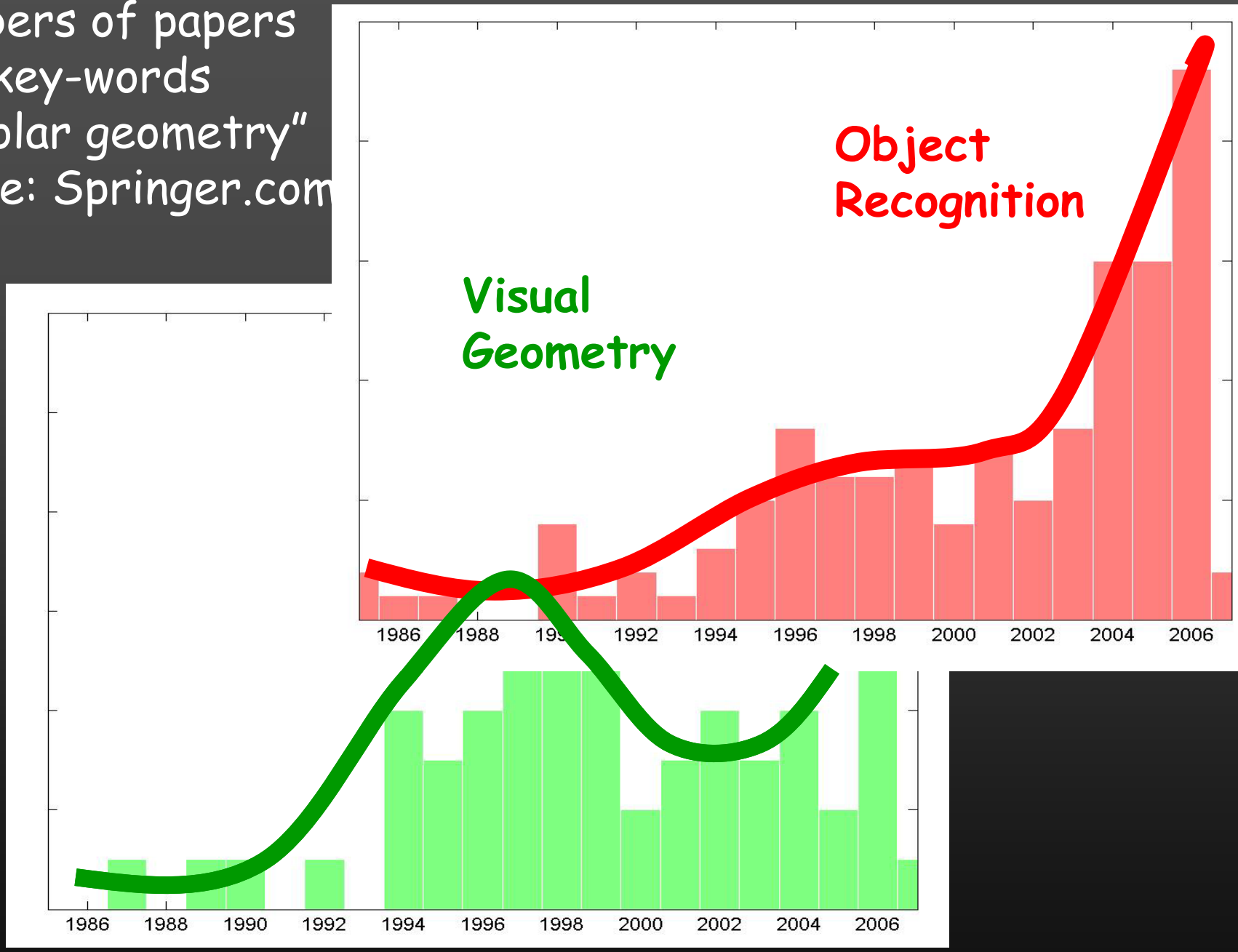
And video of course..

(Sivic, Everingham. Zisserman, CVPR'09)

Number of research papers with key-words "object recognition", source: Springer.com



Numbers of papers with key-words "epipolar geometry" source: Springer.com



Visual Geometry:

Problems: Camera calibration, 3D reconstruction,

S

T ching, …

Scale c.

Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry