High Level Computer Vision

Introduction - April 21, 2016

Bernt Schiele & Mario Fritz

Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany

mpi-inf.mpg.de/hlcv

Computer Vision and Multimodal Computing Group at the Max-Planck-Institute for Informatics

Bernt Schiele
Computer Vision

Bjoern Andres
Combinatorial Image Analysis

Andreas Bulling
Perceptual User Interfaces

Mario Fritz
Scalable Learning and Perception

Computer Vision

- Lecturer:
  - Bernt Schiele (schiele@mpi-inf.mpg.de)
  - Mario Fritz (mfritz@mpi-inf.mpg.de)
- Assistants:
  - Alina Dima (aldima@mpi-inf.mpg.de)
  - Seong Joon Oh (joon@mpi-inf.mpg.de)
  - Rakshith Shetty (rshetty@mpi-inf.mpg.de

- Language:
  - English
- mailing list for announcements etc.
  - send email to Joon (see instructions on the web)

Lecture & Exercise

- Officially: 2V (lecture) + 2Ü (exercise)
  - Lecture: Wed: 2:15pm - 4pm (room 024)
  - Exercise: Mon: :2:15pm - 4pm (room 024)

- typically 1 exercise sheet every 1-2 weeks
  - part of the final grade
  - pencil and paper, as well as matlab-based exercise,
  - reading assignment (research papers, overview papers, etc.)
- & larger project at end of lecture
  - we/you propose project, mentoring, final presentation
- 1. exercise is matlab tutorial

- Exam
  - oral exam
  - after the SS - there will be proposed dates

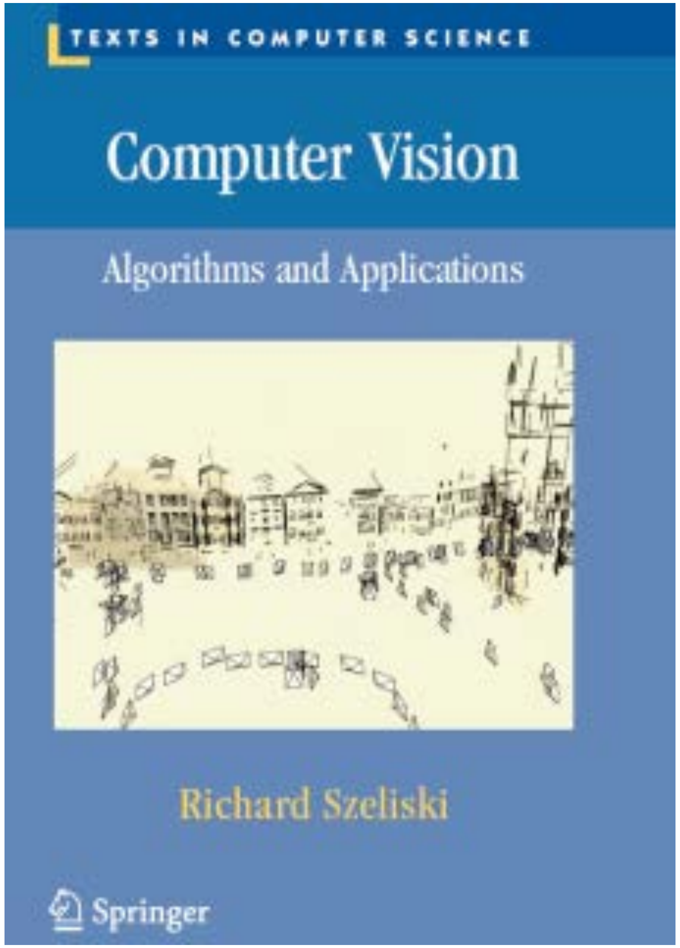
Grading

- 50% oral exam
- 50% exercises

- exercises
  - 2/3 regular exercises
  - 1/3 project

Material

- For part of the lecture: http://szeliski.org/Book/
- available online



Material

- Additional background on deep learning: Deep Learning Book
  - available online deeplearningbook.org (in preparation)

Deep Learning

An MIT Press book

Ian Goodfellow, Yoshua Bengio and Aaron Courville

Exercises Lecture Slides

The Deep Learning textbook is a resource intended to help students and practitioners enter the field of machine learning in general and deep learning in particular. The online version of the book is now complete and will remain available online for free. The print version will be available for sale soon. For up to date announcements, join our mailing list.

Citing the book

To cite this book, please use this bibtex entry:

```
@unpublished{Goodfellow-et-al-2016-Book,
    title={Deep Learning},
    author={Ian Goodfellow, Yoshua Bengio, and Aaron Courville},
    note={Book in preparation for MIT Press},
    url={http://www.deeplearningbook.org},
    year={2016}
}
```

Why Study Computer Vision

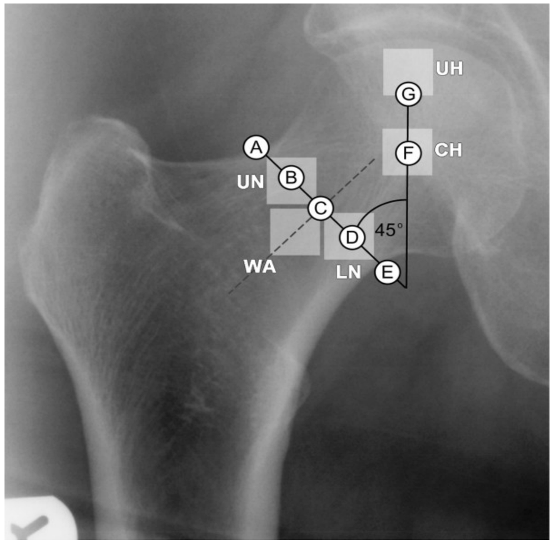
- Science
  - Foundations of perception. How do WE see?
  - computer vision to explore “computational model of human vision”

- Engineering
  - How do we build systems that perceive the world
  - computer vision to solve real-world problems
    (e.g. self-driving cars to detect pedestrians)

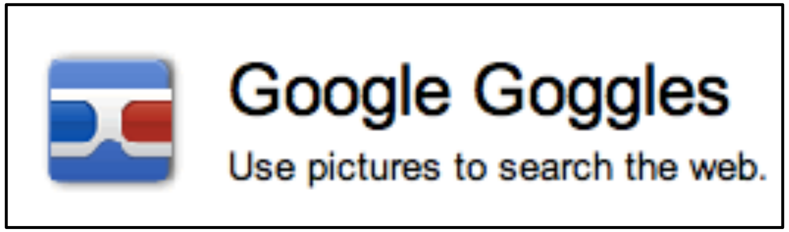
- Applications
  - medical imaging (computer vision to support medical diagnosis, visualization)
  - surveillance (to follow/track people at the airport, train-station, ...)
  - entertainment (vision-based interfaces for games)
  - graphics (image-based rendering, vision to support realistic graphics)
  - car-industry (lane-keeping, pre-crash intervention, …)

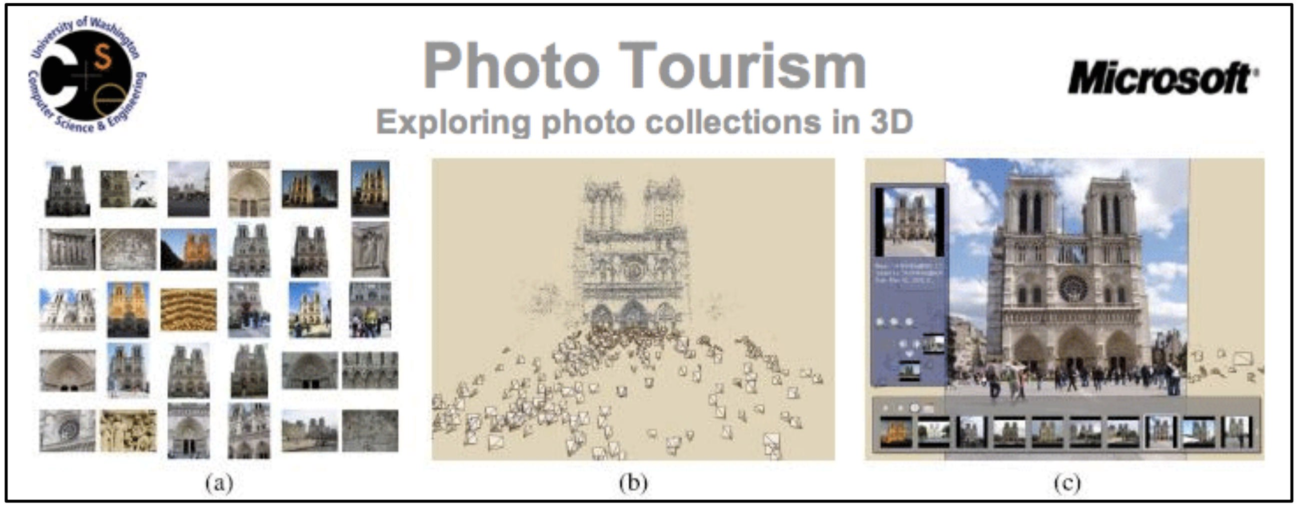
Some Applications

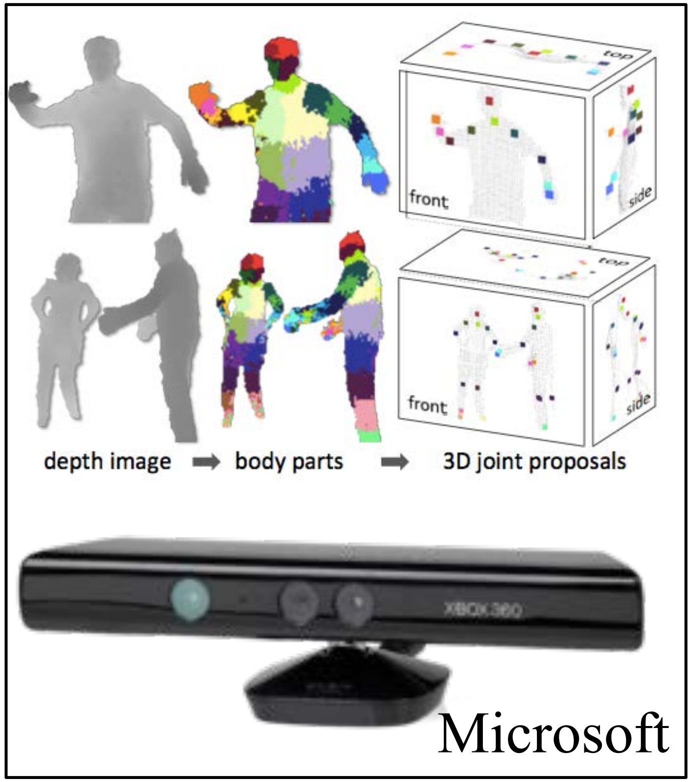
- License Plate Recognition
  - London Congestion Charge
  - http://www.cclondon.com/imagingandcameras.html
  - http://en.wikipedia.org/wiki/London_congestion_charge
- Surveillance
  - Face Recognition
  - Airport Security (People Tracking)
- Medical Imaging
  - (Semi-)automatic segmentation and measurements
- Robotics
- Driver assistance



More Applications







Goals of today's lecture

- First intuitions about
  - What is computer vision?
  - What does it mean to see and how do we (as humans) do it?
  - How can we make this computational?

- Applications & Appetizers

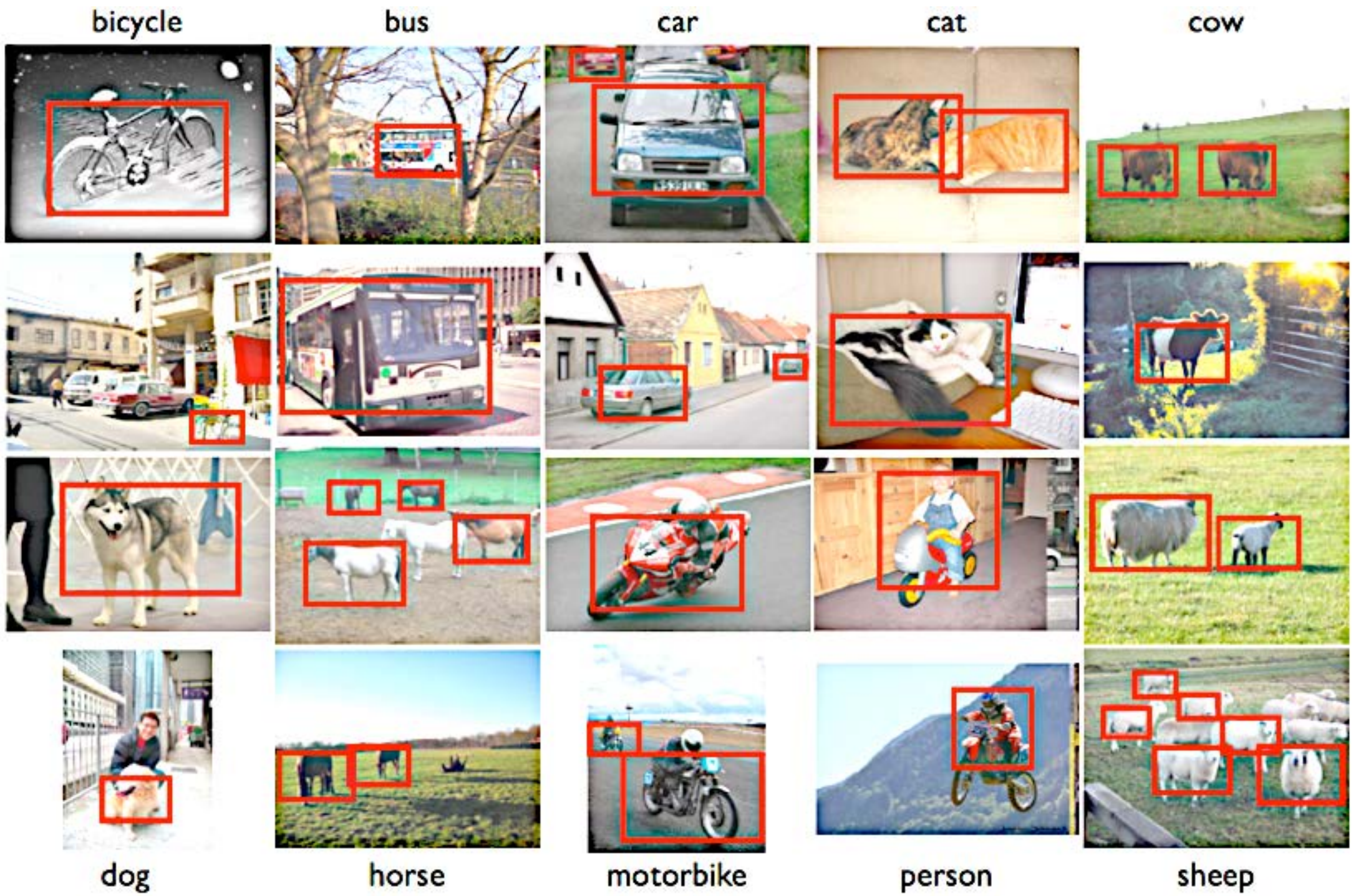
- 2 case studies:
  - Recovery of 3D structure
    - slides taken from Michael Black @ Brown University / MPI Intelligent Systems

  - Object Recognition
    - intuition from human vision...

Applications & Appetizers

... work from our group

Detection & Recognition of Visual Categories

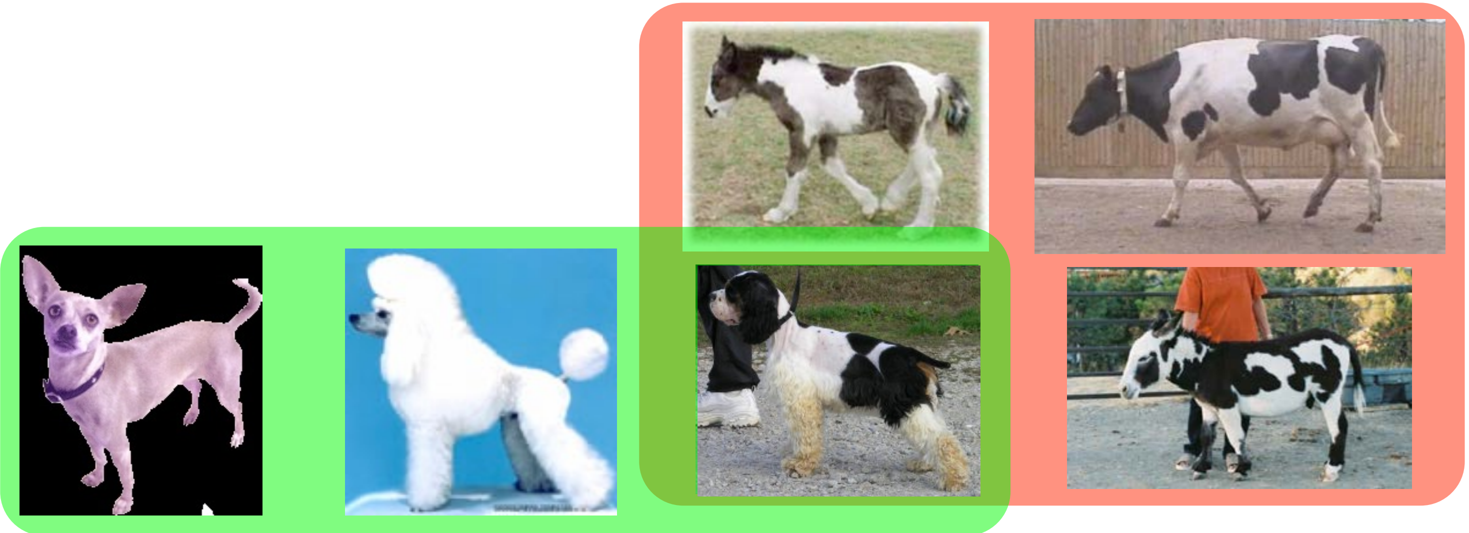


Challenges:

- multi-scale
- multi-view
- multi-class
- varying illumination
- occlusion
- cluttered background
- articulation
- high intraclass variance
- low interclass variance

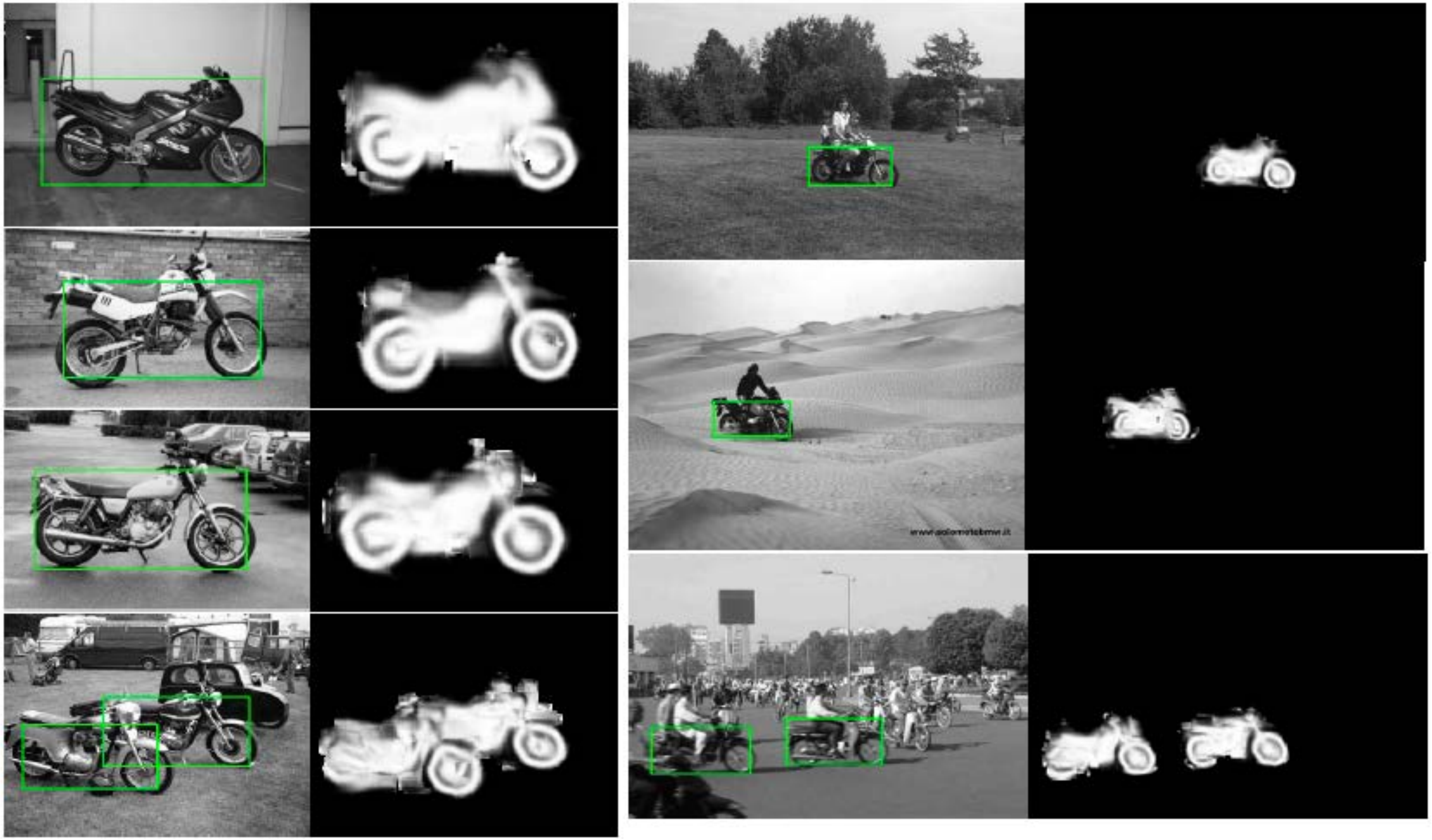
Challenges of Visual Categorization

- low inter-class variation

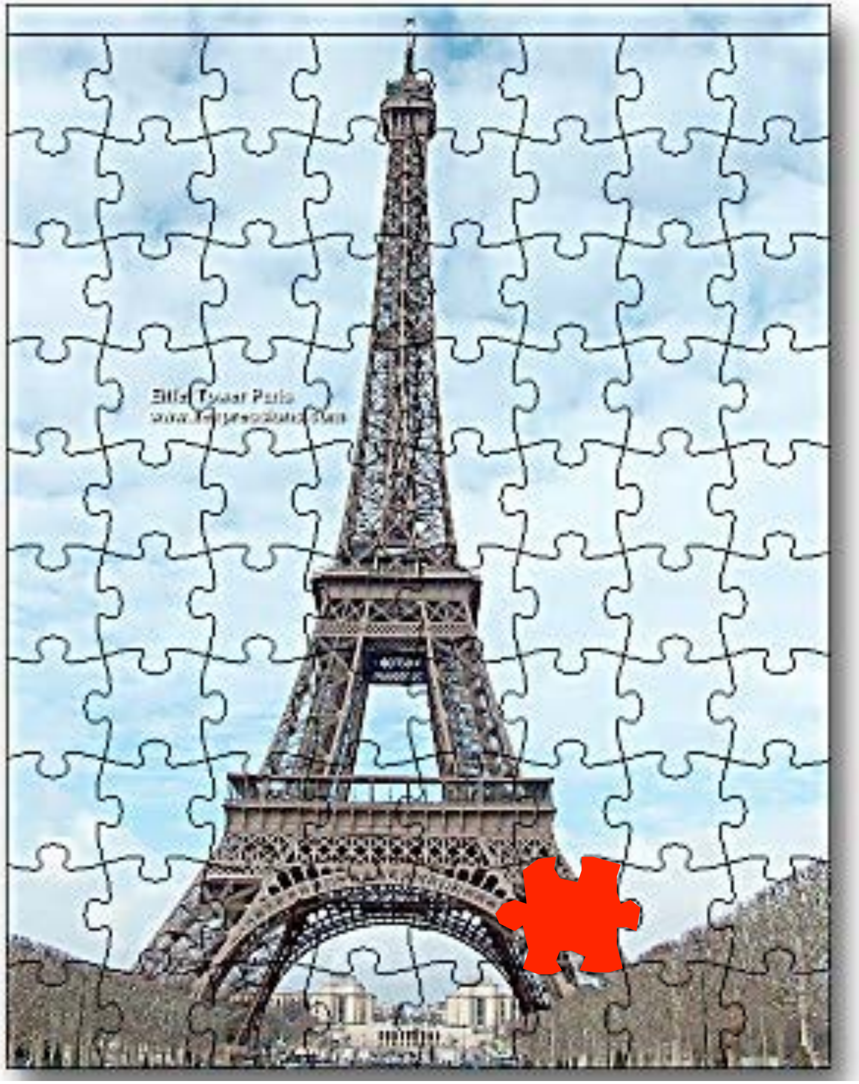


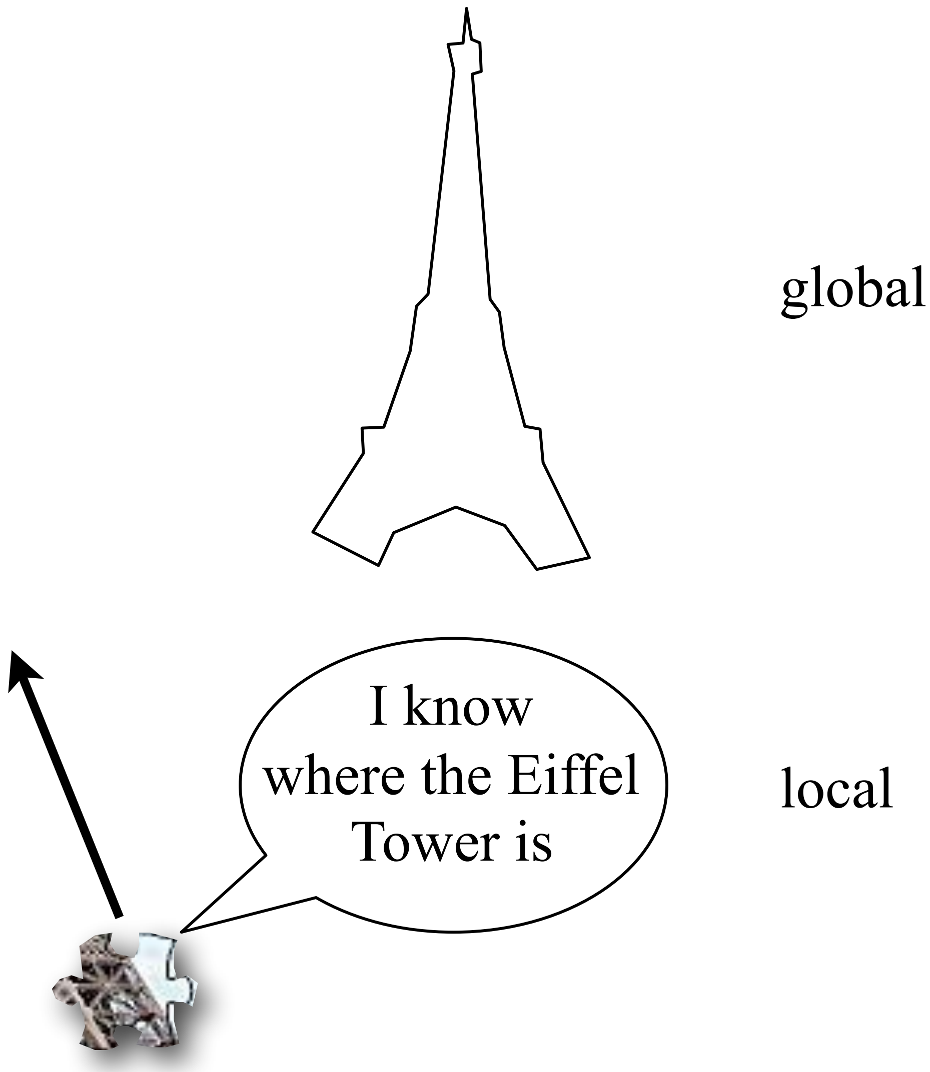
- high intra-class variation

Sample Category: Motorbikes



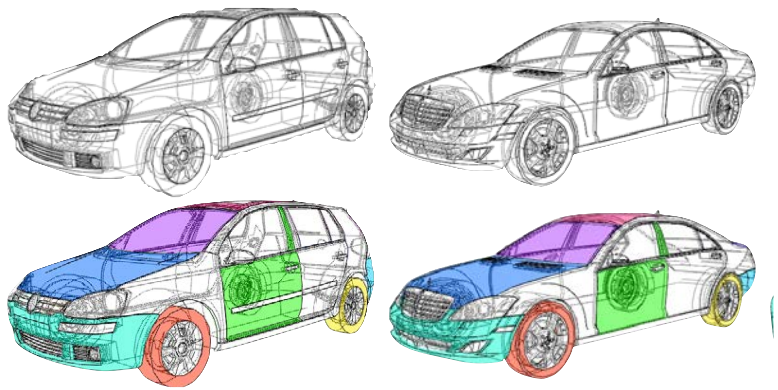
Basic Idea



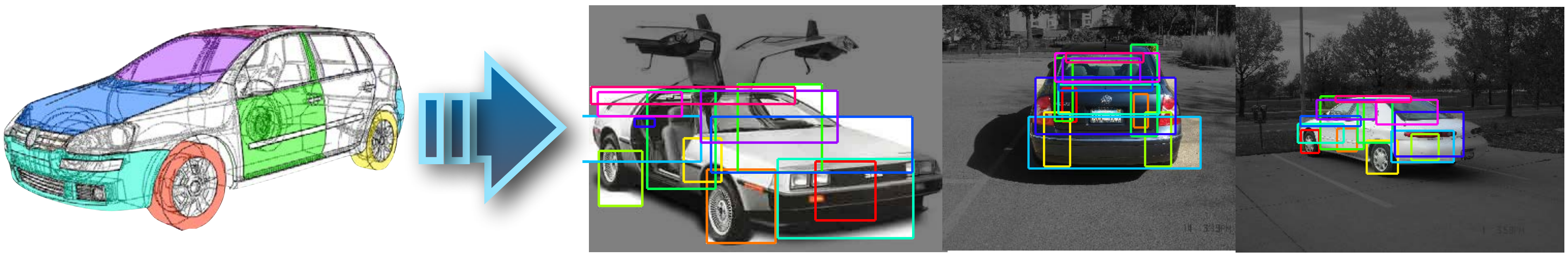


Large Scale Object Class Recognition

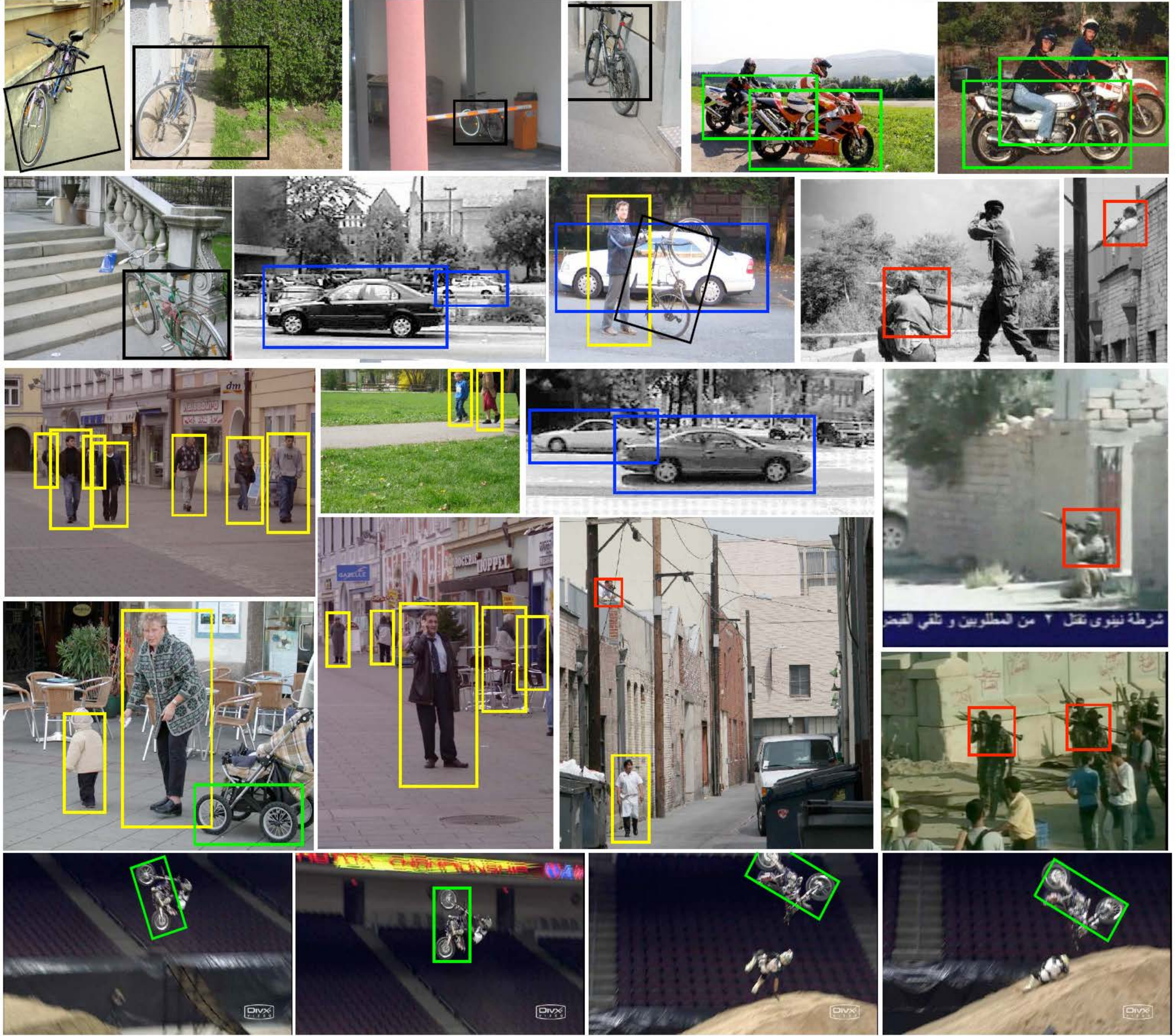
- Learning Shape Models from 3D CAD Data
  - 3D Computer Aided Design (CAD) Models for
    - computer graphics, game design
    - polygonal meshes + texture descriptions
    - semantic part annotations (may) exist



  - Learning Object Class Model directly from 3D CAD-data:



Michael Stark

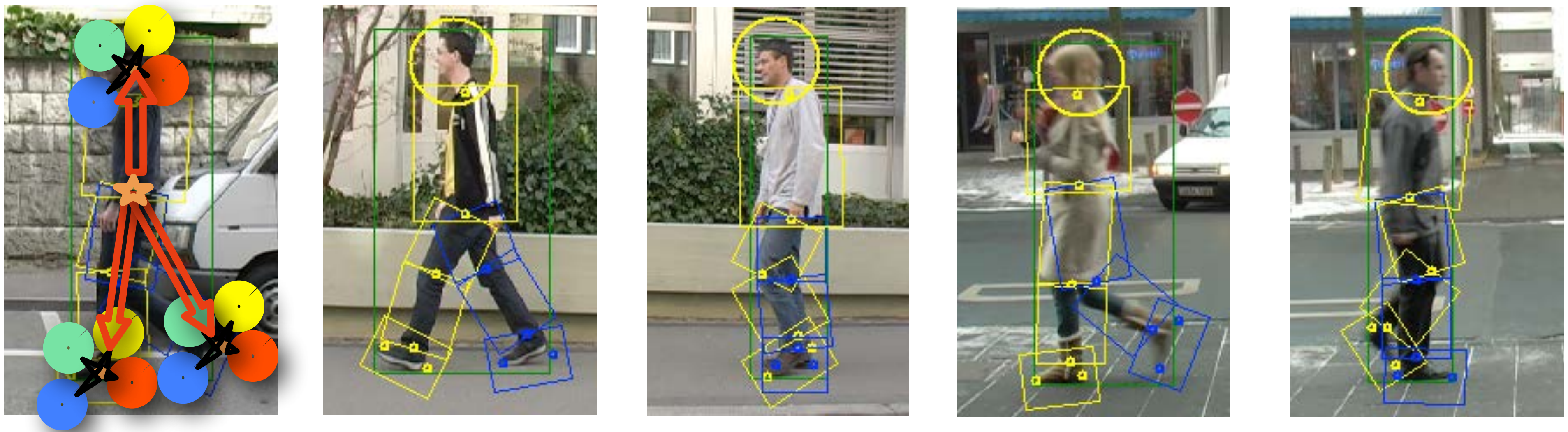

dm
شرطة نينوى تقتل ٢ من المطلوبين و تلقي القبض

Video...

Articulation Model

$$p(L|a) = p(\mathbf{x}^o) \prod_{i=1}^{N} p(\mathbf{x}^i|\mathbf{x}^o, a)$$

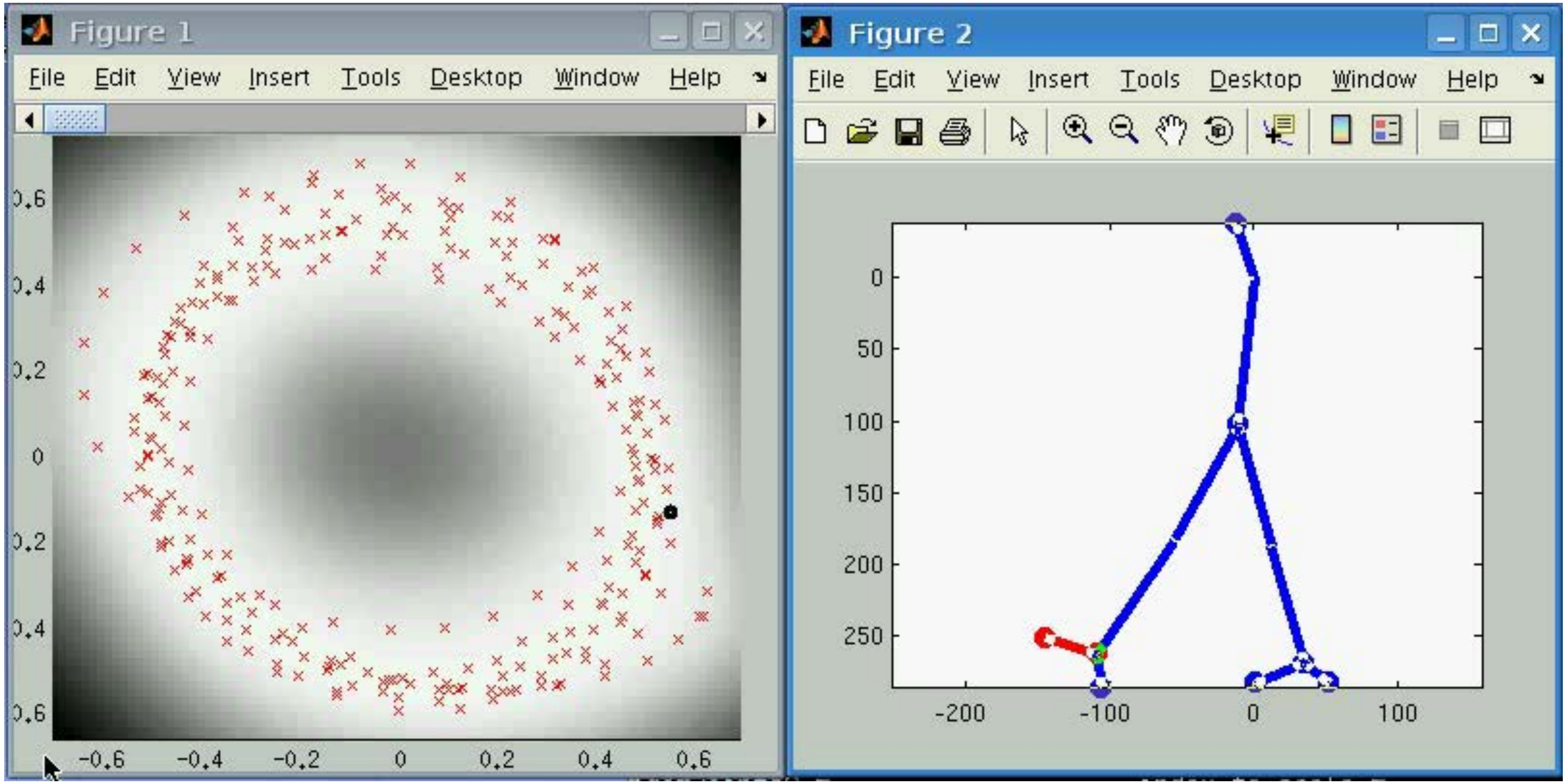
- Assume uniform position prior for the whole body
- Learn the conditional relation between part position and body center from data:

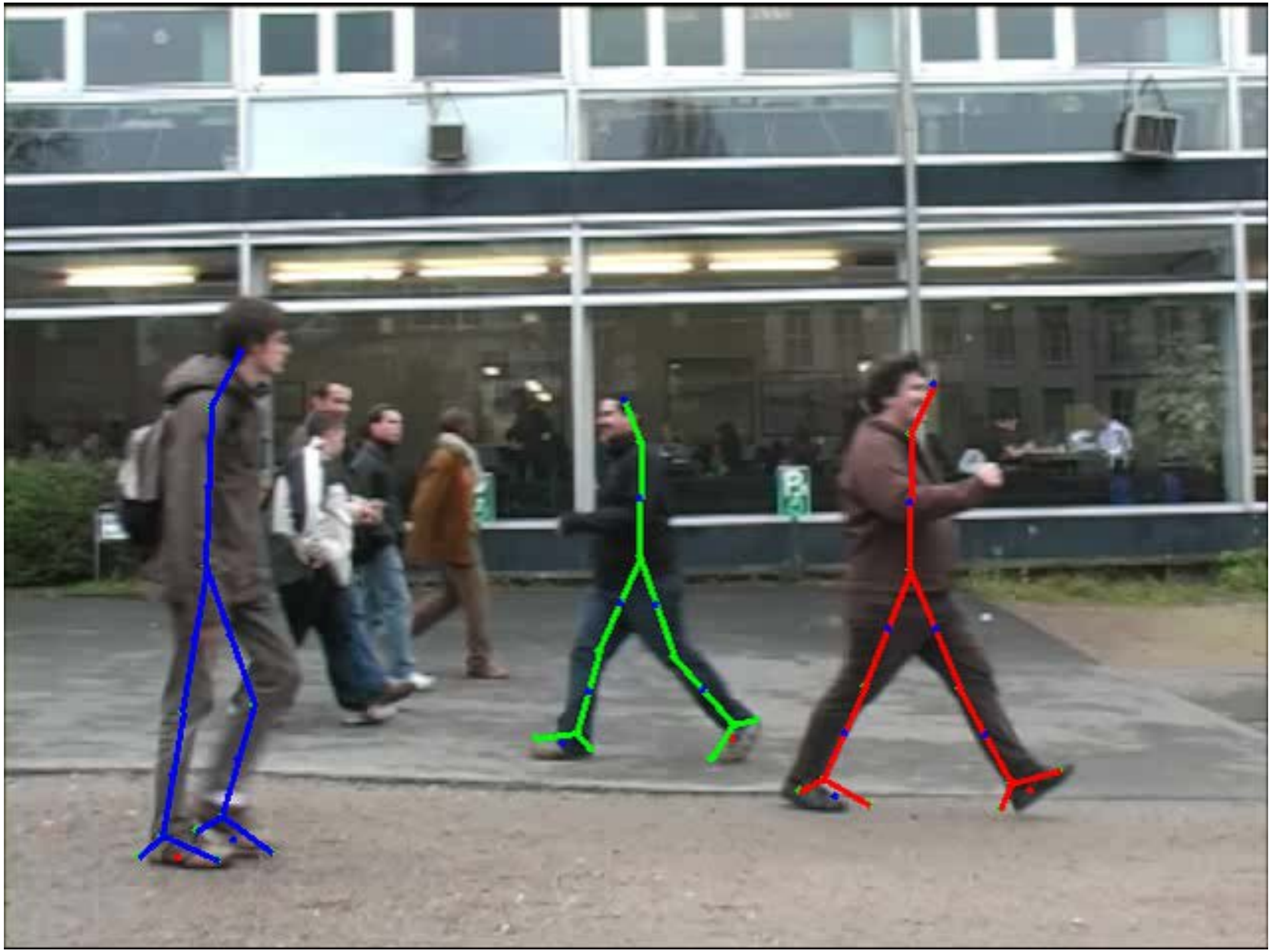


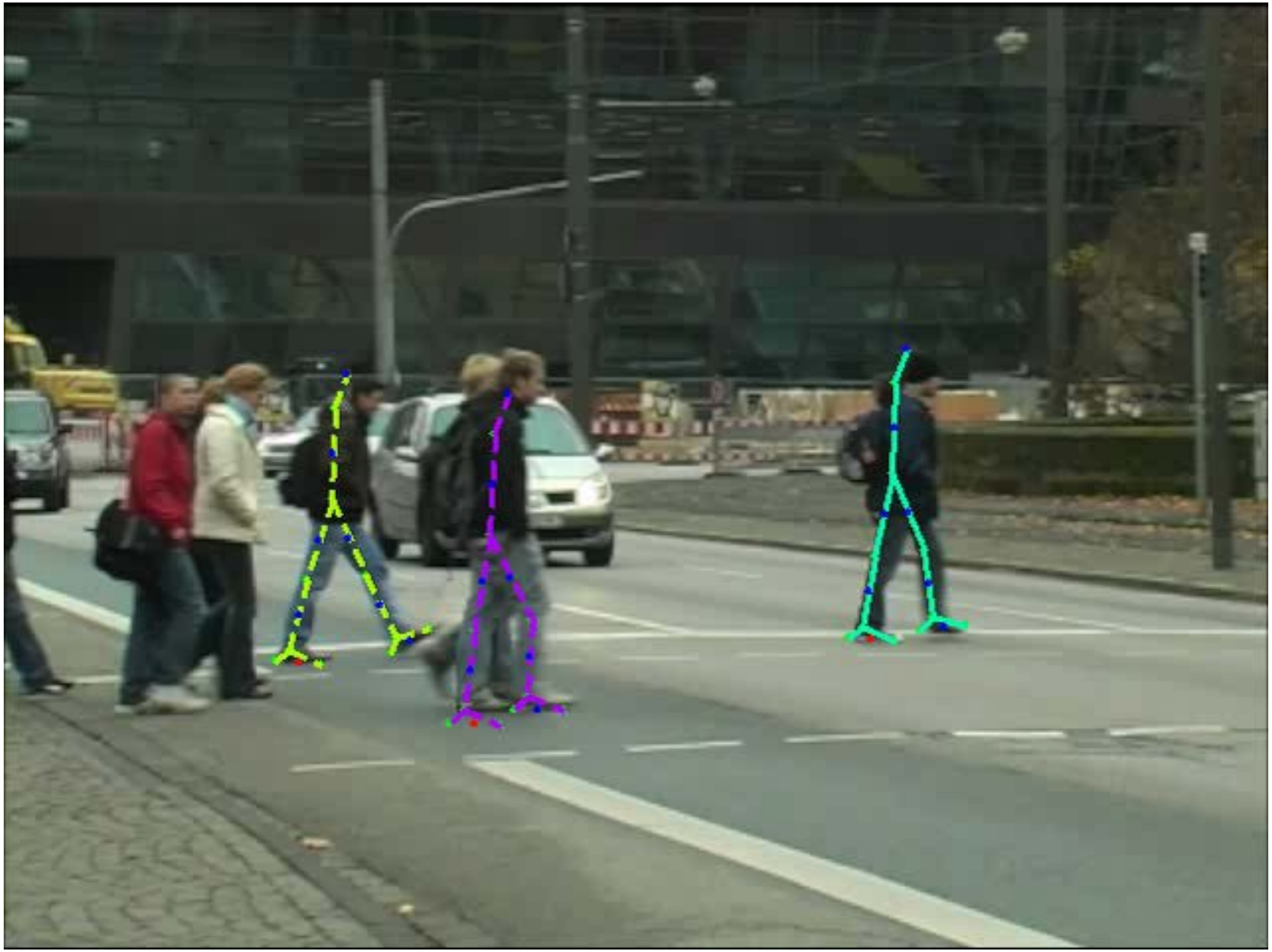
400 annotated training images

Modeling Body Dynamics

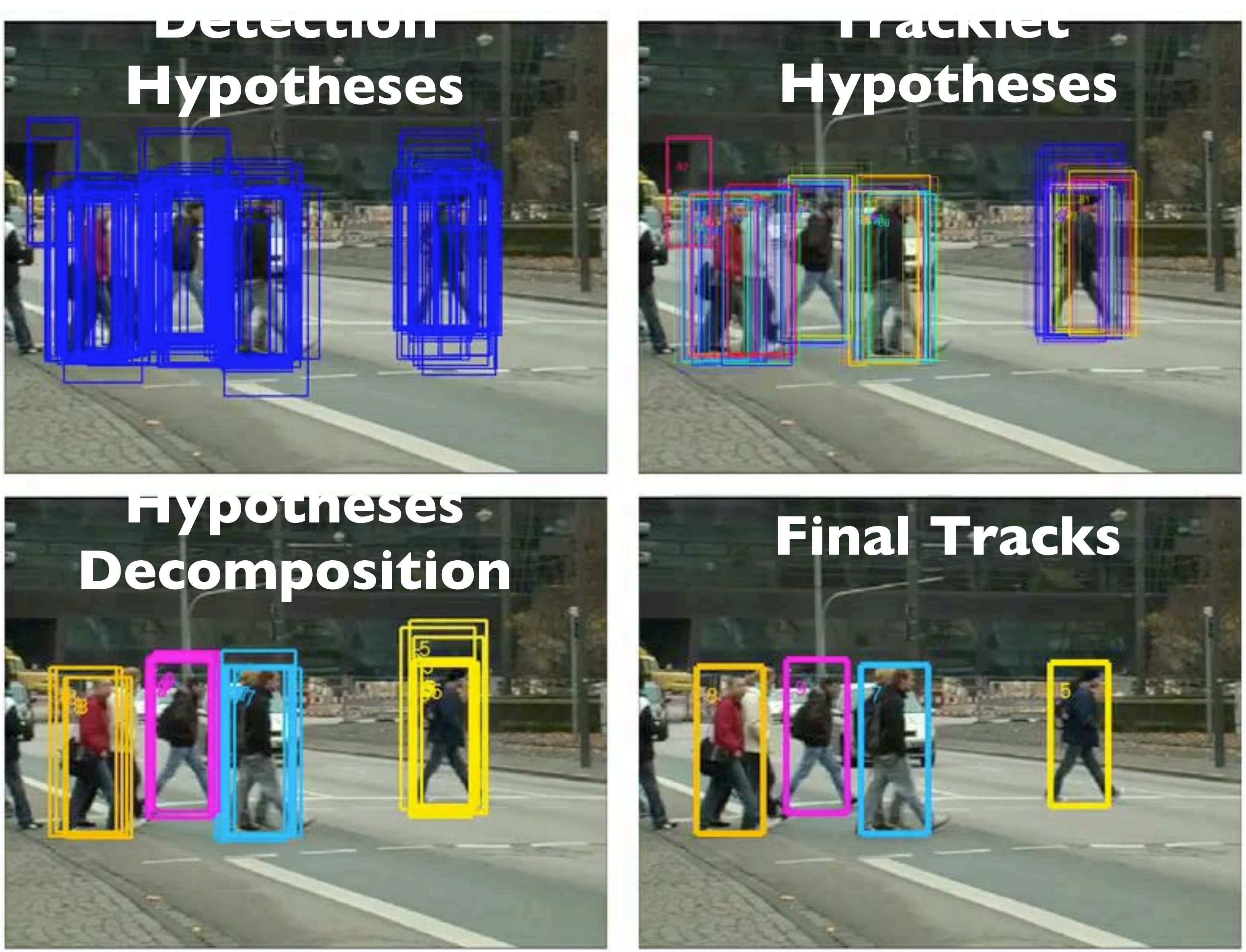
- Visualization of the hierarchical Gaussian process latent variable model (hGPLVM)







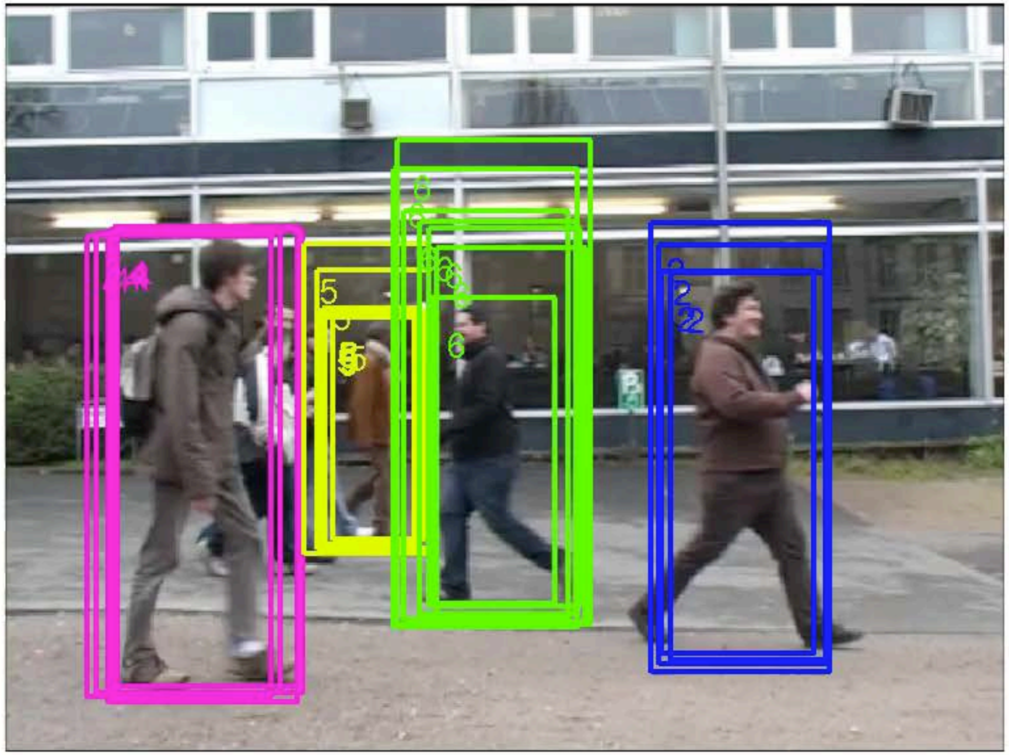
Our Subgraph Multicut Tracking Results



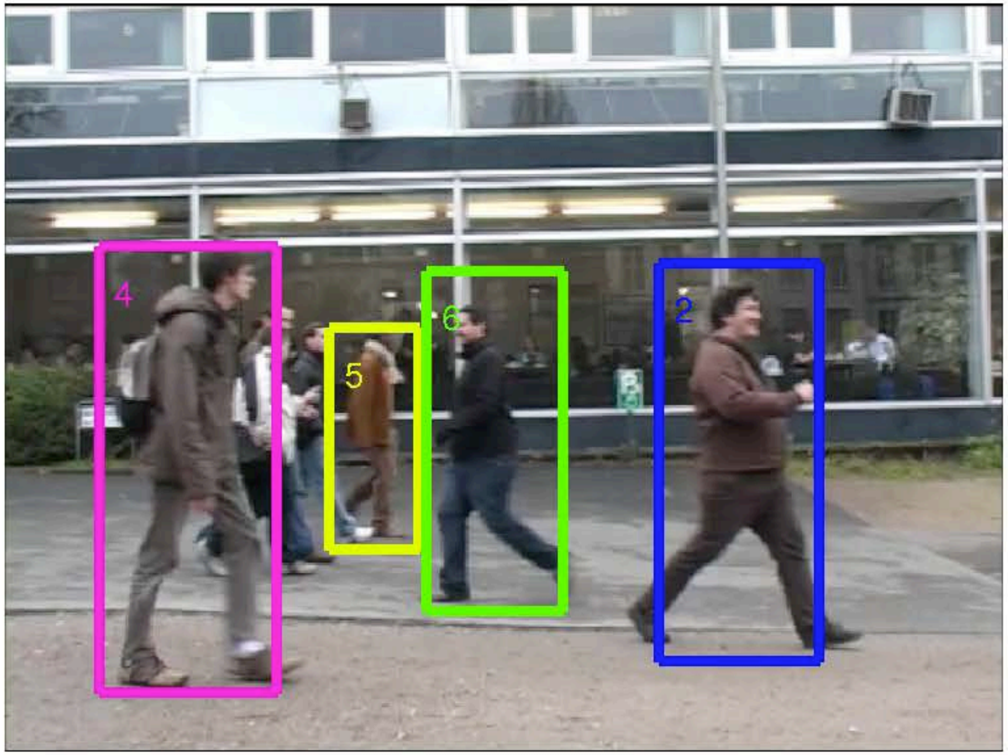
Dotted rectangles are interpolated tracks.

More Results

Dotted rectangles are interpolated tracks.



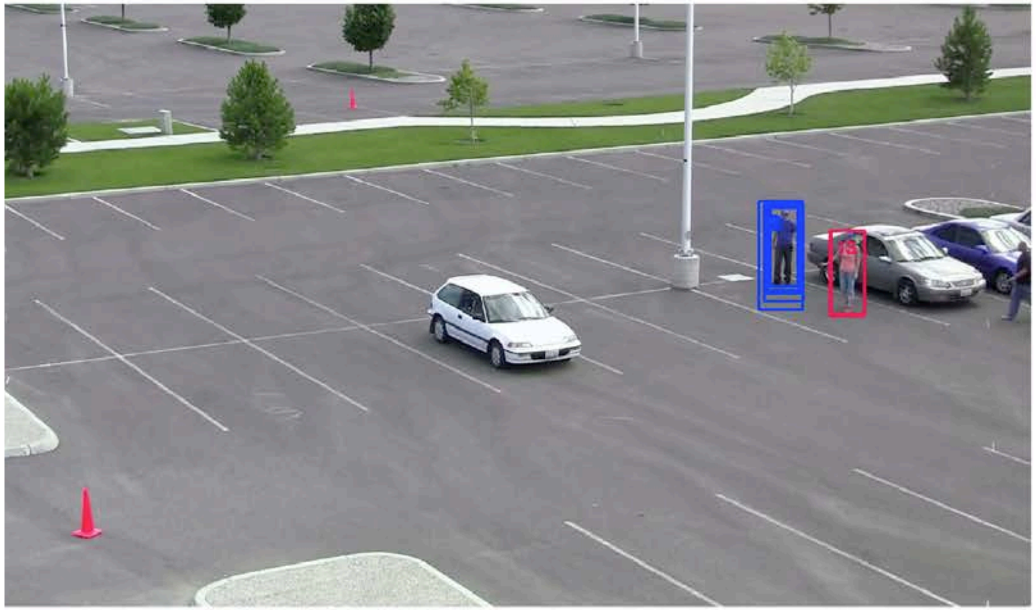
Decompositions (clusters)



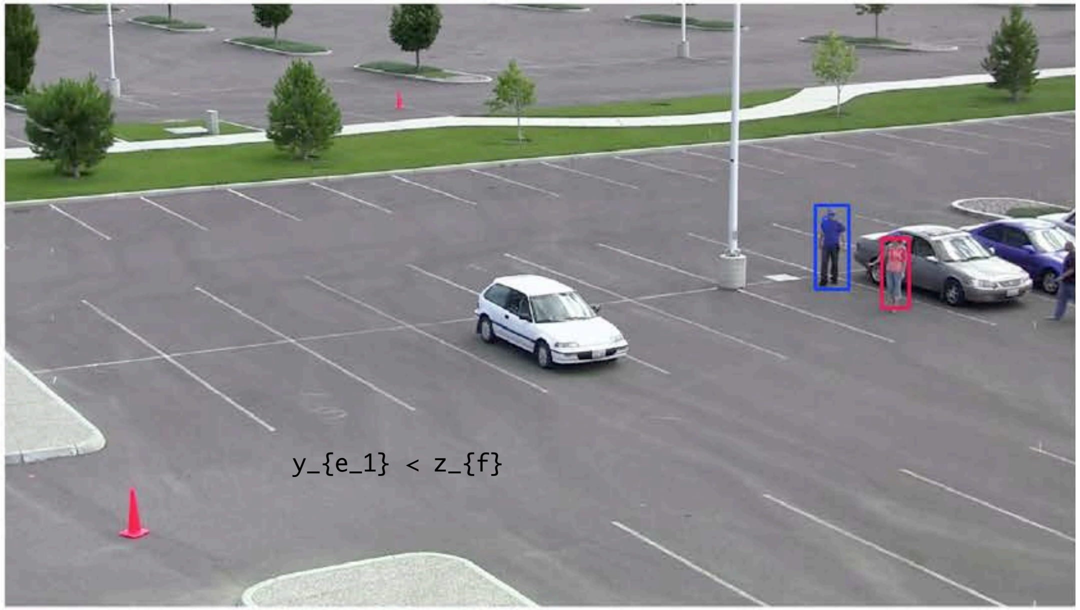
Tracks

More Results

Dotted rectangles are interpolated tracks.



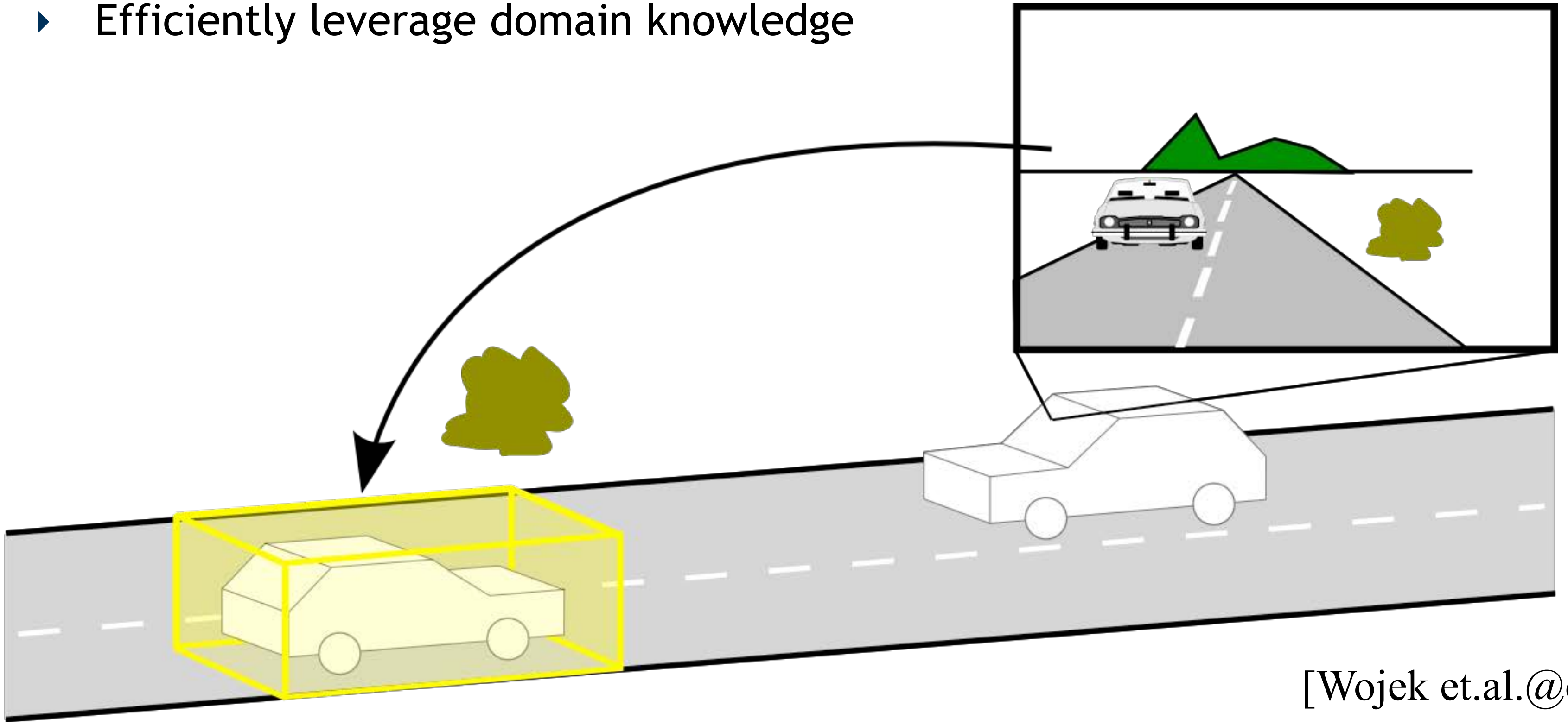
Decompositions (clusters)



Tracks

Complete 3D Scene Modeling

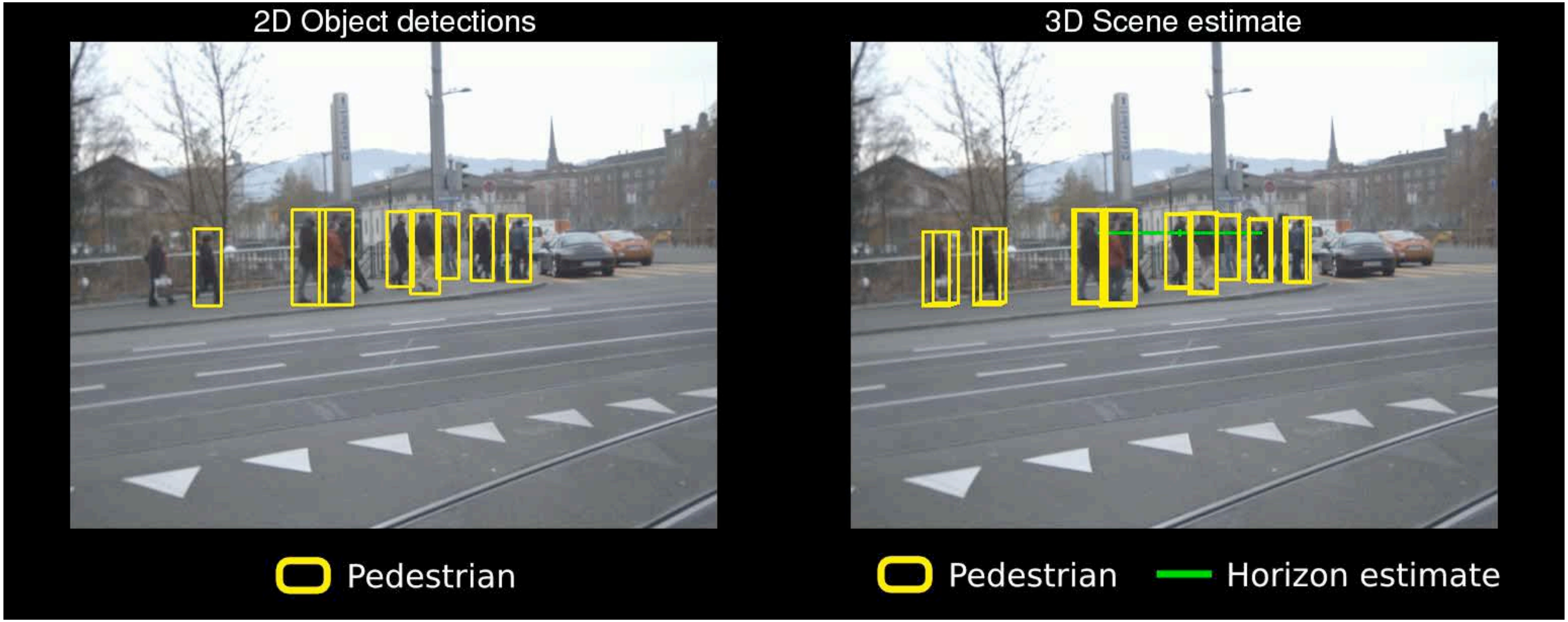
- Goal: Infer consistent 3D world hypothesis from 2D image sequences with a moving monocular camera
  - Tracking 3D Scene Model
  - Integrate SoA object detectors, scene labeling
  - Efficiently leverage domain knowledge



[Wojek et.al.@eccv10]

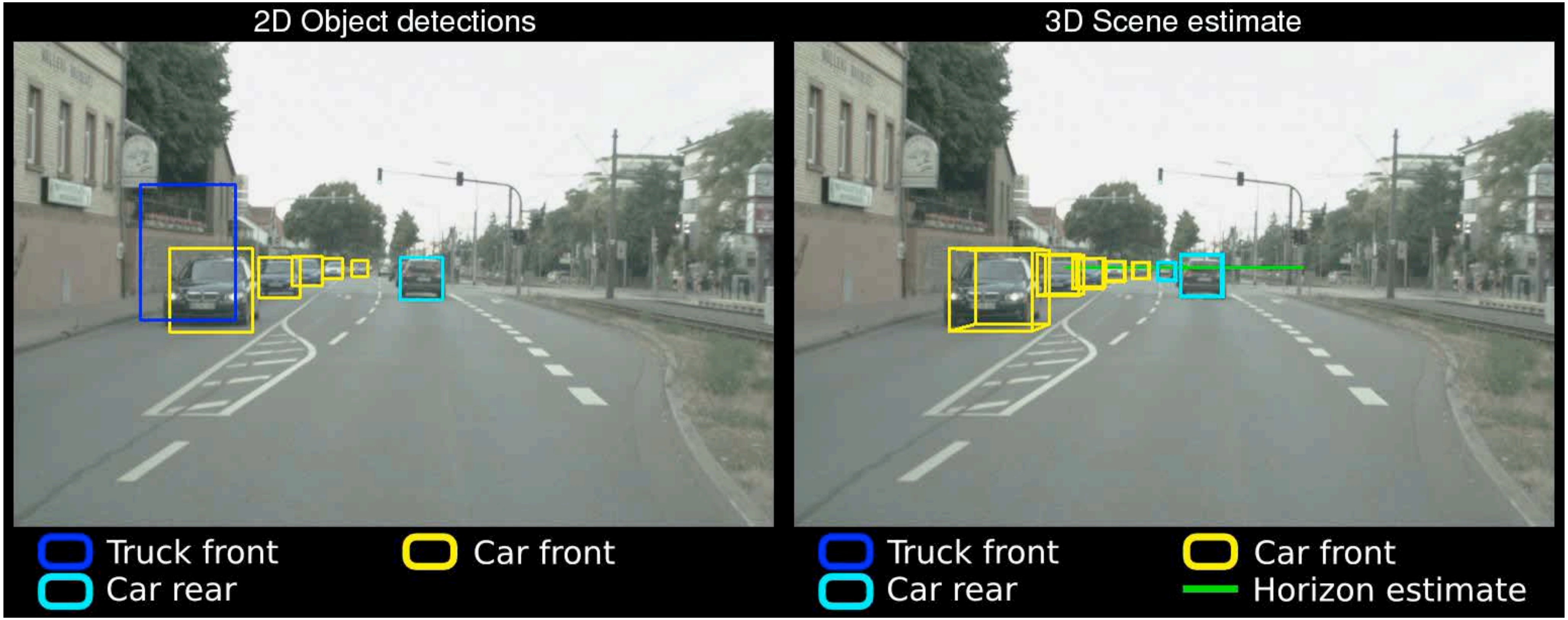
[Wojek et.al.@eccv10]

System sample video (pedestrians)



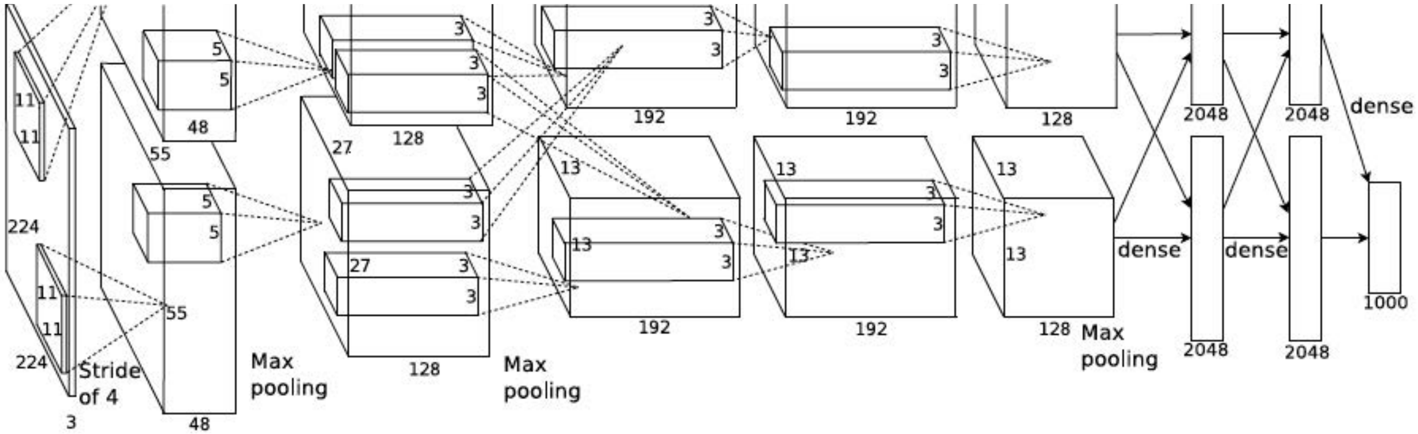
ETH-Loewenplatz sequence: By courtesy of ETH Zürich [Ess et al., PAMI '09]

System sample video (vehicles)



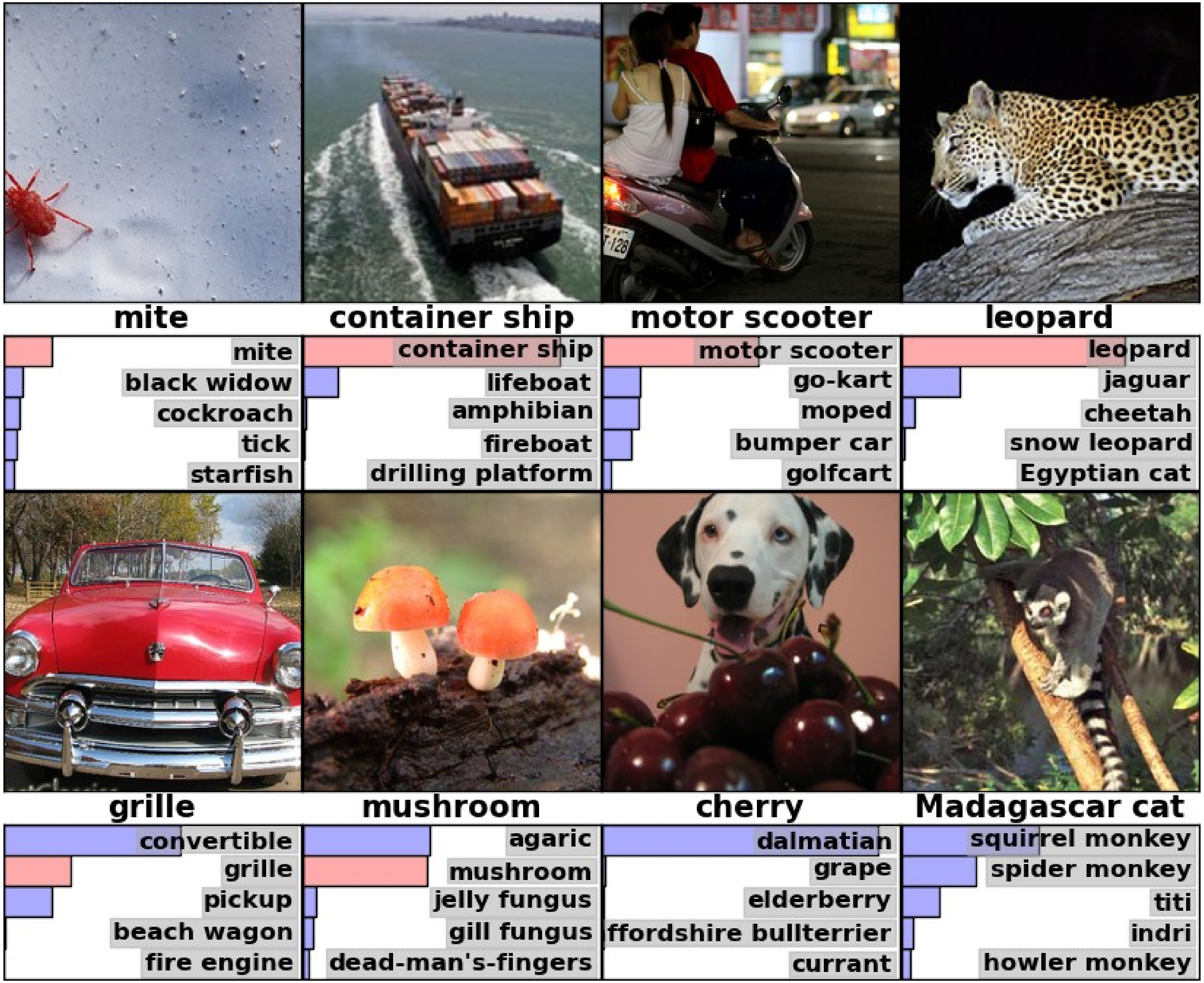
Deep Neural Networks

- Same model as LeCun’98 — BUT
  - Bigger model (8 layers)
  - More data (106 vs 103 images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)

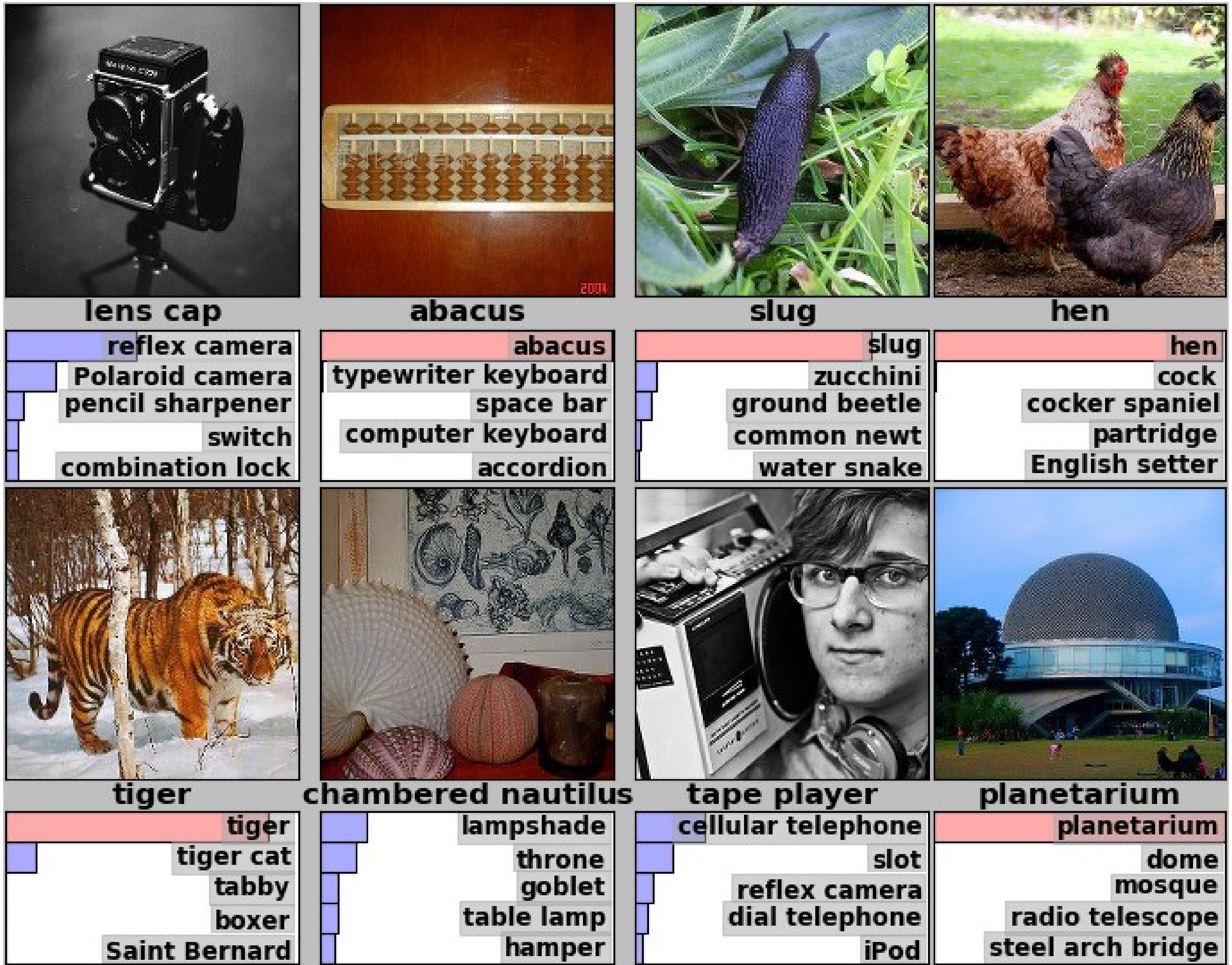


- resulting in:
  - 7 hidden layers, 650,000 neurons, 60,000,000 parameters
  - Trained on 2 GPUs for a week

Validation classification

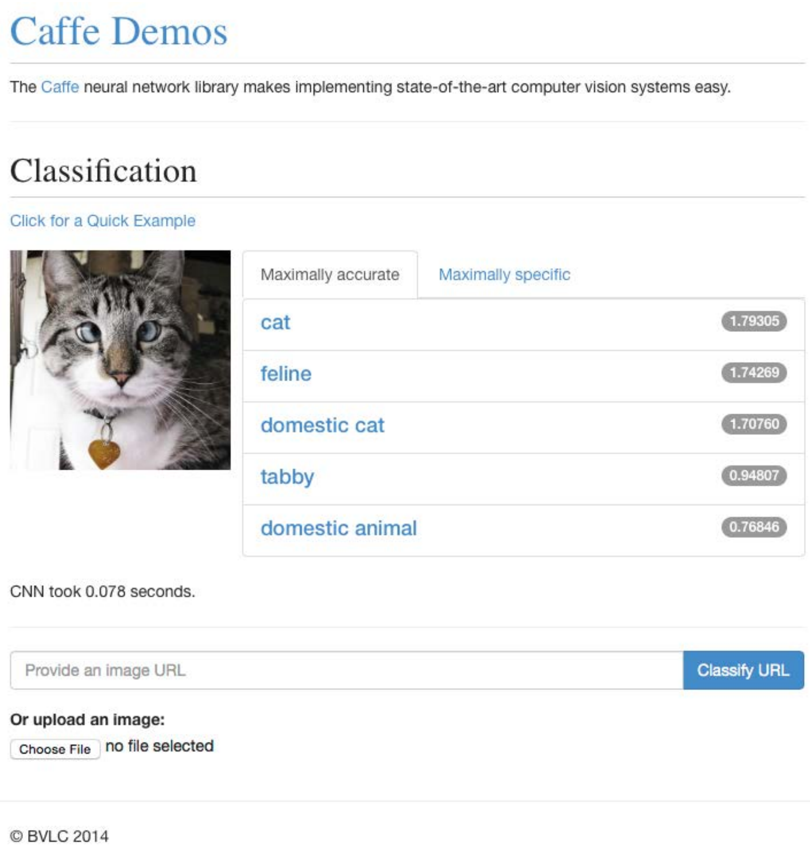


Validation classification



Try it out yourself

- Caffe ist an open implementation from the Berkeley Vision Group
  - http://caffe.berkeleyvision.org
  - http://demo.caffe.berkeleyvision.org

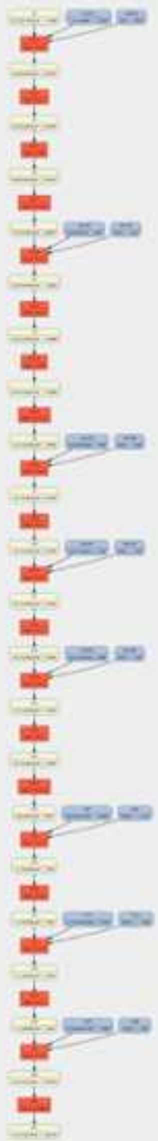


How deep is enough?

AlexNet (2012)

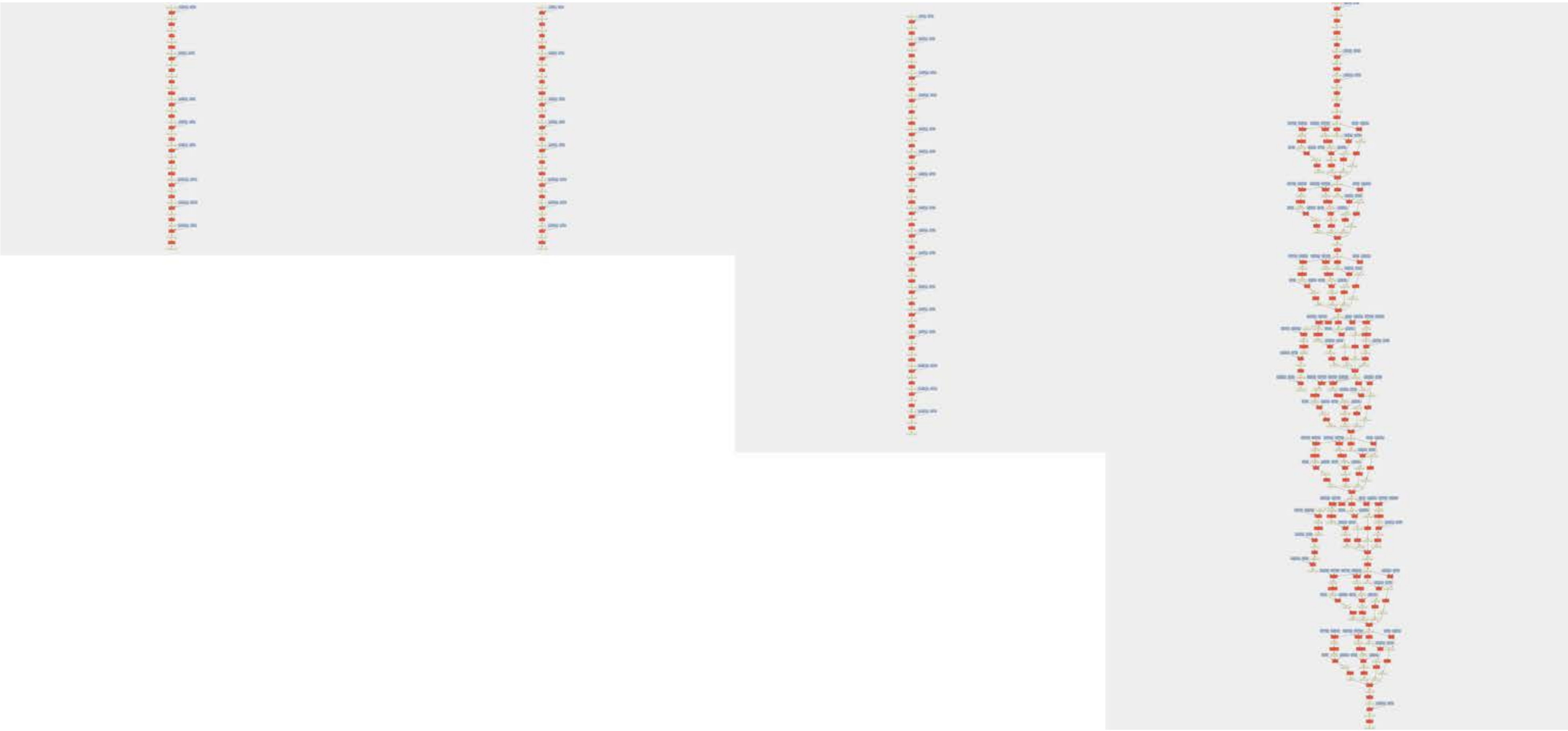
5 convolutional layers

3 fully-connected layers

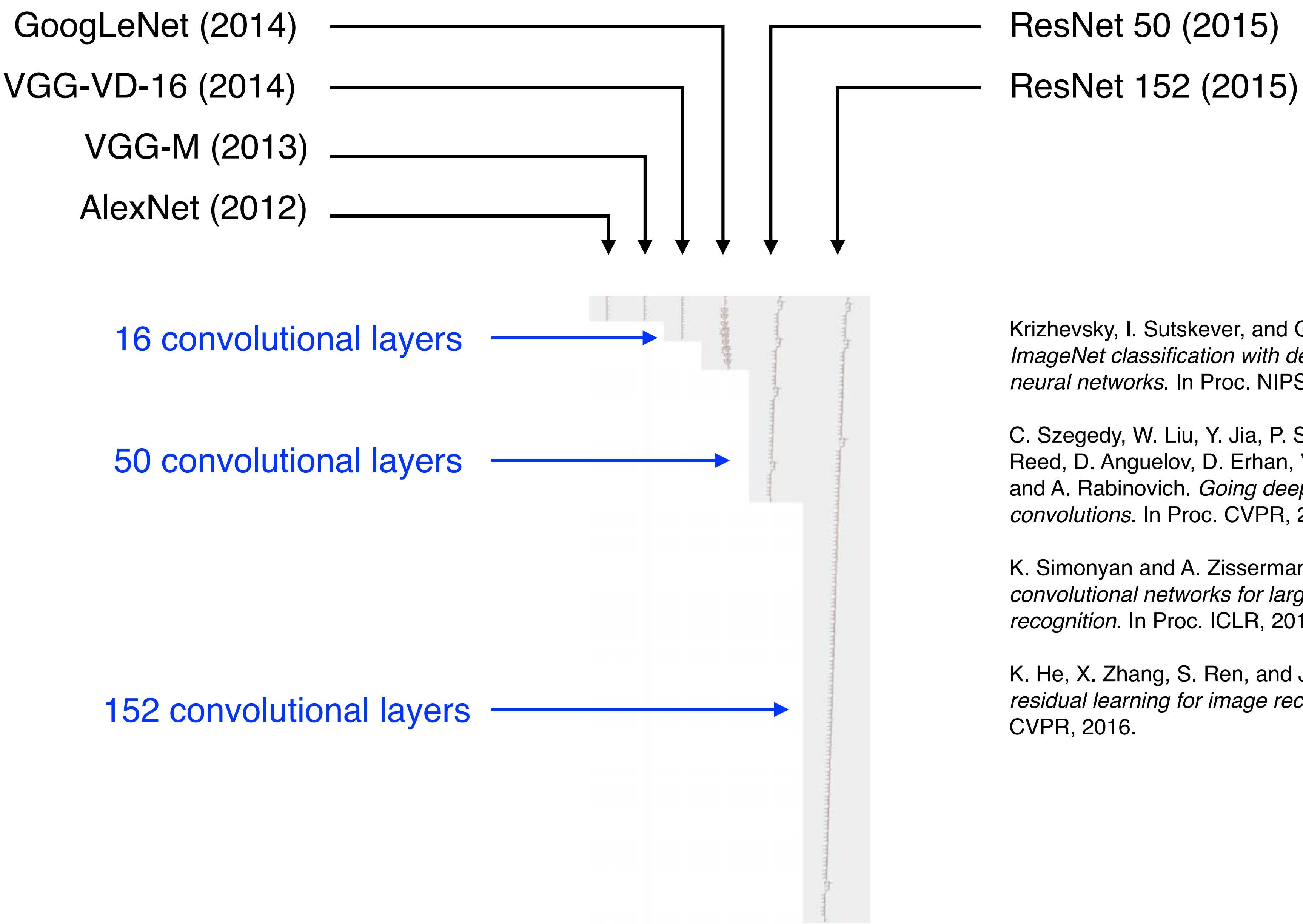


How deep is enough?

AlexNet (2012) VGG-M (2013) VGG-VD-16 (2014) GoogLeNet (2014)



How deep is enough?



Krizhevsky, I. Sutskever, and G. E. Hinton. *ImageNet classification with deep convolutional neural networks*. In Proc. NIPS, 2012.

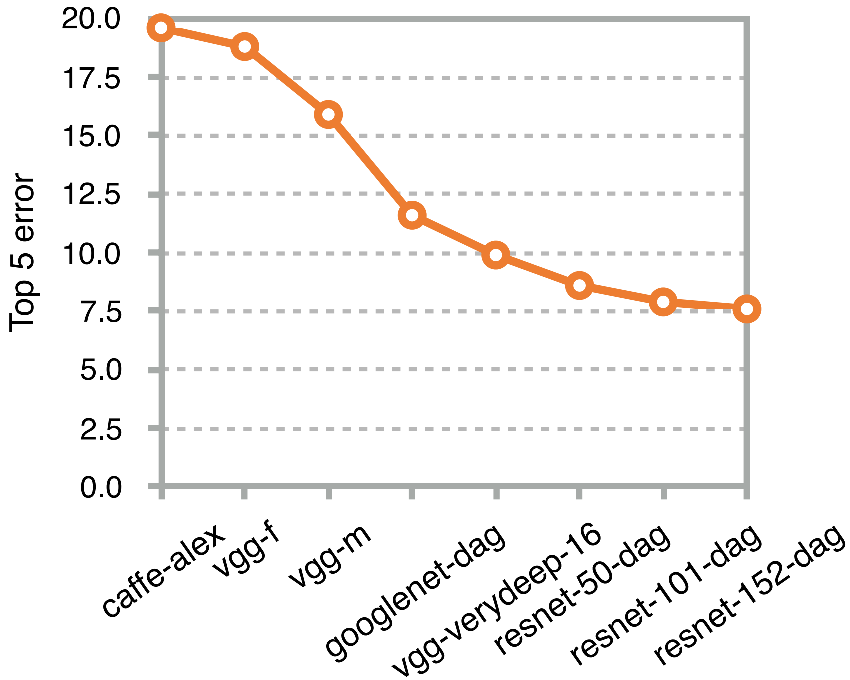
C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. *Going deeper with convolutions*. In Proc. CVPR, 2015.

K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. In Proc. ICLR, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. *Deep residual learning for image recognition*. In Proc. CVPR, 2016.

Accuracy

3 × more accurate in 3 years



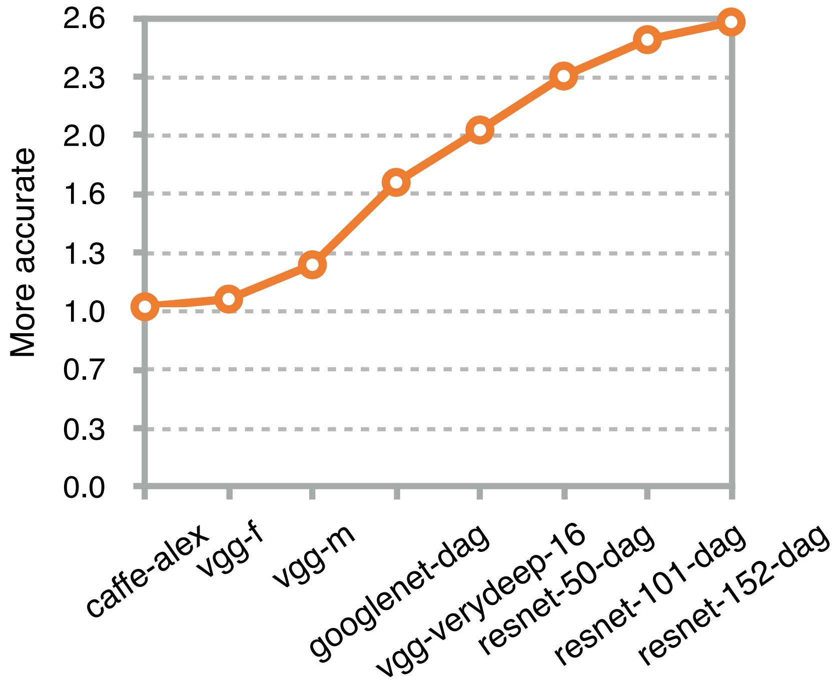


Image Description

A female tennis player in action on the court.

A group of young men playing a game of soccer.

A man riding a wave on top of a surfboard.

Image Description

A black and white cat is sitting on a chair.

A large clock mounted to the side of a building.

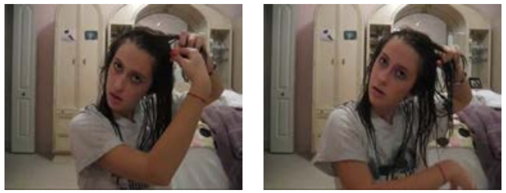
A bunch of fruit that are sitting on a table.

Video Description

Correct descriptions.

S2VT: A man is doing stunts on his bike.

S2VT: A herd of zebras are walking in a field.

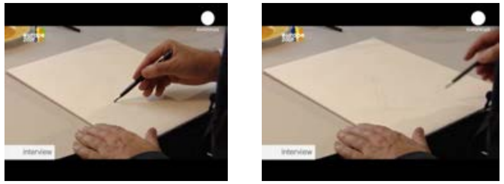


S2VT: A young woman is doing her hair.

S2VT: A man is shooting a gun at a target.

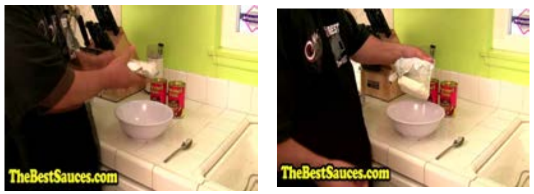
Relevant but incorrect descriptions.

S2VT: A small bus is running into a building.



S2VT: A man is cutting a piece of a pair of a paper.

S2VT: A cat is trying to get a small board.



S2VT: A man is spreading butter on a tortilla.

Irrelevant descriptions.

S2VT: A man is pouring liquid in a pan.

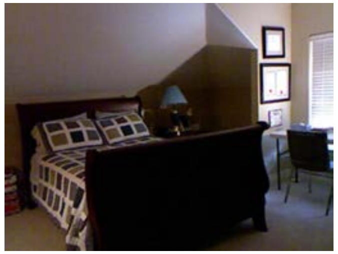
S2VT: A polar bear is walking on a hill.

S2VT: A man is doing a pencil.

S2VT: A black clip to walking through a path.

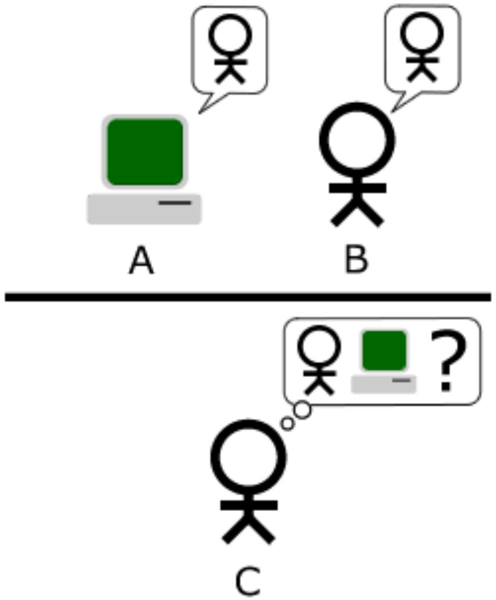
Towards a Visual Turing Challenge

Q: What is the object on the counter in the corner?
A: micro wave



What is the color of the largest object in the scene?
A: brown

Q:How many lights are on?
A: 6



- 1449 RGB-D images (NYU depth dataset)
- 12500 question-answer-pairs
- Publicly available

Question Answering Results

What is on the right side of the cabinet?
Vision + Language: **bed**
Language Only: **bed**

What objects are found on the bed?
Vision + Language: **bed sheets, pillow**
Language Only: **doll, pillow**

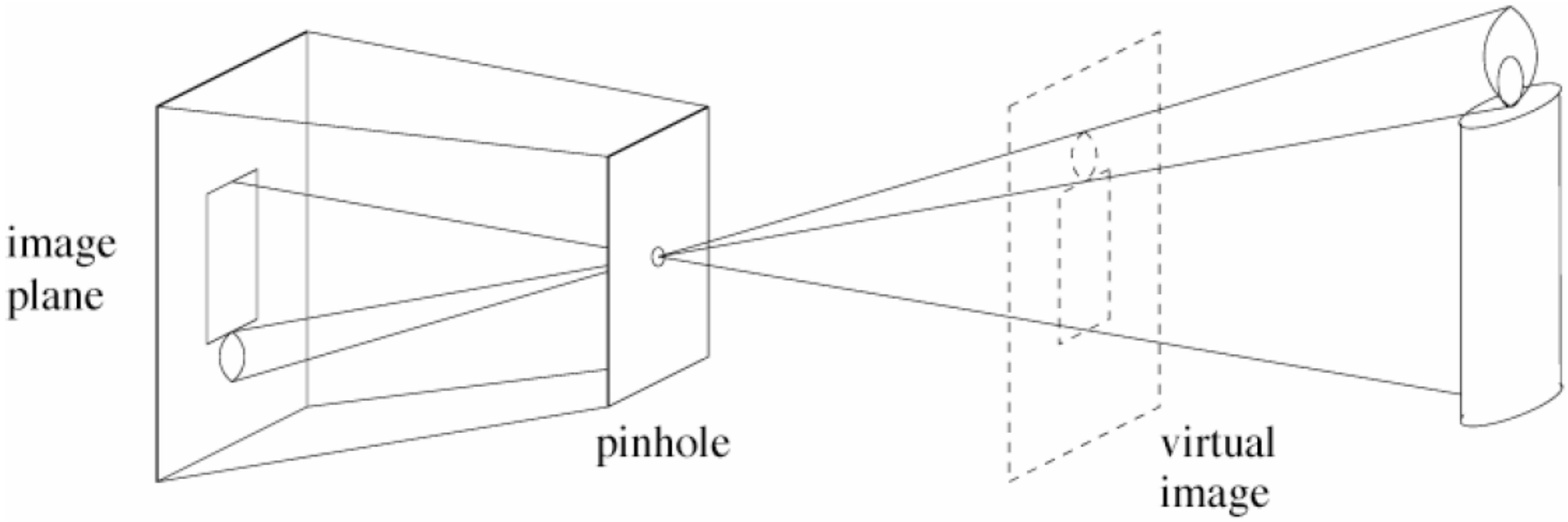
How many burner knobs are there?
Vision + Language: **4**
Language Only: **6**

Basic Concepts and Terminology

Computer Vision vs. Computer Graphics

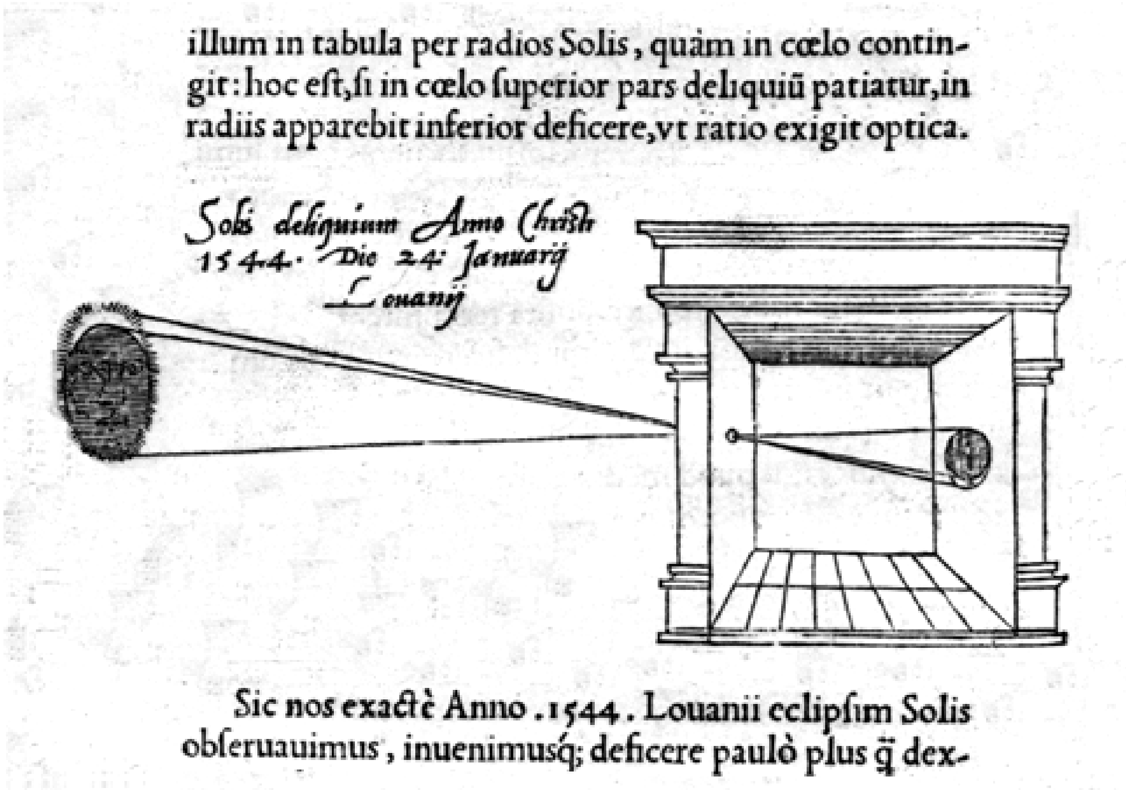
Pinhole Camera (Model)

- (simple) standard and abstract model today
  - box with a small hole in it



Camera Obscura

- around 1519, Leonardo da Vinci (1452 - 1519)
  - http://www.acmi.net.au/AIC/CAMERA_OBSCURA.html
  - "when images of illuminated objects … penetrate through a small hole into a very dark room … you will see [on the opposite wall] these objects in their proper form and color, reduced in size … in a reversed position owing to the intersection of the rays"

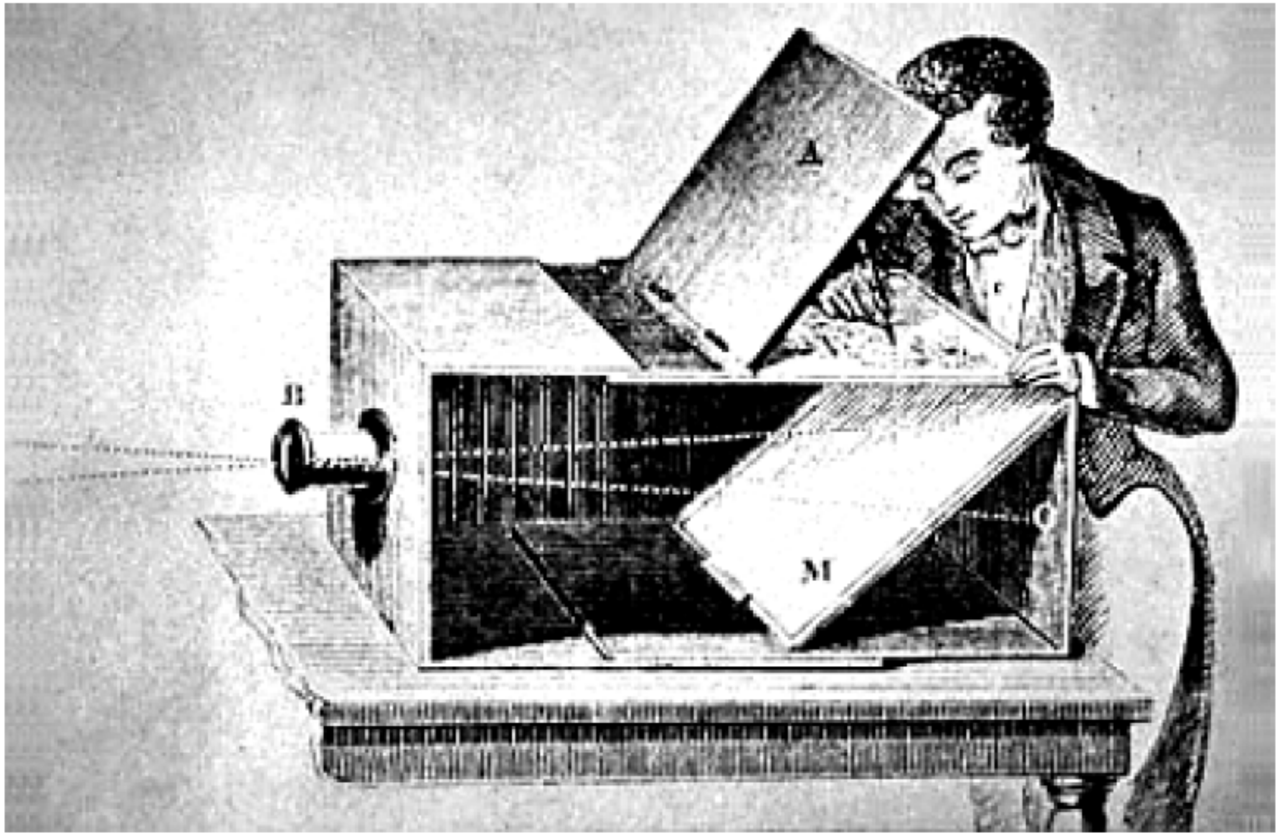

illum in tabula per radios Solis, quàm in cœlo contin-
git: hoc eſt, ſi in cœlo ſuperior pars deliquiũ patiatur, in
radiis apparebit inferior deficere, vt ratio exigit optica.

Solis deliquium Anno Christi
1544. Die 24 Januarij
Louanij

Sic nos exactè Anno .1544. Louanii eclipſim Solis
obſeruauimus, inuenimusq; deficere paulò plus q̃ dex-

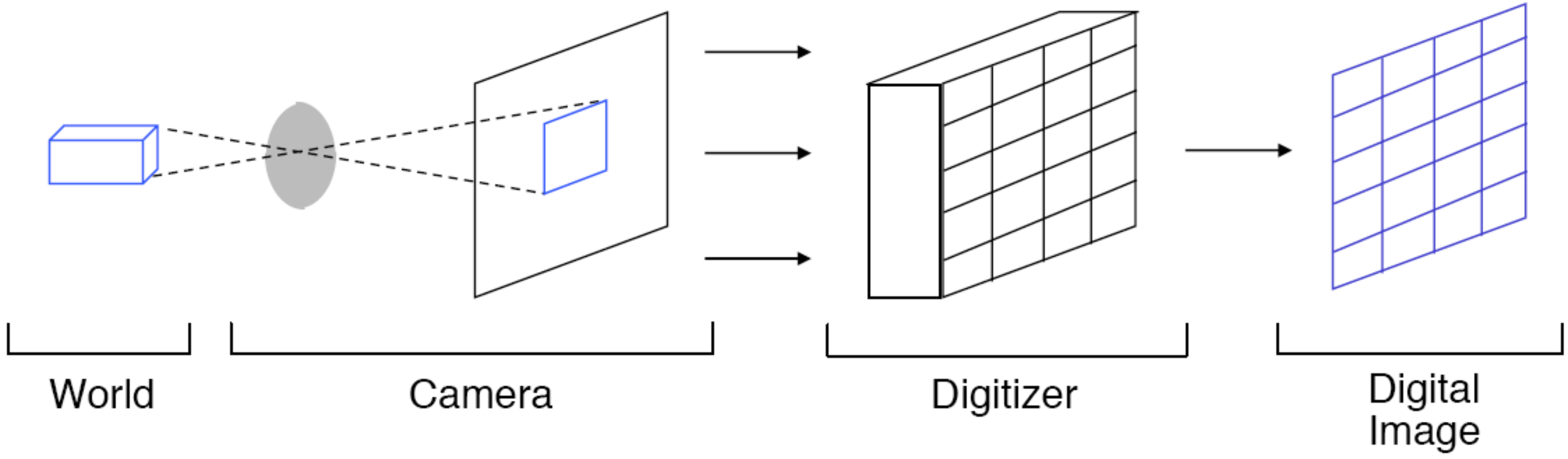
Principle of pinhole....

- ...used by artists
  - (e.g. Vermeer 17th century, dutch)
- and scientists



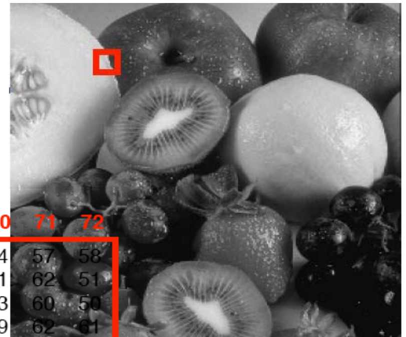
Digital Images

- Imaging Process:
  - (pinhole) camera model
  - digitizer to obtain digital image



(Grayscale) Image

- ‘Goals’ of Computer Vision
  - how can we recognize fruits from an array of (gray-scale) numbers?
  - how can we perceive depth from an array of (gray-scale) numbers?
  - …
- ‘Goals’ of Graphics
  - how can we generate an array of (gray-scale) numbers that looks like fruits?
  - how can we generate an array of (gray-scale) numbers so that the human observer perceives depth?
  - …



- computer vision = the problem of ‘inverse graphics’ …?

y = \ x =	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
41	210	209	204	202	197	247	143	71	64	80	84	54	54	57	58
42	206	196	203	197	195	210	207	56	63	58	53	53	61	62	51
43	201	207	192	201	198	213	156	69	65	57	55	52	53	60	50
44	216	206	211	193	202	207	208	57	69	60	55	77	49	62	61
45	221	206	211	194	196	197	220	56	63	60	55	46	97	58	106
46	209	214	224	199	194	193	204	173	64	60	59	51	62	56	48
47	204	212	213	208	191	190	191	214	60	62	66	76	51	49	55
48	214	215	215	207	208	180	172	188	69	72	55	49	56	52	56
49	209	205	214	205	204	196	187	196	86	62	66	87	57	60	48
50	208	209	205	203	202	186	174	185	149	71	63	55	55	45	56
51	207	210	211	199	217	194	183	177	209	90	62	64	52	93	52
52	208	205	209	209	197	194	183	187	187	239	58	68	61	51	56
53	204	206	203	209	195	203	188	185	183	221	75	61	58	60	60
54	200	203	199	236	188	197	183	190	183	196	122	63	58	64	66
55	205	210	202	203	199	197	196	181	173	186	105	62	57	64	63

Visual Cues for Image Analysis

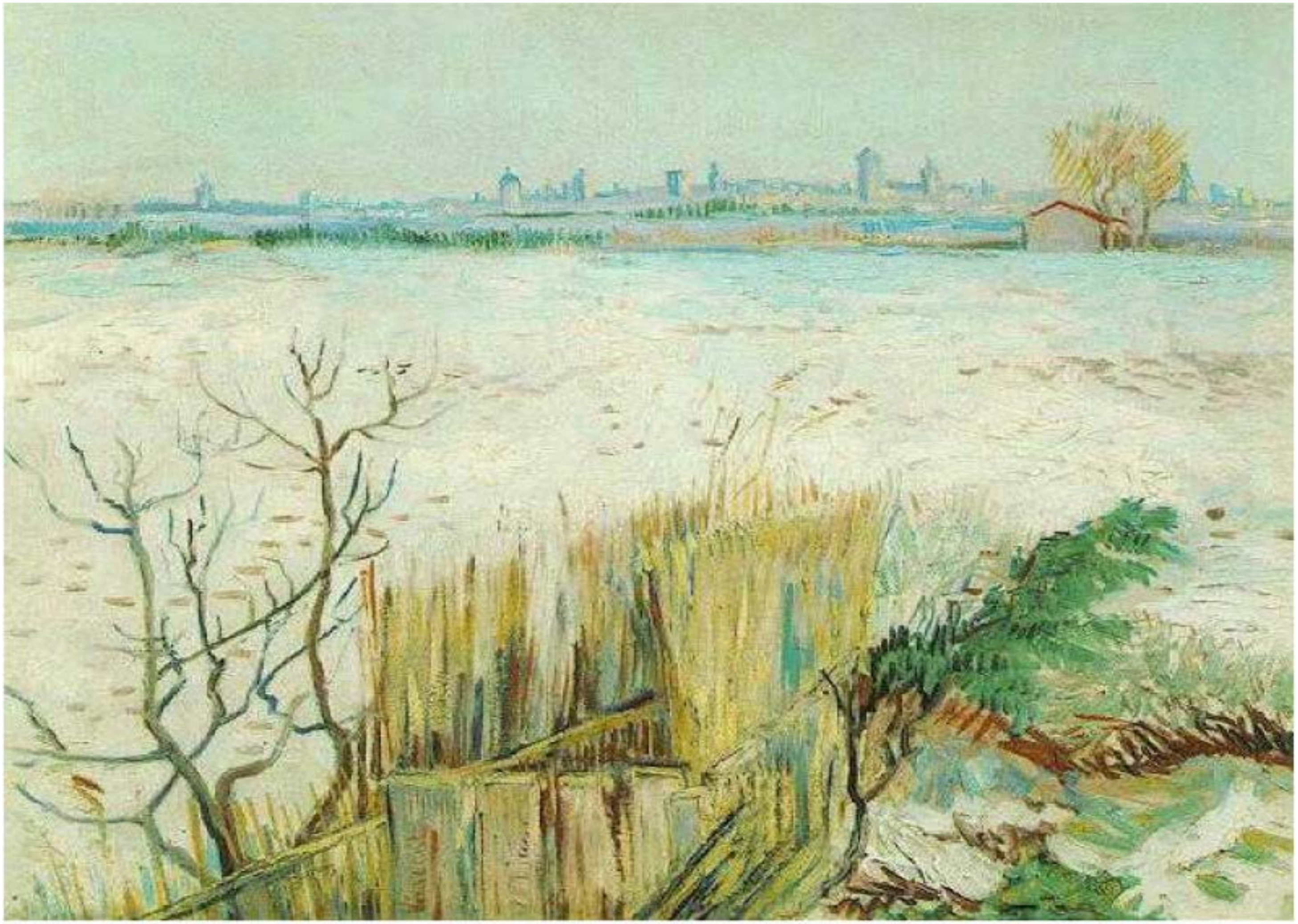
... in art and visual illusions

1. Case Study:
Human & Art - Recovery of 3D Structure

1. Case Study: Human & Art - Recovery of 3D Structure

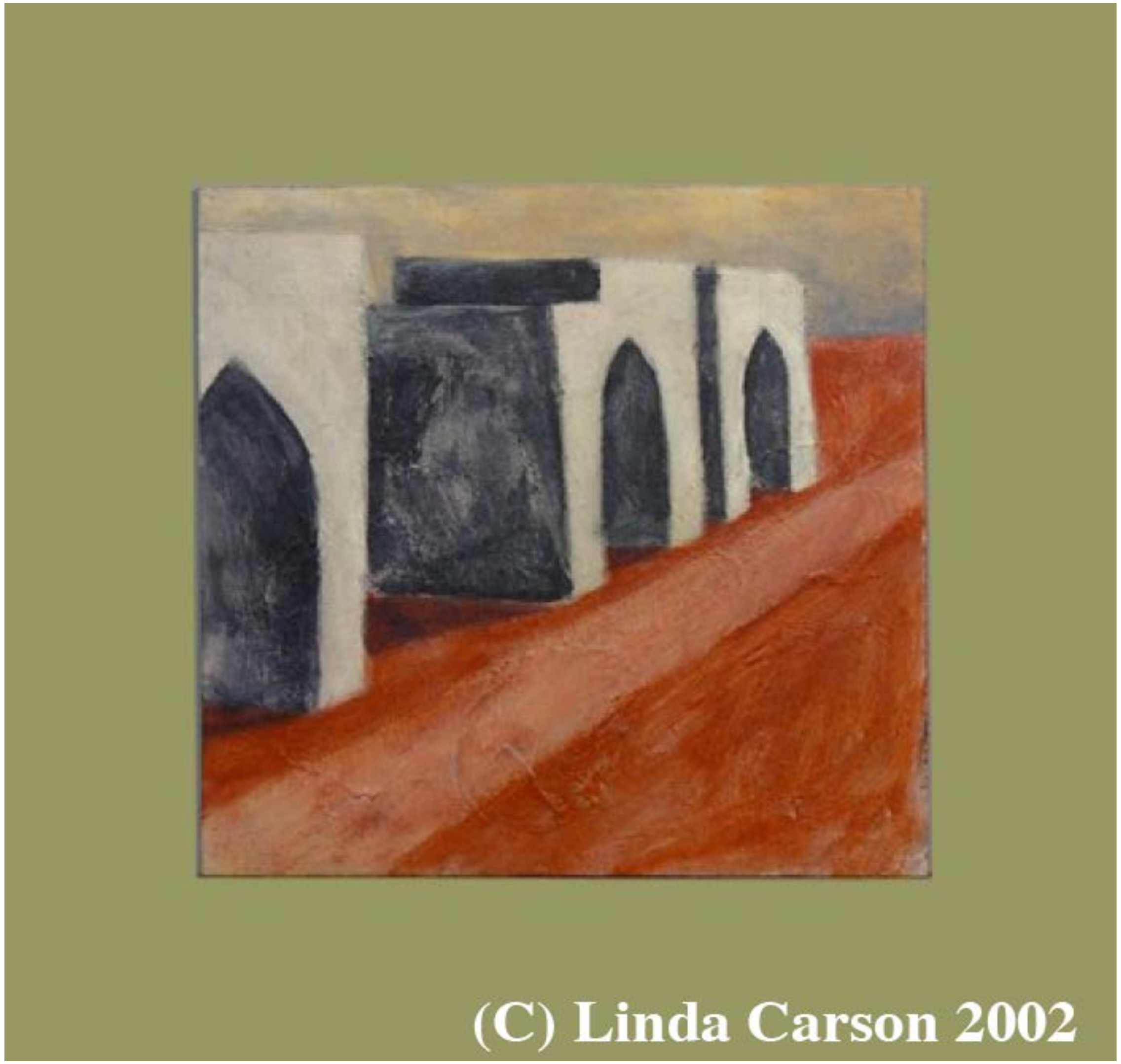
Vincent van Gogh *Interior of a Restaurant at Arles* 1888

1. Case Study: Human & Art - Recovery of 3D Structure

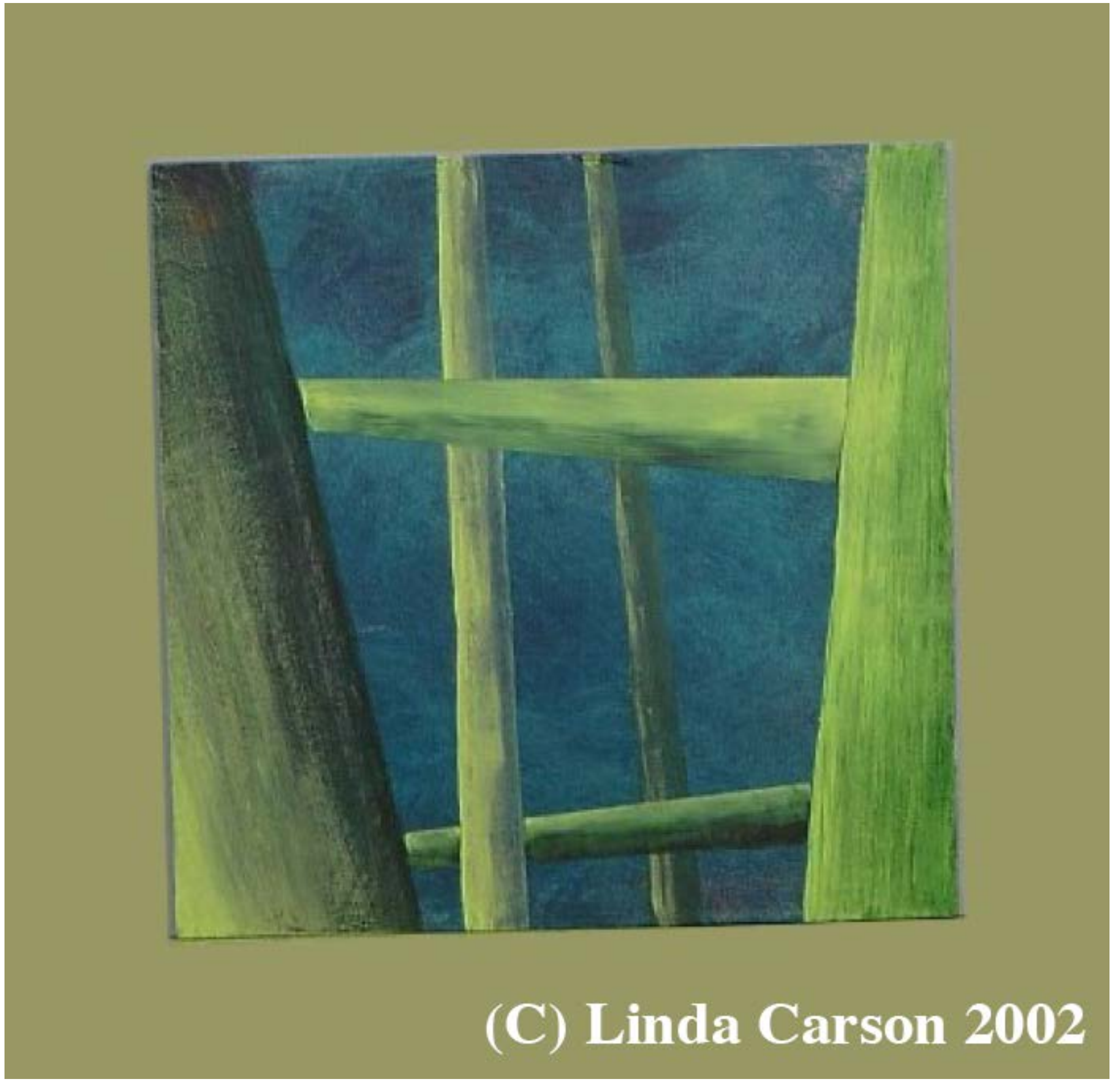


Vincent van Gogh *Snowy Landscape with Arles in the Background* 1888

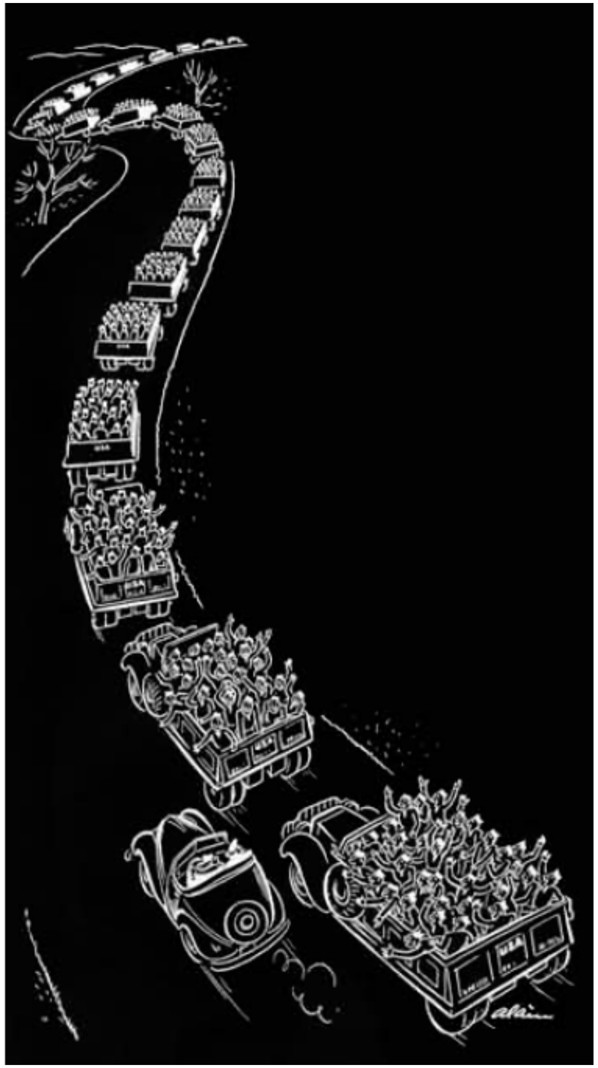
1. Case Study:
Human & Art - Recovery of 3D Structure

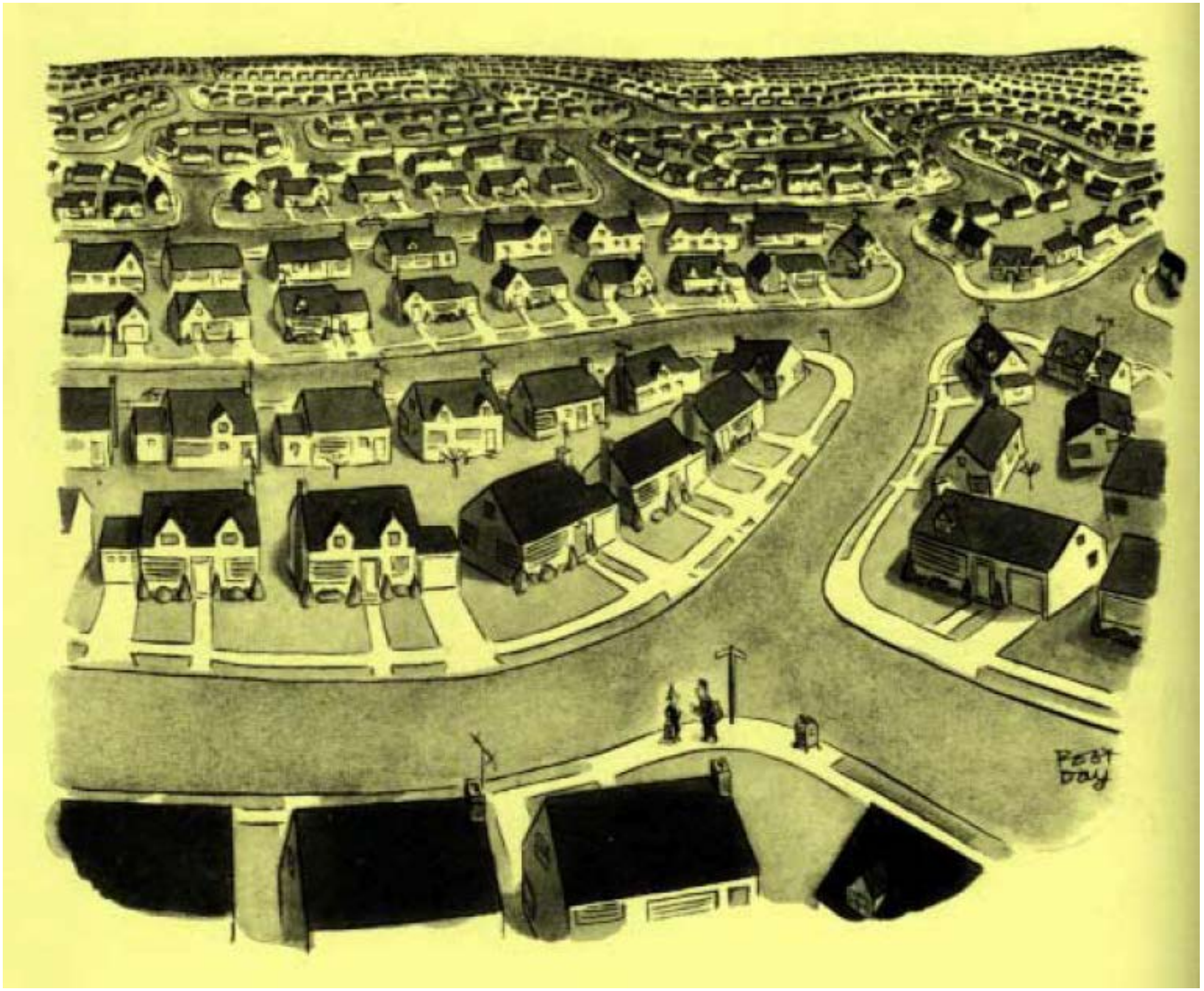


1. Case Study:
Human & Art - Recovery of 3D Structure



1. Case Study:
Human & Art - Recovery of 3D Structure

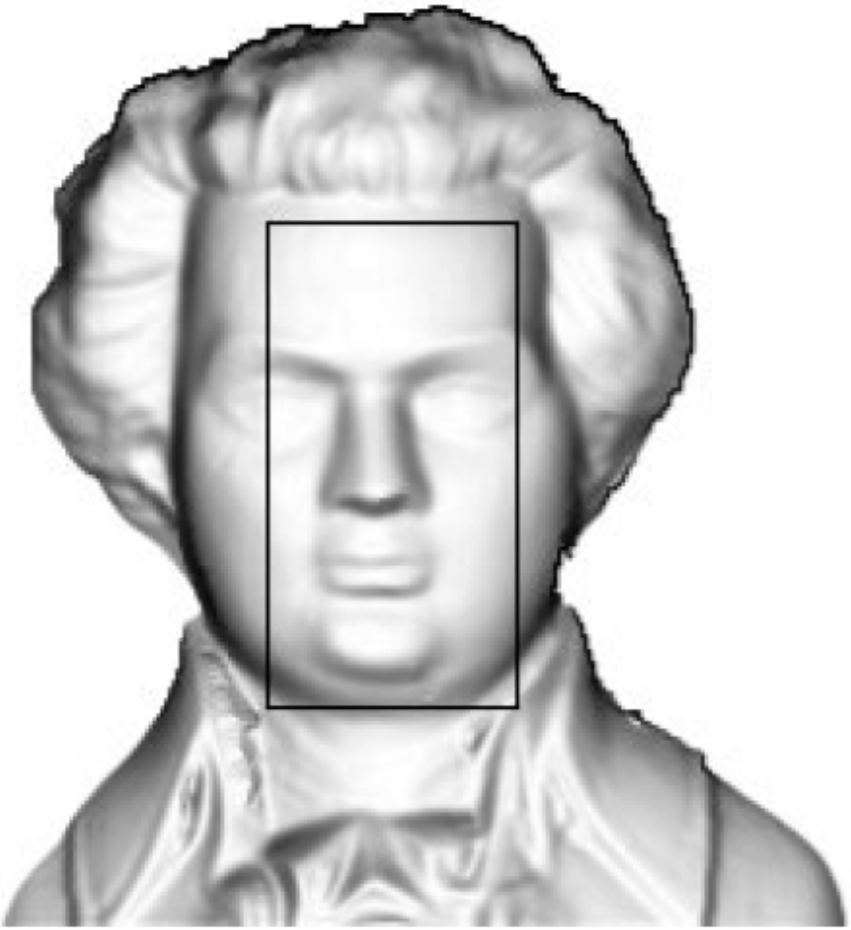




1. Case Study
Computer Vision - Recovery of 3D Structure

- take all the cues of artists and 'turn them around'
  - exploit these cues to **infer** the structure of the world
  - need **mathematical** and **computational models** of these cues

- sometimes called 'inverse graphics'



http://www.vrvis.at/ar2/adm/shading/

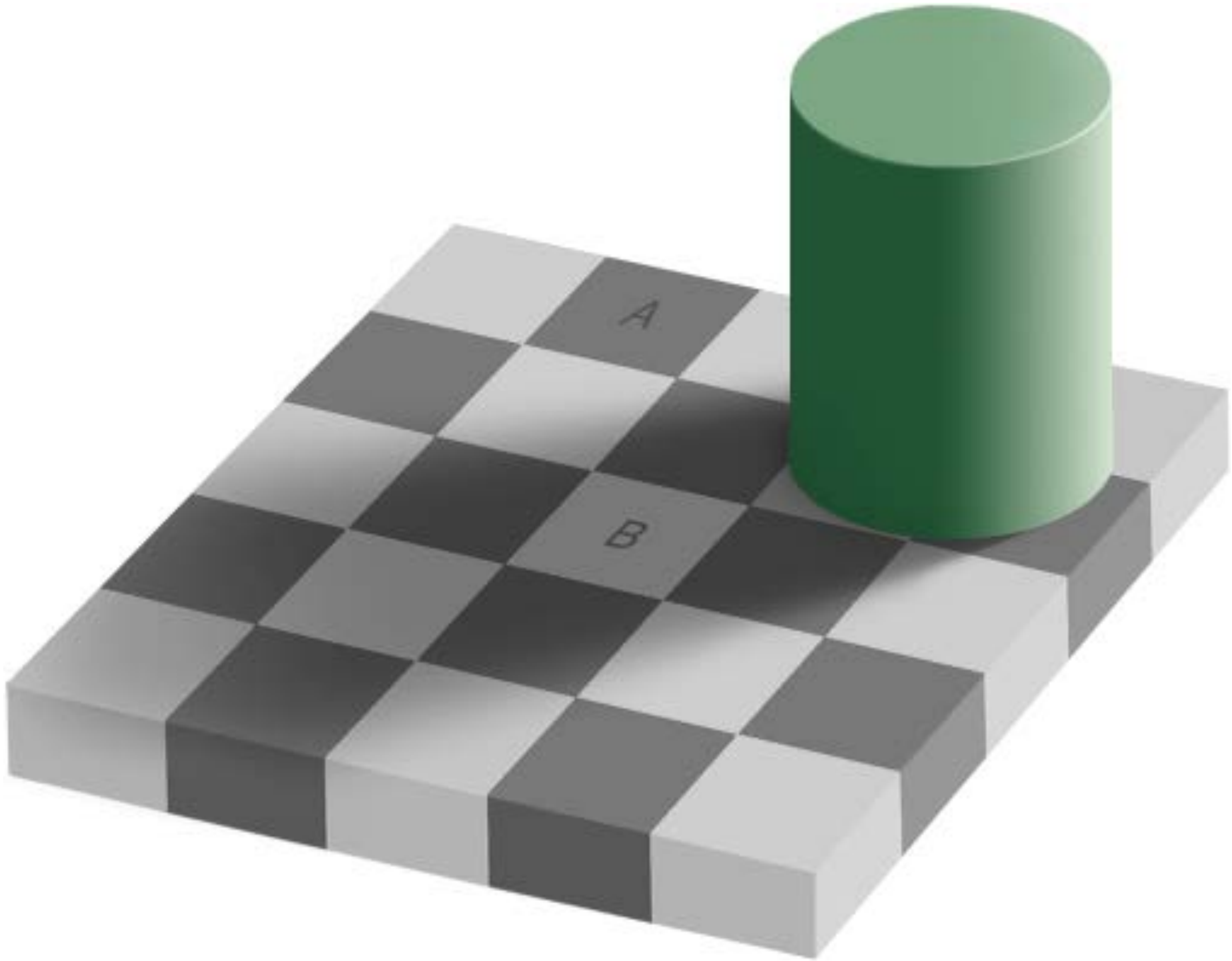
A ‘trompe l’oeil’

- depth-perception
  - movement of ball stays the same
  - location/trace of shadow changes

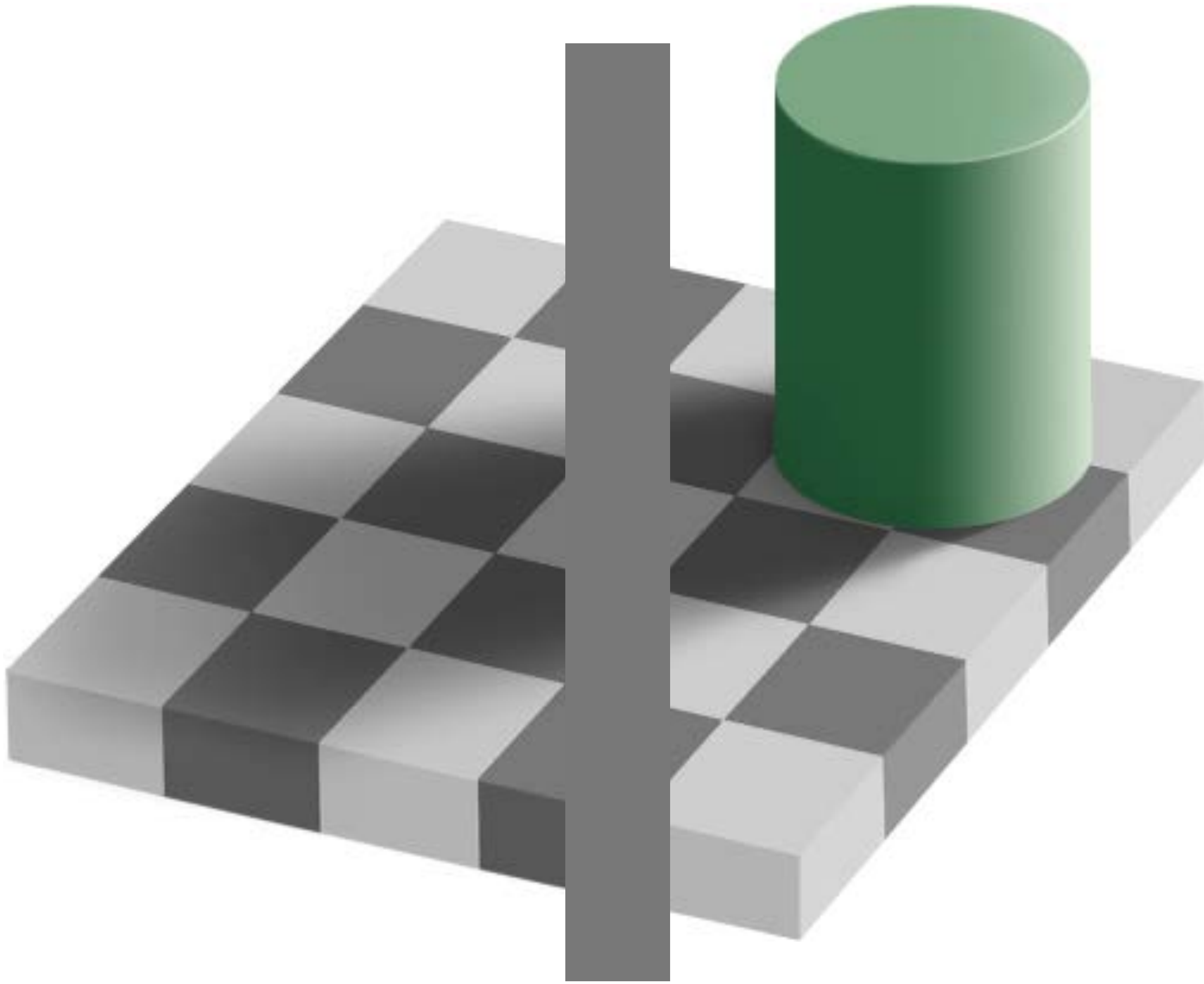
Another ‘trompe l’oeil’

- illusory motion
  - only shadows changes
  - square is stationary

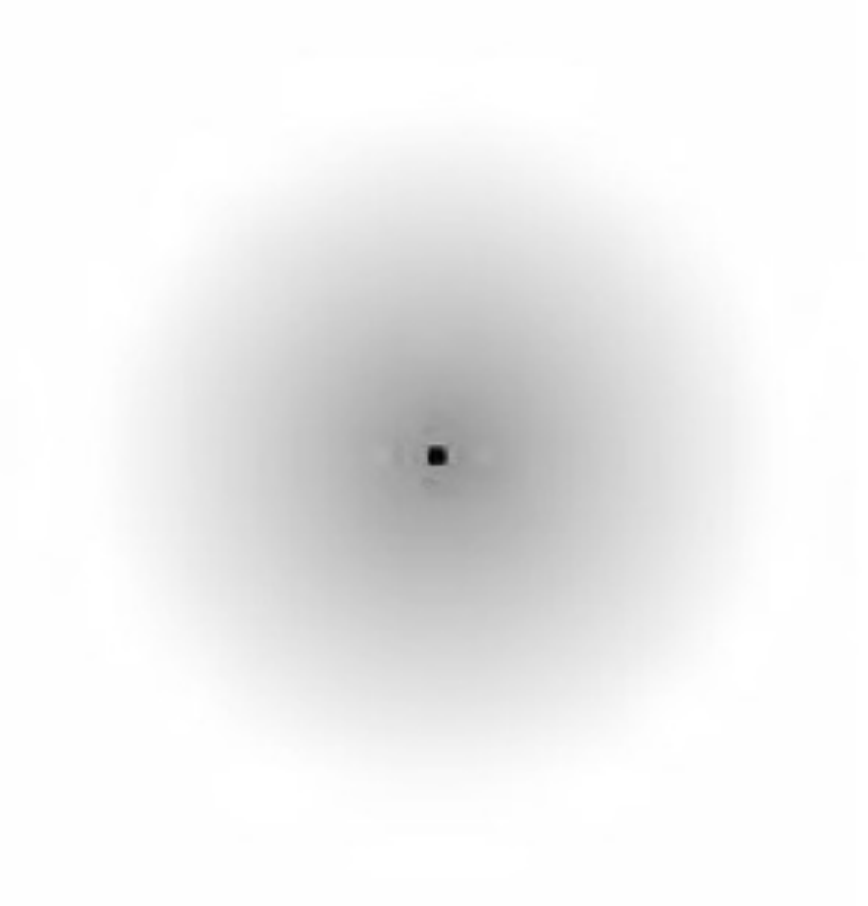
Color & Shading



Color & Shading



Keep staring at the black dot. After a while the gray haze around it will appear to shrink.



Do you still believe what you see?

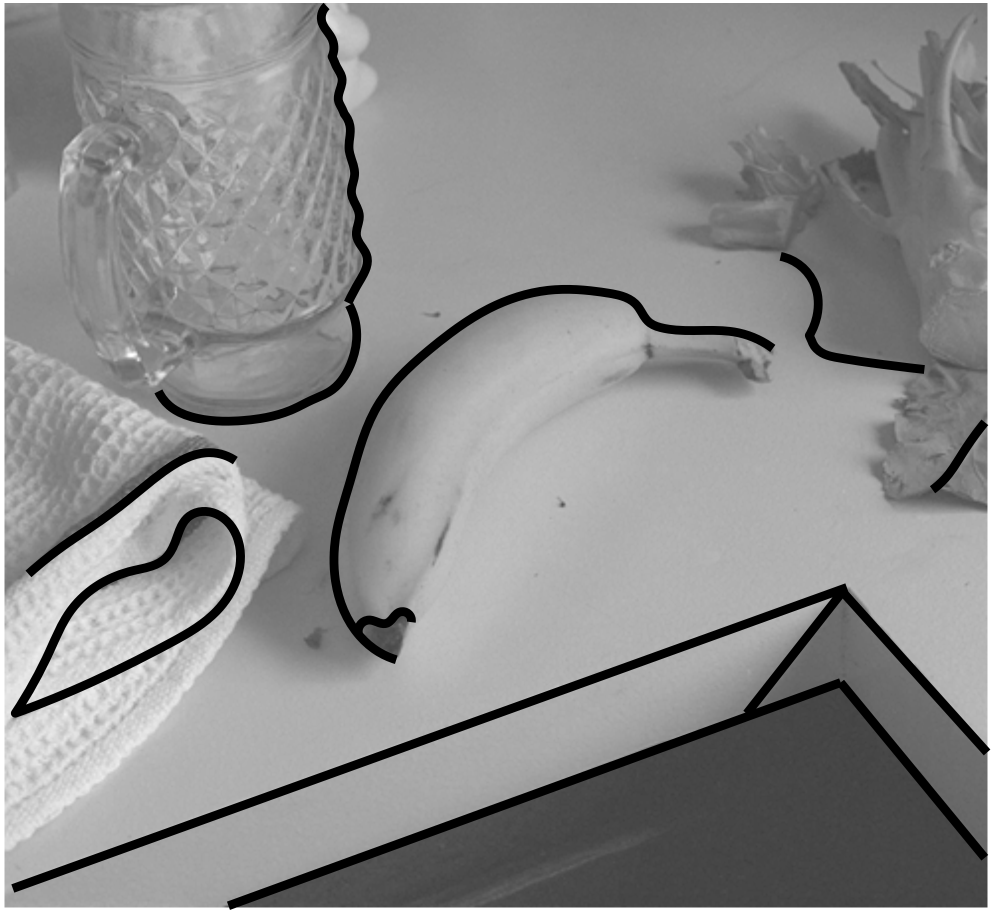
- Experiment
  - carefully point flash light into your eye from one corner
  - don’t hurt yourself!

- Observation
  - you’ll see your own blood vessels
  - they are actually in front of the retina
  - we’ve adapted to their usual shadow

2. Case Study: Computer Vision & Object Recognition

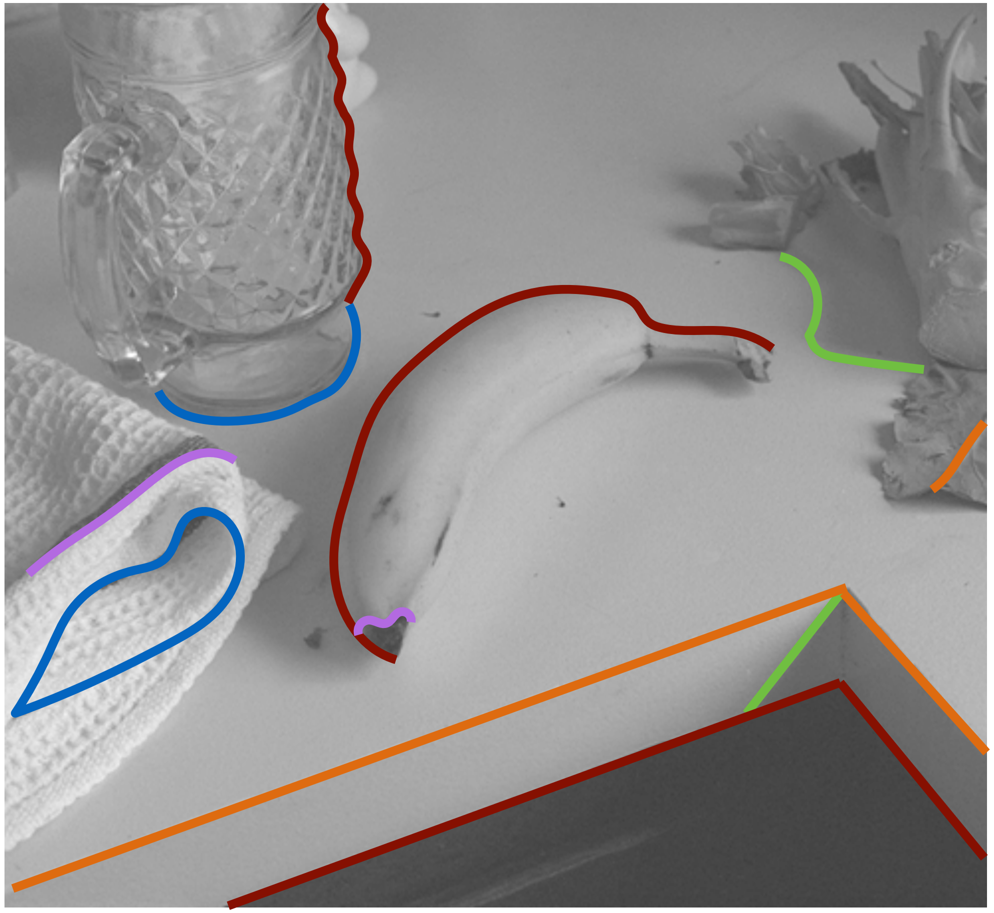
- is it more than inverse graphics?
- how do you recognize
  - the banana?
  - the glas?
  - the towel?
- how can we make computers to do this?
- ill posed problem:
  - missing data
  - ambiguities
  - multiple possible explanations

Image Edges:
What are edges? Where do they come from?



- Edges are changes in pixel brightness

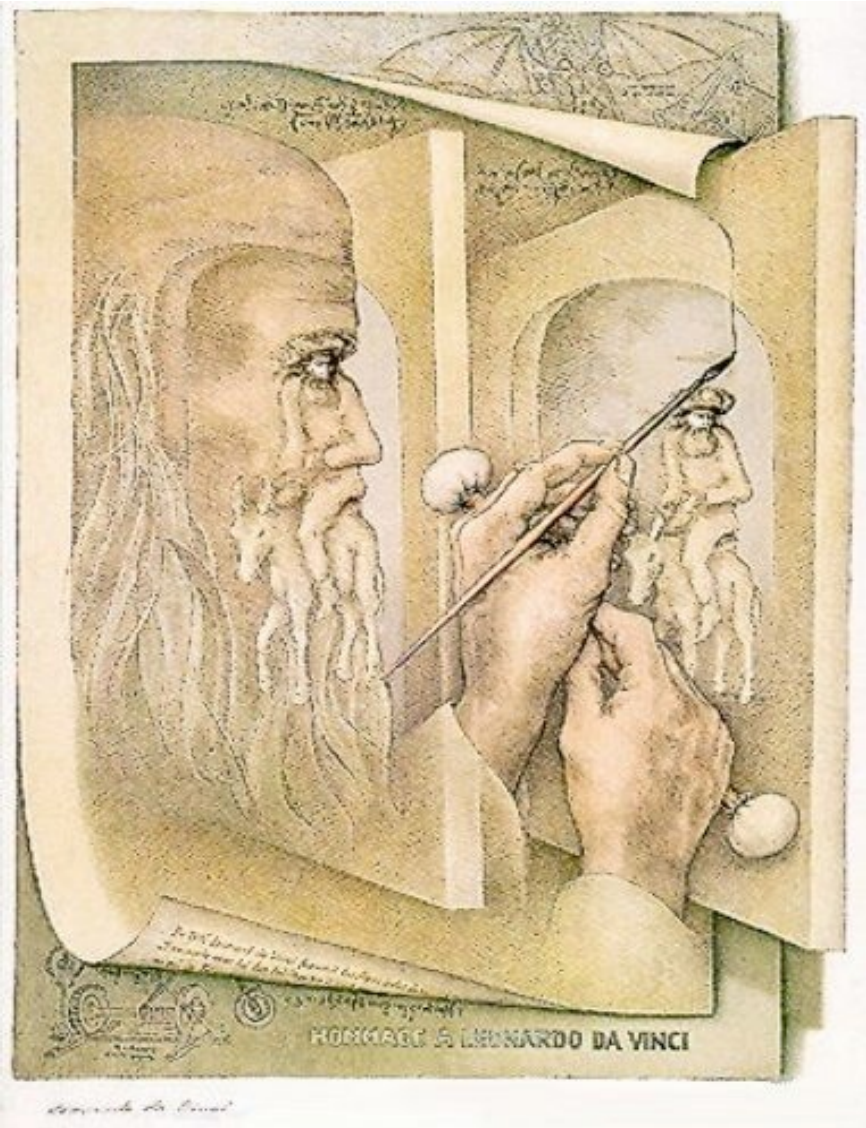
Image Edges:
What are edges? Where do they come from?



- Edges are changes in pixel brightnes
  - Foreground/Background Boundaries
  - Object-Object-Boundaries
  - Shadow Edges
  - Changes in Albedo or Texture
  - Changes in Surface Normals

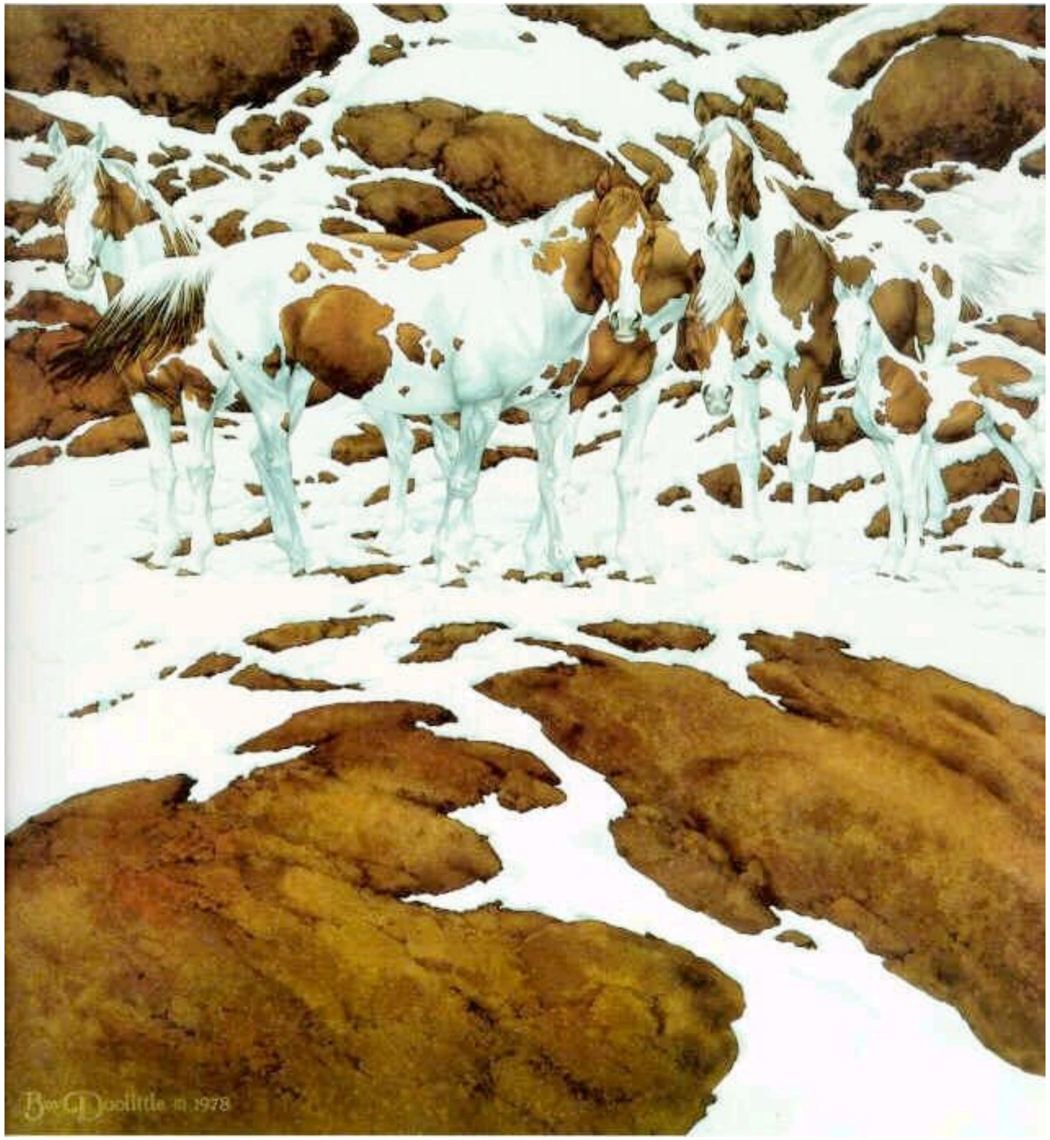
Line Drawings: Good Starting Point for Recognition?

Complexity of Recognition

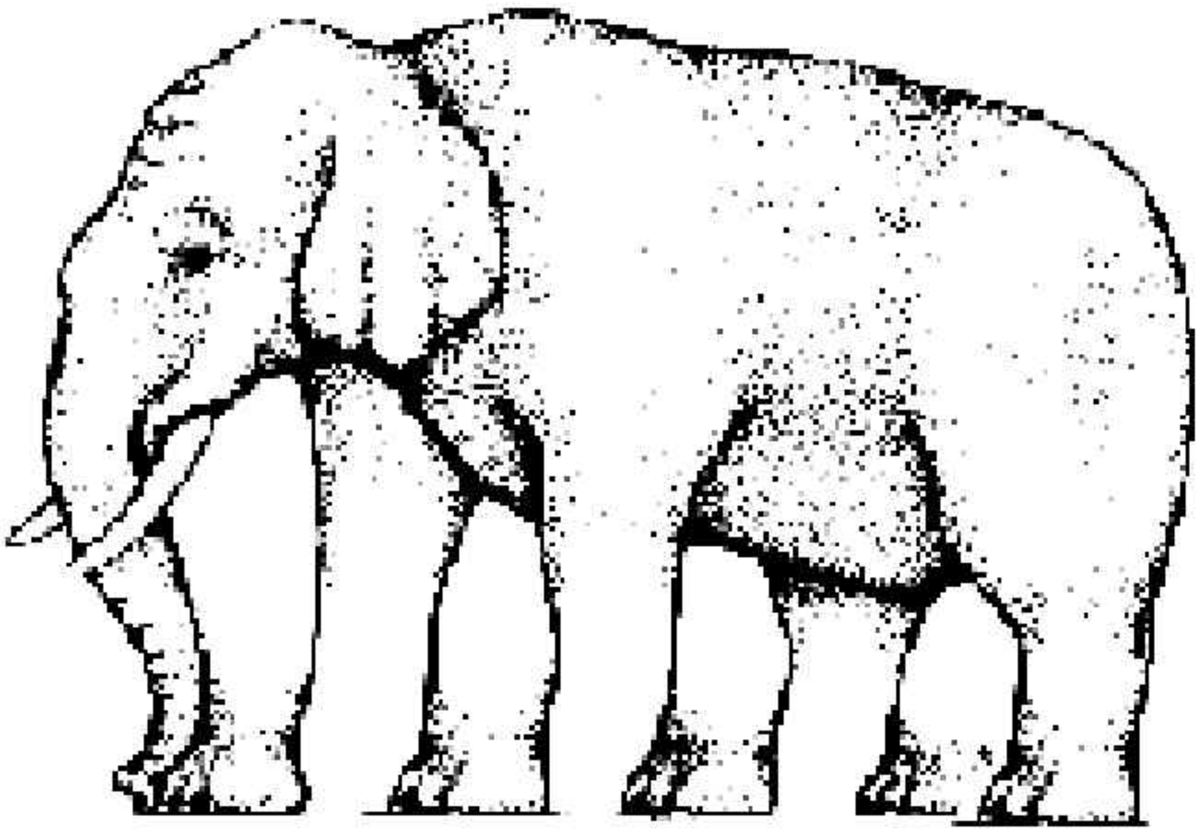


Complexity of Recognition

Complexity of Recognition



Complexity of Recognition



Recognition: the Role of Context

- Antonio Torralba

Recognition: the role of Prior Expectation

- Guiseppe Arcimboldo

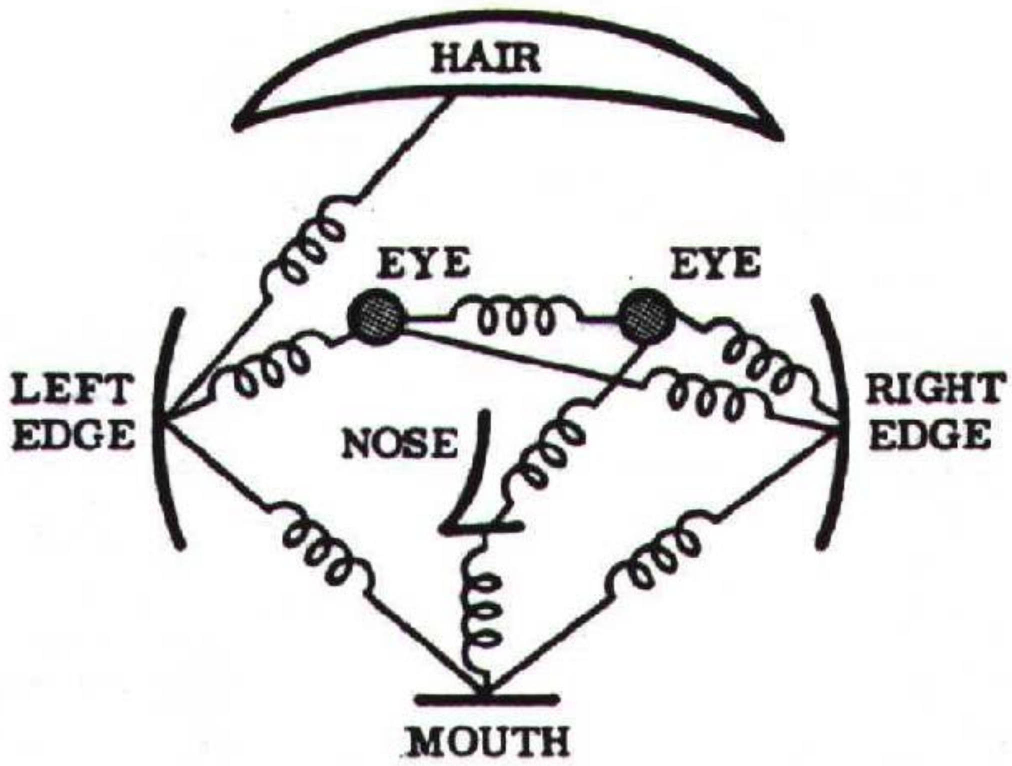
Complexity of Recognition

Complexity of Recognition

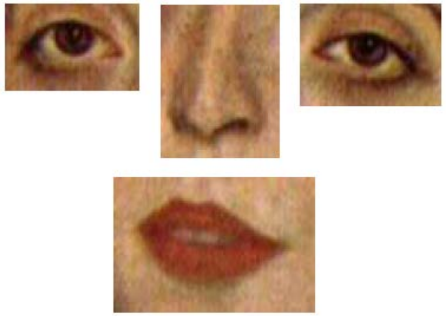
One or Two Faces ?

Class of Models: Pictorial Structure

- Fischler & Elschlager 1973
- Model has two components
  - parts
    (2D image fragments)
  - structure
    (configuration of parts)



Deformations



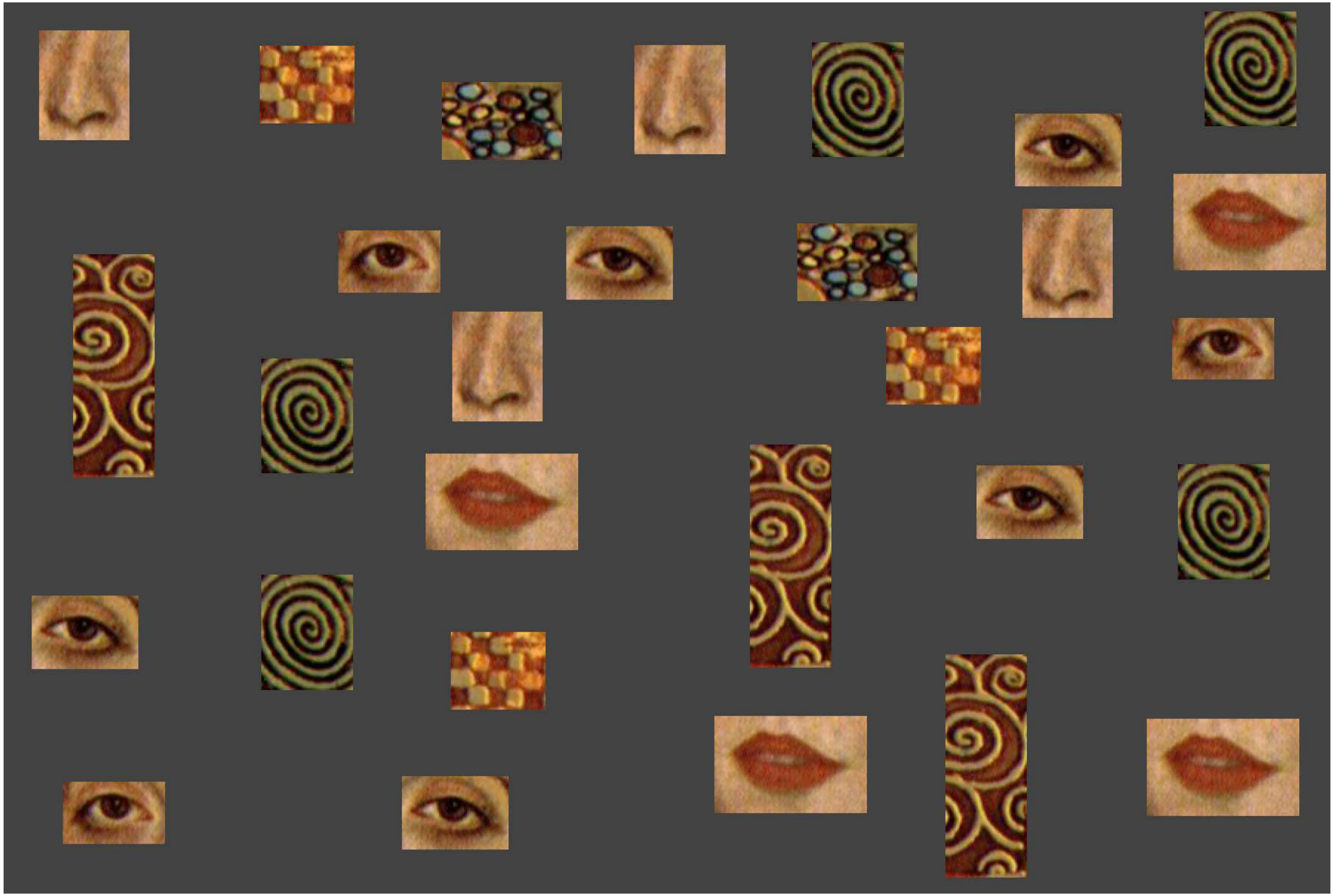
A

B

C

D

Clutter



Example

Recognition, Localization, and Segmentation

a few terms

… let’s briefly define what we mean by that

Object Recognition: First part of this Computer Vision class

- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your pencil, your dog, your car
  - Object **Classification**
    - recognize any pencil, any dog, any car
    - also called: generic object recognition, object categorization, …

- Recognition and
  - **Segmentation**: separate pixels belonging to the foreground (object) and the background
  - **Localization/Detection**: position of the object in the scene, pose estimate (orientation, size/scale, 3D position)

Object Recognition:
First part of this Computer Vision class

- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your apple, your cup, your dog
  - Object **Classification**
    - recognize any apple, any cup, any dog
    - also called: generic object recognition, object categorization, …
    - typical definition: ‘basic level category’

Which Level is right for Object Classes?

- Basic-Level Categories
  - the highest level at which category members have **similar perceived shape**
  - the highest level at which a **single mental image** can reflect the entire category
  - the highest level at which a person uses similar **motor actions** to interact with category members
  - the level at which human subjects are usually **fastest** at identifying category members
  - the first level named and understood by **children**

  - (while the definition of basic-level categories depends on culture there exist a remarkable consistency across cultures...)

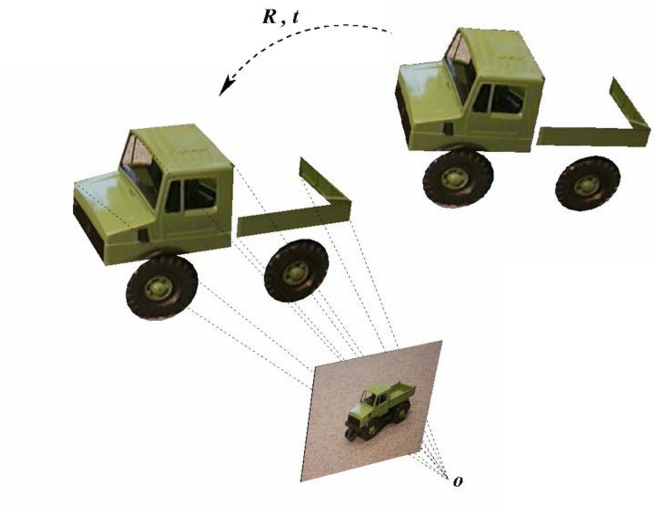
- Most recent work in object recognition has focused on this problem
  - we will discuss several of the most successful methods in the lecture :-)

Object Recognition: First part of this Computer Vision class

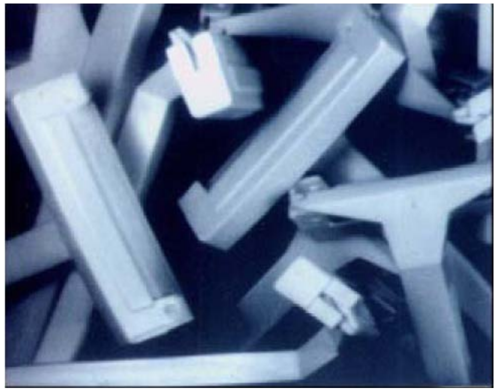
- Recognition and
  - Segmentation: separate pixels belonging to the foreground (object) and the background

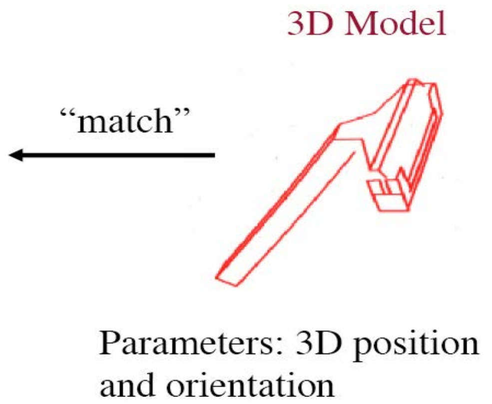
Object Recognition: First part of this Computer Vision class

- Recognition and
  - **Localization**: to position the object in the scene, estimate the object's pose (orientation, size/scale, 3D position)

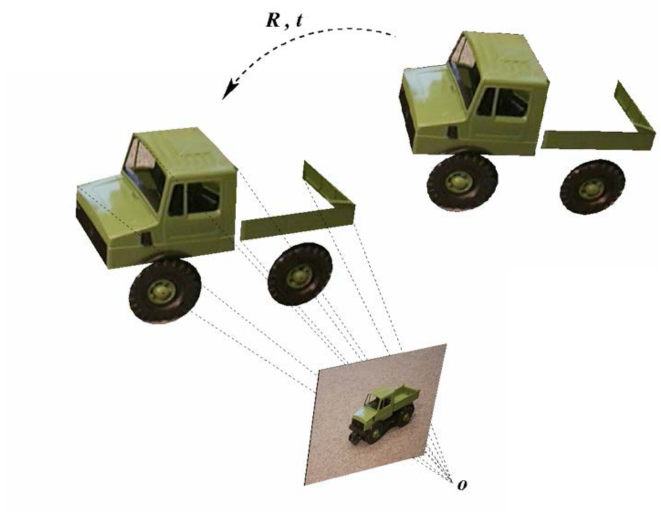


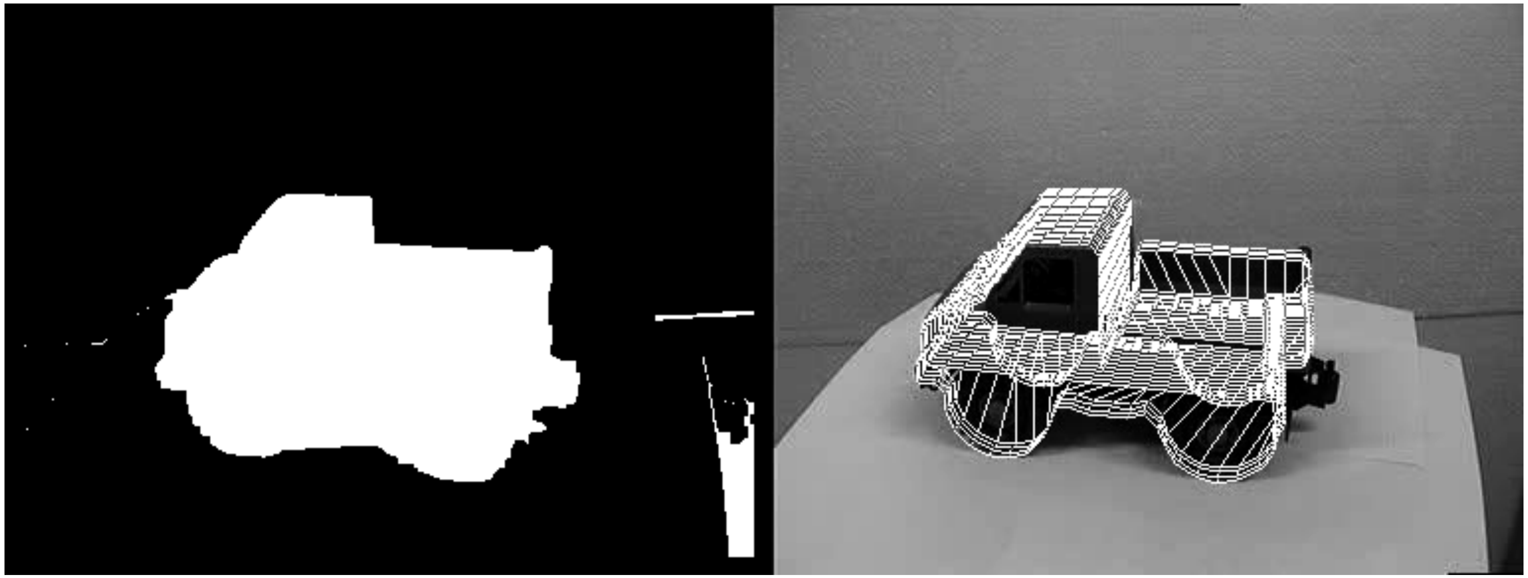
  - Example from David Lowe:





Localization: Example Video 1





Localization: Example Video 2

Object Recognition: First part of this Computer Vision class

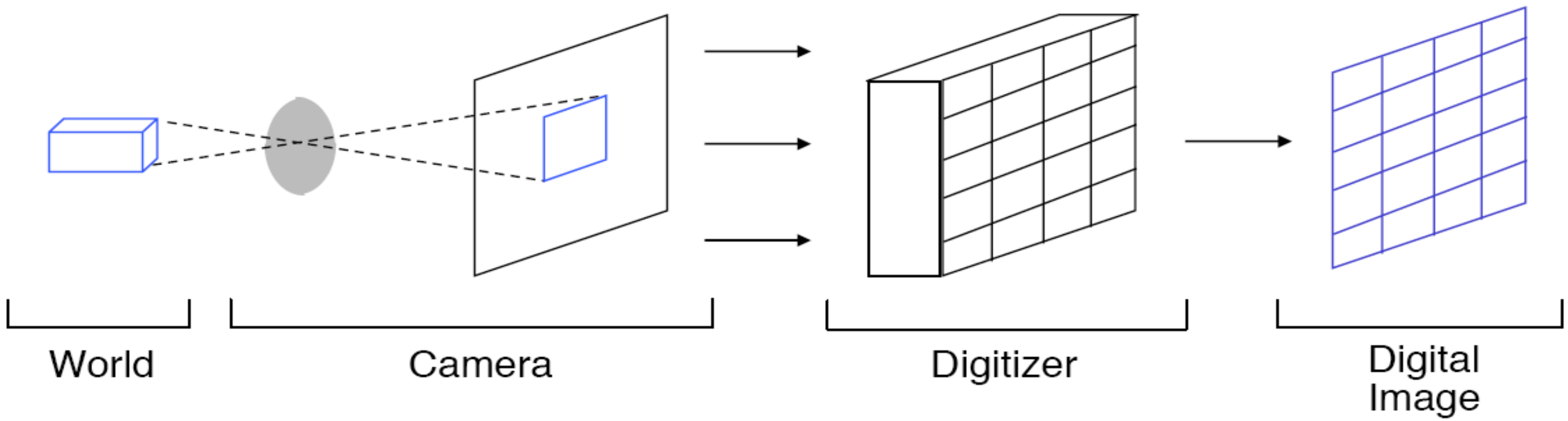
- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your pencil, your dog, your car
  - Object **Classification**
    - recognize any pencil, any dog, any car
    - also called: generic object recognition, object categorization, …

- Recognition and
  - **Segmentation**: separate pixels belonging to the foreground (object) and the background
  - **Localization**: position the object in the scene, estimate pose of the object (orientation, size/scale, 3D position)

Basic Filtering

Computer Vision and Fundamental Components

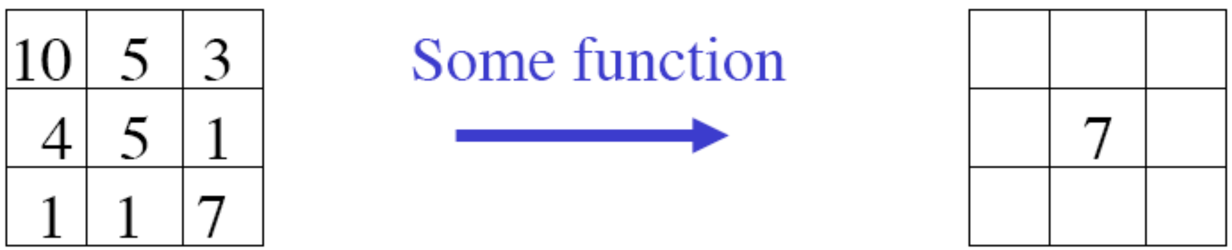
- computer vision: ‘reverse’ the imaging process
  - 2D (2-dimensional) digital image processing
  - ‘pattern recognition’ / 3D image analysis
  - image understanding



Digital Image Processing

- Some Basics
  - (digital signal processing, FFT, …)
  - Image Filtering
    - (taken from a class by Bill Freeman @MIT)

- Image Filtering
  - take some local image patch (e.g. 3x3 block)
  - image filtering: apply some function to local image patch



Local image data

Modified image data

Image Filtering

- Some Examples:
  - what assumptions are you making to infer the center value?

3	3	3
3	?	3
3	3	3

3

- Goals of Image Filtering:
  - reduce noise
  - fill-in missing values/ information
  - extract image features (e.g.edges/corners)
  - ...

3	4	3
2	?	5
5	4	2

3 or 4

Image Filtering

- simplest case: linear filtering:
  - replace each pixel by a linear combination of its neighbors

10	5	3
4	5	1
1	1	7

Local image data

0	0	0
0	0.5	0
0	1	0.5

kernel

	7	

Modified image data

- the prescription for the linear combination is called the 'convolution kernel'

2D signals and convolution

- Components of ‘convolution’:
  - Image:
    - continuous: $I(x,y)$
    - discrete: $I[k,l]$ or $I_{k,l}$
  - filter ‘kernel’: $g[k,l]$
  - ‘filtered’ image: $f[m,n]$

- 2D convolution (discrete):

$$f[m,n] = I \otimes g = \sum_{k,l} I[m-k, n-l] g[k,l]$$

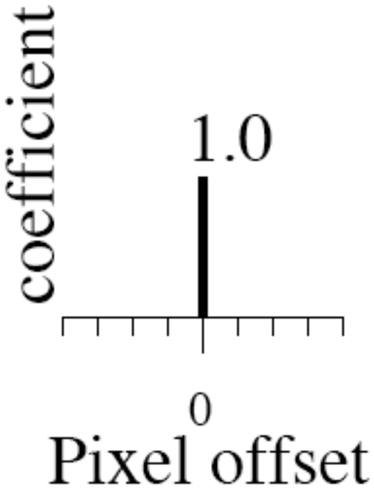
- special case:
  - convolution (discrete) of a 2D-image with a 1D-filter

$$f[m,n] = I \otimes g = \sum_{k} I[m-k, n] g[k]$$

Linear Filtering (warm-up slide)

$$f[m,n] = I \otimes g = \sum_k I[m-k,n]g[k]$$

original



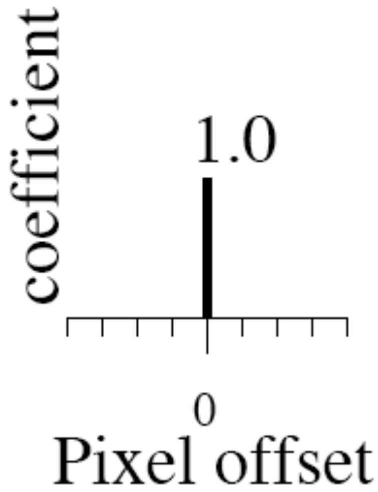
$= ?$

$I \qquad \otimes \qquad g \qquad = f$

Linear Filtering (warm-up slide)

$$f[m,n] = I \otimes g = \sum_k I[m-k,n]g[k]$$

original



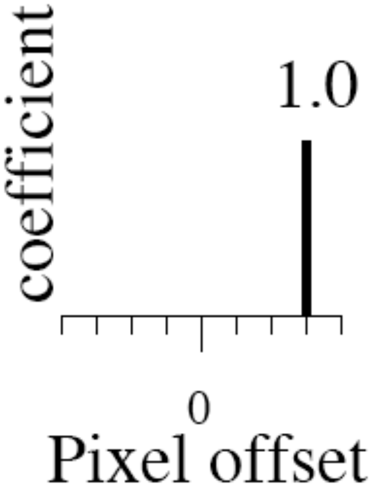
Filtered
(no change)

$I \qquad \otimes \qquad g \qquad = f$

Linear Filtering

$$f[m,n] = I \otimes g = \sum_k I[m-k,n]g[k]$$

original

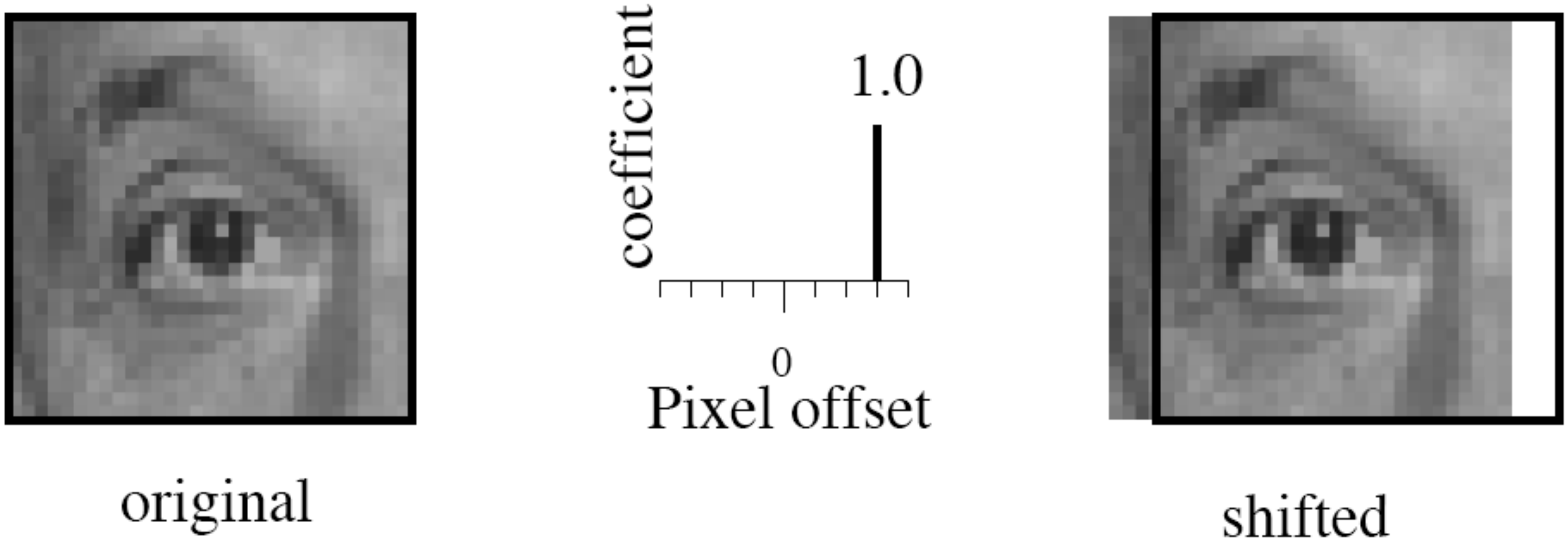


?

$I \quad \otimes \quad g \quad = f$

Linear Filtering

$$f[m,n] = I \otimes g = \sum_k I[m-k,n]g[k]$$

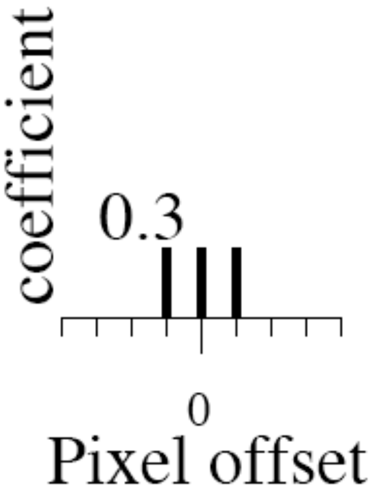


$I \qquad \otimes \qquad g \qquad = f$

Linear Filtering

$$f[m,n] = I \otimes g = \sum_k I[m-k,n]g[k]$$

original



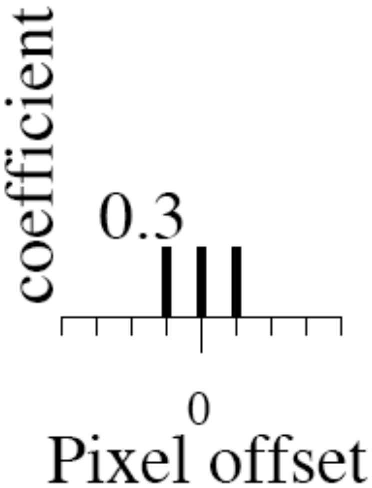
?

$I \quad \otimes \quad g \quad = f$

Blurring

$$f[m,n] = I \otimes g = \sum_k I[m-k,n]g[k]$$

original



Blurred (filter applied in both dimensions).

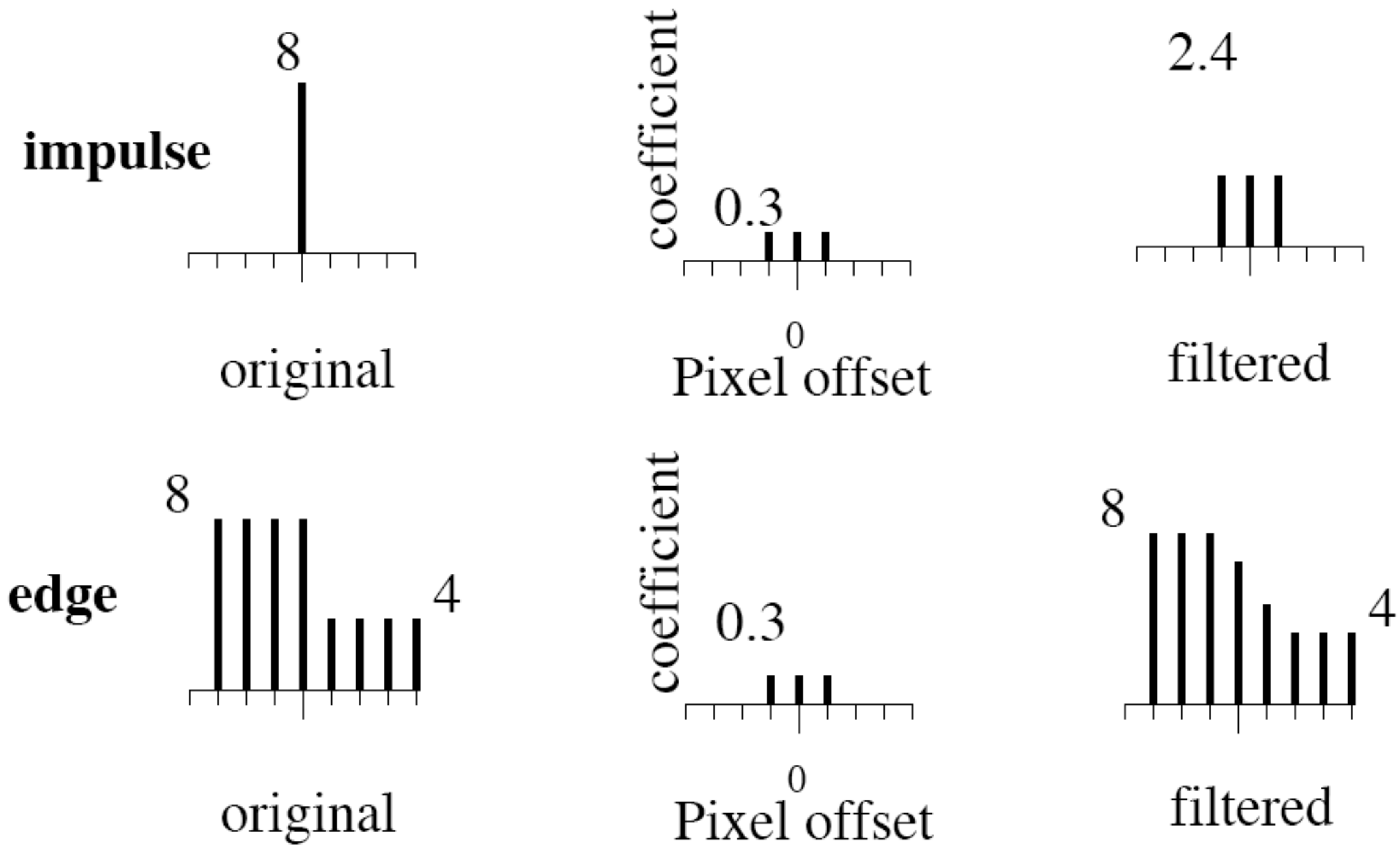
$I \otimes g$

Try it out in GIMP

- You can try out linear filter kernels in the free image manipulation tool GIMP - availble at gimp.org

- open image
- from the menu pick:
  - Filters
    - Generic
      - Convolution Matrix ...
- enter filter kernel in “Matrix”
- press “ok” to apply

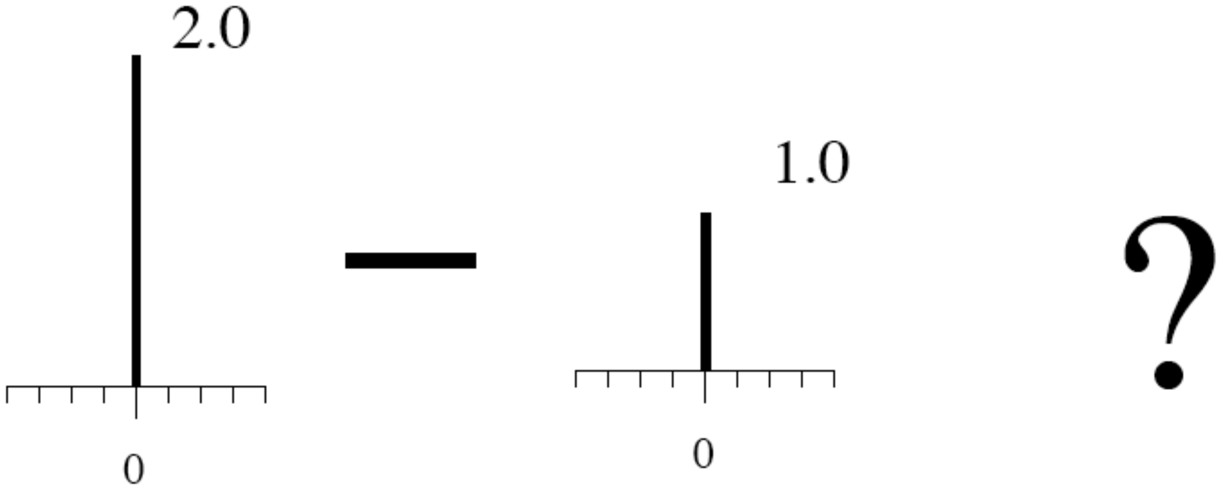
Blurring Examples



Linear Filtering (warm-up slide)

$$f[m,n] = I \otimes g_1 - I \otimes g_2 = I \otimes (g_1 - g_2)$$

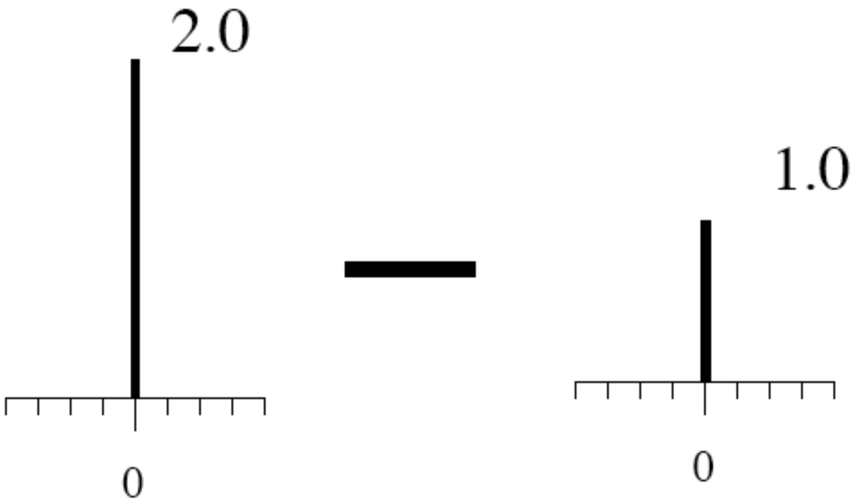
original



Linear Filtering (warm-up slide)

$$f[m,n] = I \otimes g_1 - I \otimes g_2 = I \otimes (g_1 - g_2)$$

original

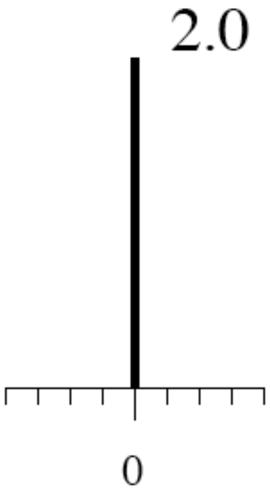


Filtered
(no change)

Linear Filtering

$$f[m,n] = I \otimes g_1 - I \otimes g_2 = I \otimes (g_1 - g_2)$$

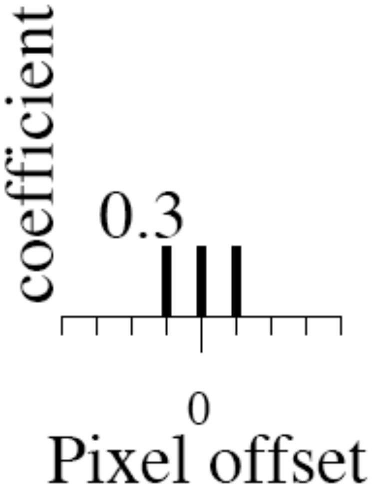
original



−

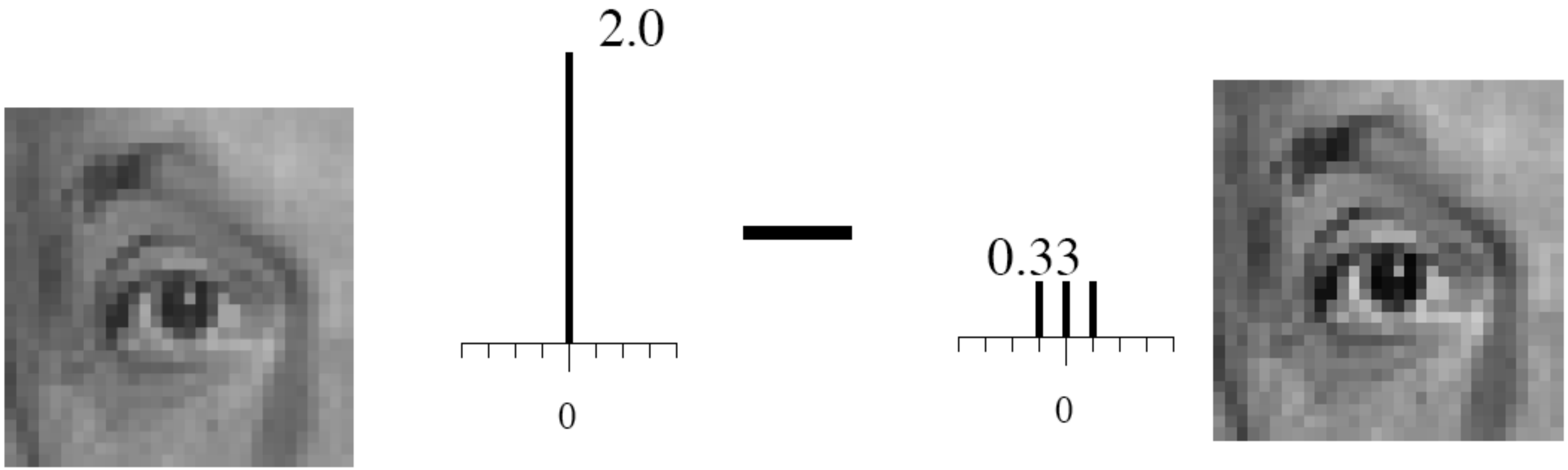
?

original



Blurred (filter applied in both dimensions).

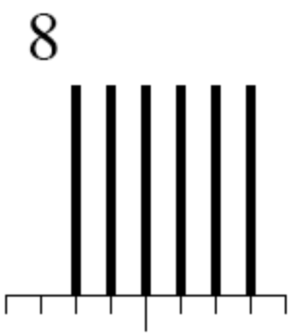
Sharpening



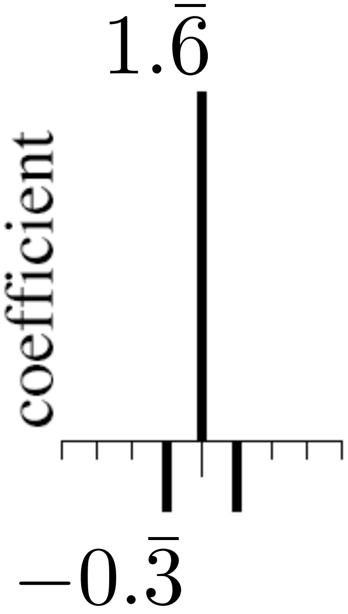
original

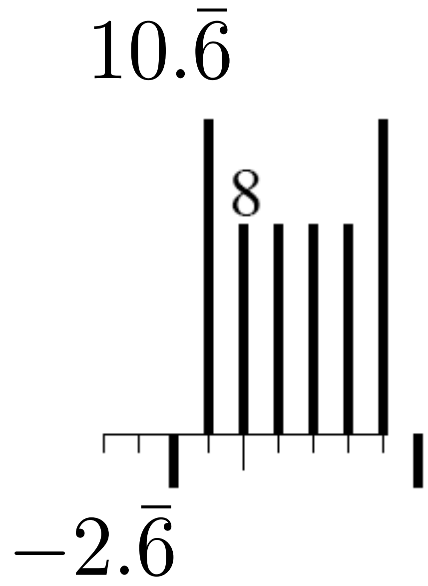
Sharpened original

Sharpening Example



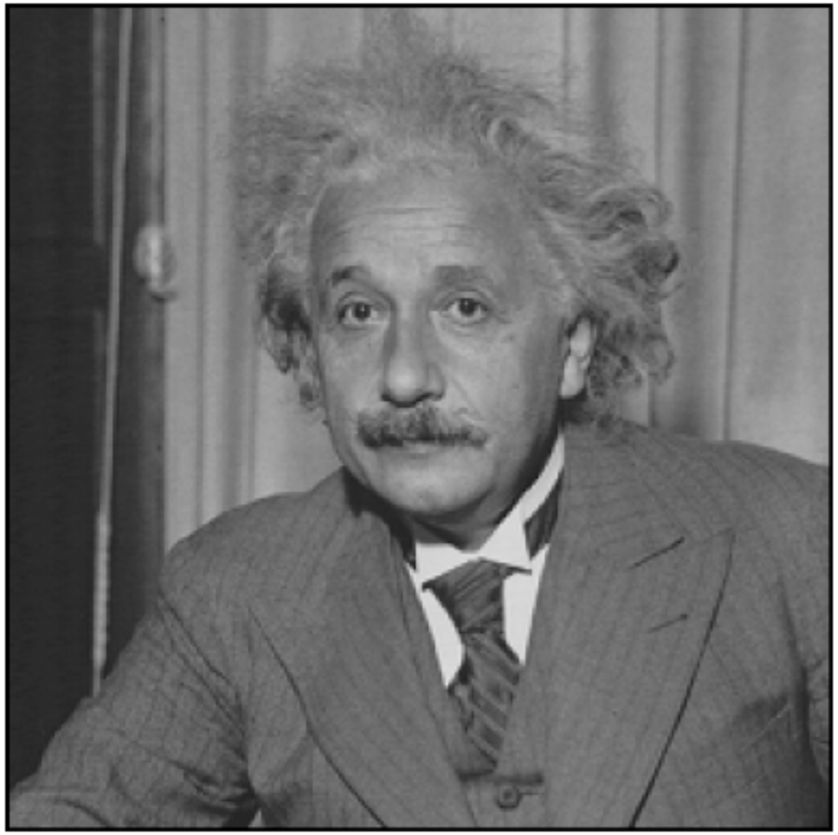
original



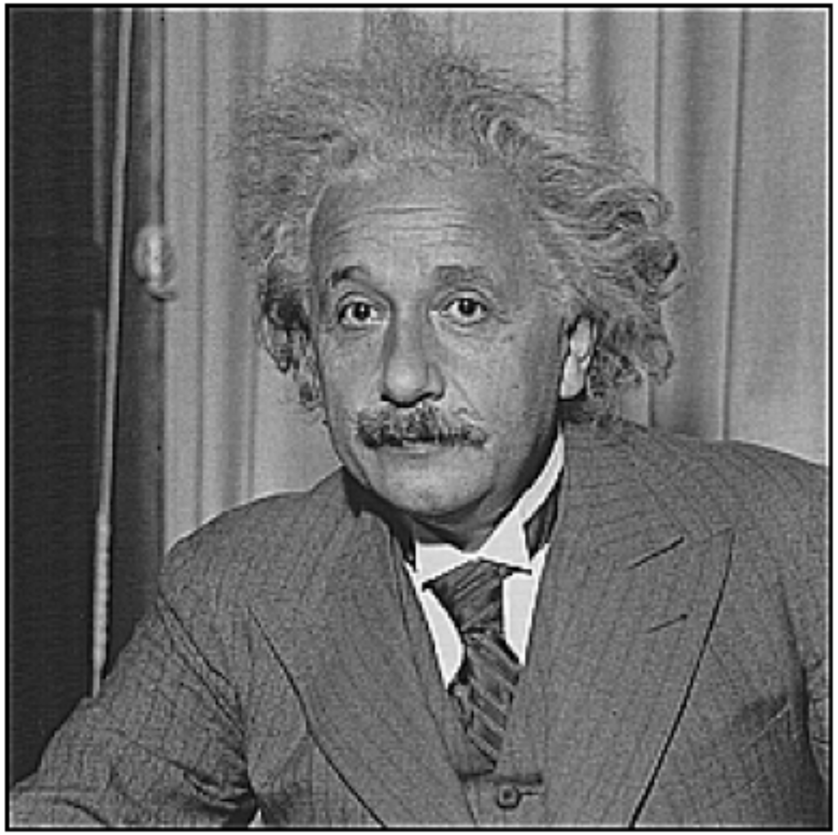


Sharpened
(differences are
accentuated; constant
areas are left untouched).

Sharpening



before



after