Computer Vision I - *Introduction & Image Processing*

Carsten Rother

01/11/2016

Admin Stuff

- Language: German/English; Slides: English
  (all the terminology and books are in English)
- Lecturer: Carsten Rother
- Exercises: Dmitri Schlesinger, Holger Heidrich, Alex Krull
- Staff Email: dmytro.shlezinger@tu-dresden.de
- Announcements: on our webpage

- Course Books:
  - For Image Processing and Geometry:
    Computer Vision: Algorithms and Applications
    by Rick Szeliski; Springer 2011. An earlier version of the book is online:
    http://szeliski.org/Book/
  - Geometry:
    Multiple View Geometry; Hartley and Zisserman;
    Cambridge Press 2004. Second edition. Parts of book are online:
    http://www.robots.ox.ac.uk/~vgg/hzbook/
  - Also pointers to conference and journal articles

Course Overview (total 14 lectures)

Lecture 1 (21.10): Introduction & Image Processing
Ex1: Intro to OpenCV

Lecture 2 (28.10): Image Processing
Ex2: Intro to exercise: Filtering

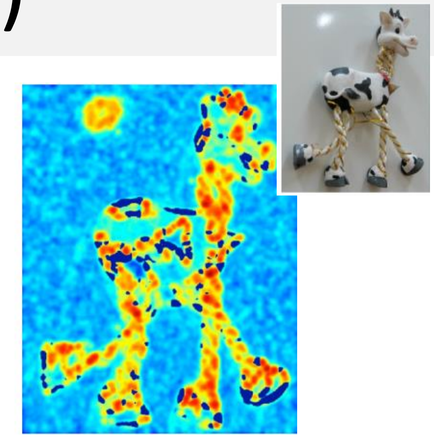
Lecture 3 (4.11): Image Formation Process
Ex3: homework

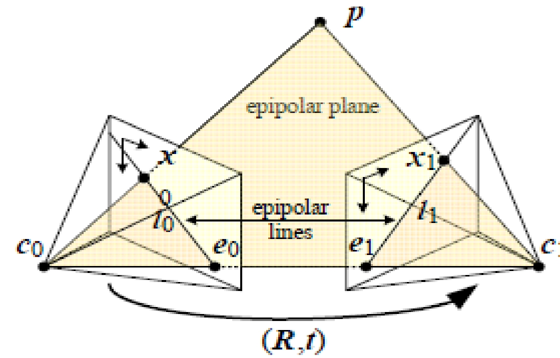
Lecture 4 (11.11): Geometry
Ex4: homework

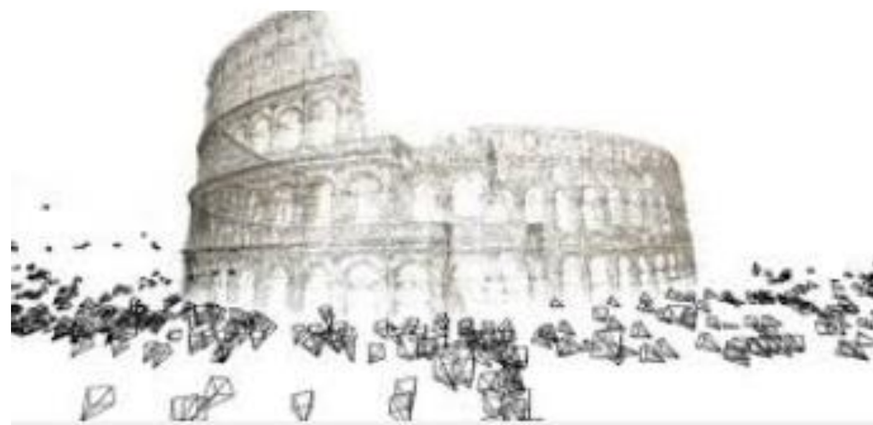
Lecture 5 (16.11 **NOT** 18.11): Geometry
Ex5: Intro to exercise: Panoramic Stitching (Geometry)

Lecture 6 (25.11): Geometry
Ex6: homework

Lecture 7 (2.12): Geometry
Ex7: homework







Course Overview (total 14 lectures)

Lecture 8 (11.12): Object Tracking
Ex8: Intro to exercise: Object Tracking

Lecture 9 (18.12): Object Tracking
Ex9: homework

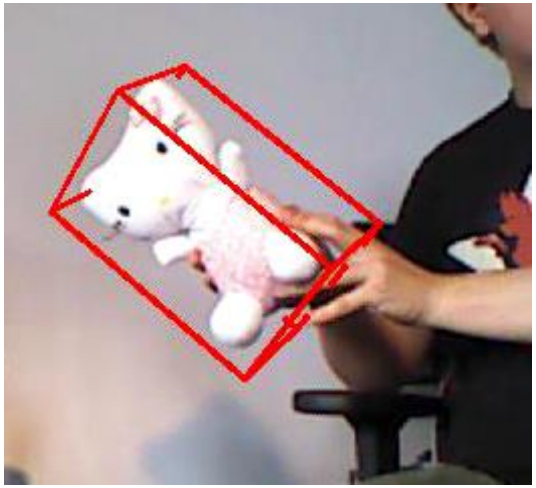
Lecture 10 (8.1): Object Tracking
Ex10: homework

Lecture 11 (15.1): Dense Matching – Stereo
Ex11: homework

Lecture 12 (22.1): Dense Matching – Stereo
Ex12: homework

Lecture 13 (29.1): Dense Matching – Flow
Ex13: Prepare Poster Session

Lecture 14 (5.2): Poster Session
Ex14: homework



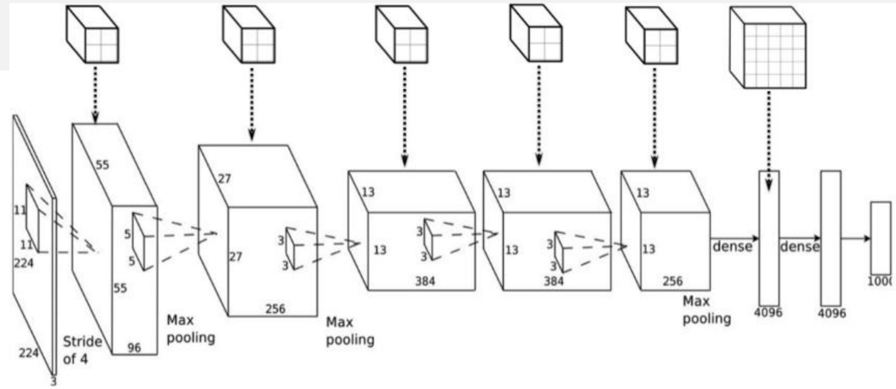
Exams and Exercises

- Exam: in Person

- Exercises/homework:
    - There are three blocks
    - In each block you have to do a certain amount of exercises
    - The exercises should to be handed-in until end of semester (ideally after each block)

- Exam:
    - Exercises are not mandatory for the exam, i.e. you can sit the exam without having done the exercises
    - In the exam I may ask questions about the exact exercise you have done. If you have not done any exercise then this may result in a worse mark

CVLD Lectures



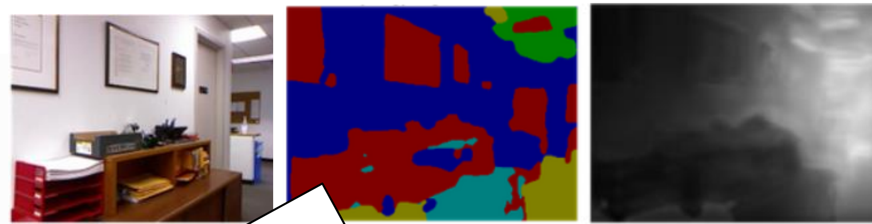
- **<u>WS:</u>**
  - Computer Vision 1 (2+2)
  - Machine Learning 1 (2+2)

- **<u>SS:</u>**
  - Computer Vision 2 (2+2)
  - Machine Learning 2 (2+2)

- For doing a Master/PhD in the CVLD one should do the computer vision or machine learning track

- Computer Graphics (Prof. Gumhold) (Introduction, I, II)
  3D Scanning with structured light; Illumination models; Geometry

Before we start … some Advertisement



Optimization & Learning

with a focus on graphical models, deep models, and large-scale optimization.

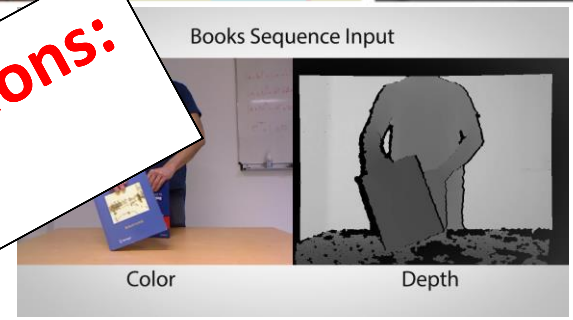
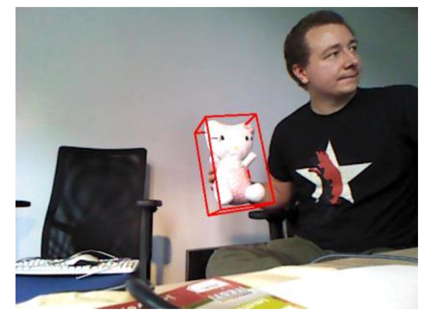


Image Matching

with a focus on 3-9DoF scene f…
jointly recovering multip…

Thesis and Practical Sessions:
Looking for People !

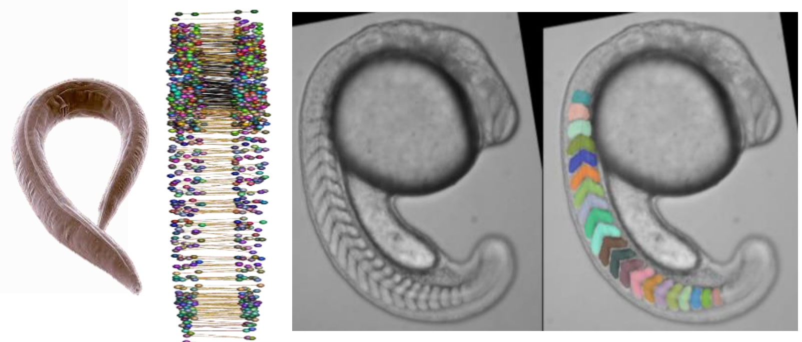


Scene…

with a focus … …ntic scene under-
standing, and …del-based vision.

Bio Imaging

with a focus on segmentation, tracking, and matching – jointly with teams from MPI-CBG.



Your Future in Computer Vision

A project work in the CVLD is a good stepping stone if you:

- want to do a PhD in computer vision, graphics, machine learning
- Becoming a researcher or software developer in a research lab (Facebook, Google, Microsoft Research, Daimler, Bosch, Adobe, TechniColor, etc etc)
- If you are interested in doing a Start-up
- Other “computer vision related” industry

Introduction to Computer Vision

What is computer Vision?

(Potential) Definition:
Developing computational models and algorithms to interpret digital images and visual data in order to understand the visual world we live in.

Introduction to Computer Vision

What is computer Vision?

(Potential) Definition:
Developing computational models and algorithms to interpret digital images and visual data in order to understand the visual world we live in.

What does it mean to “understand”?

(Potential) Definition:

Developing computational models and algorithms to interpret digital images and visual data in order to understand the visual world we live in.

Physics-based vision:

Geometry

Segmentation

Camera parameters

Emitted light (sun)

Surface properties: Reflectance, material

Semantic-based vision:

Objects: class, pose

Scene: outdoor,…

Attributes/Properties:

- old-fashioned train
- A-on-top-of-B

Image-formation model

Very many sources of variability

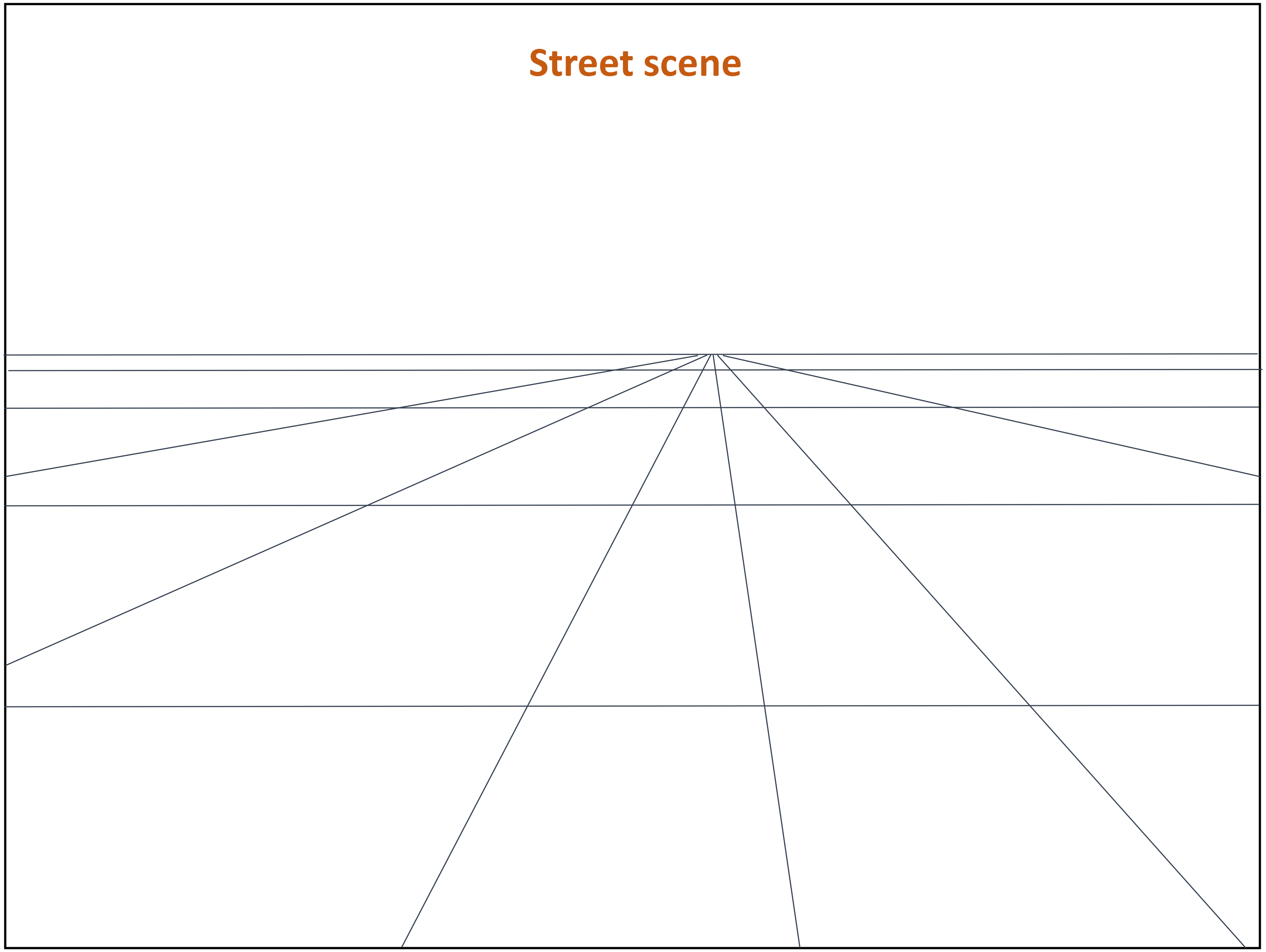
Image

[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

Scene geometry



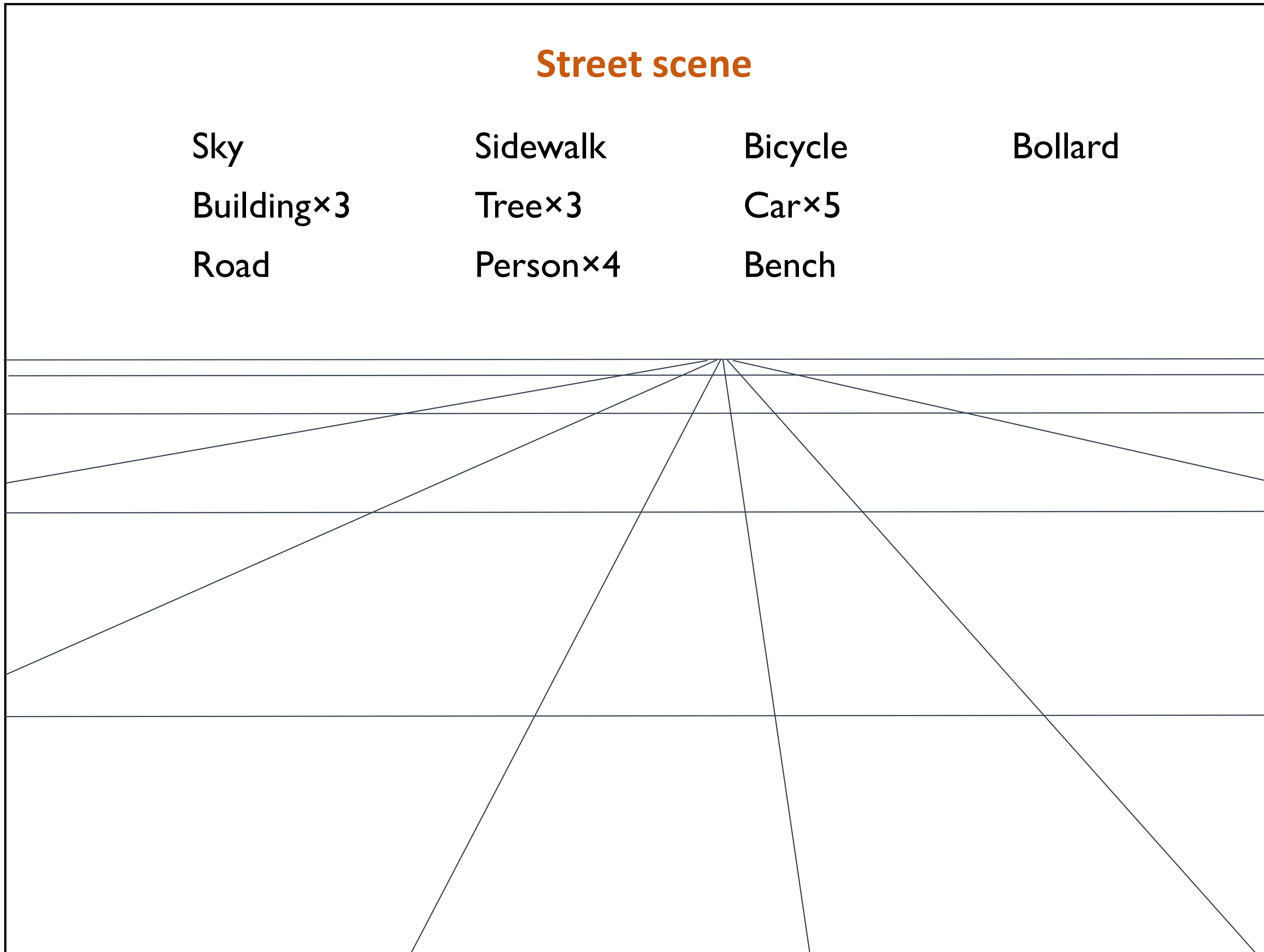
[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

Scene geometry

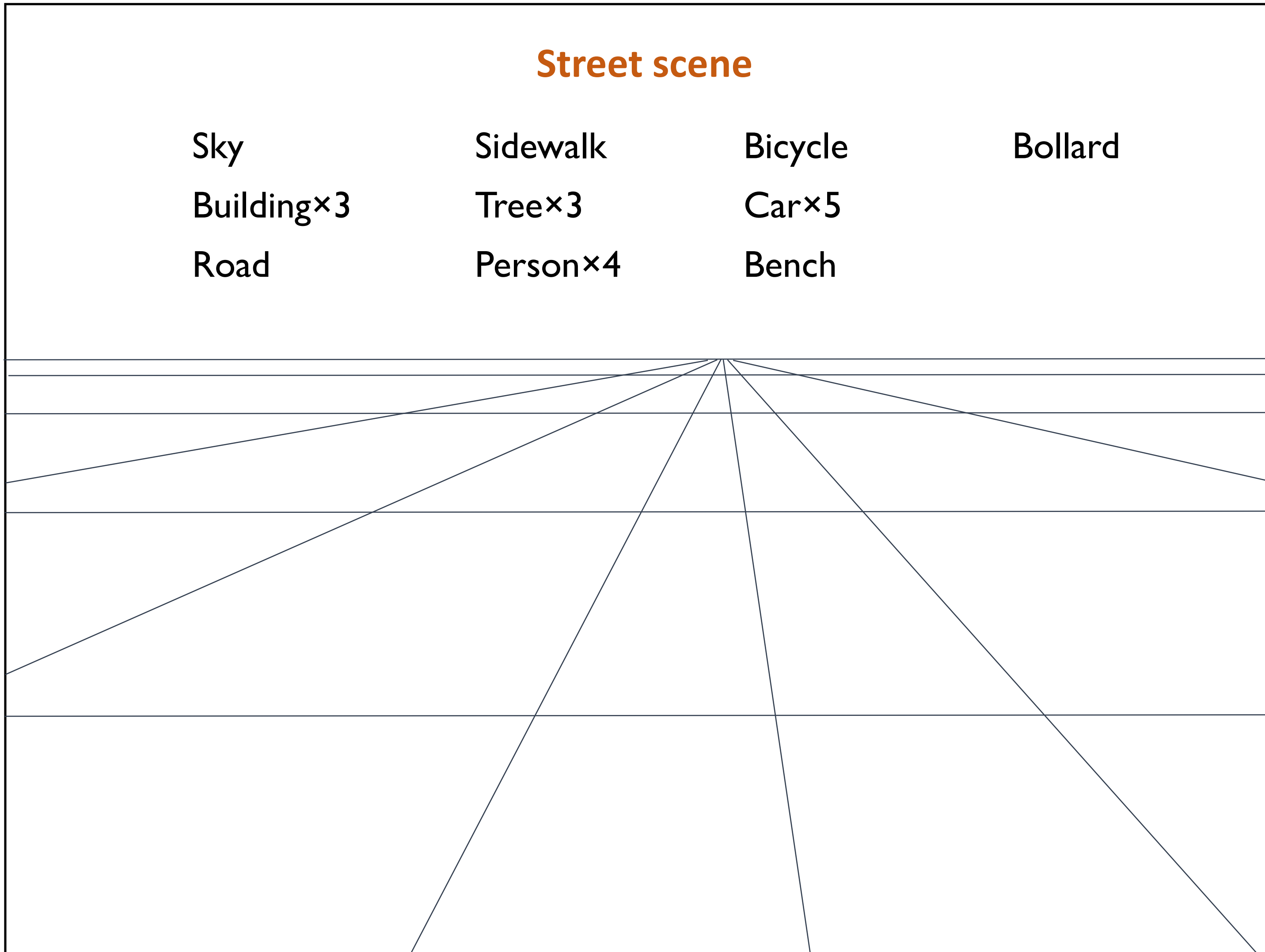
Object classes



[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type
Scene geometry
Object classes
Object position
Object orientation



[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

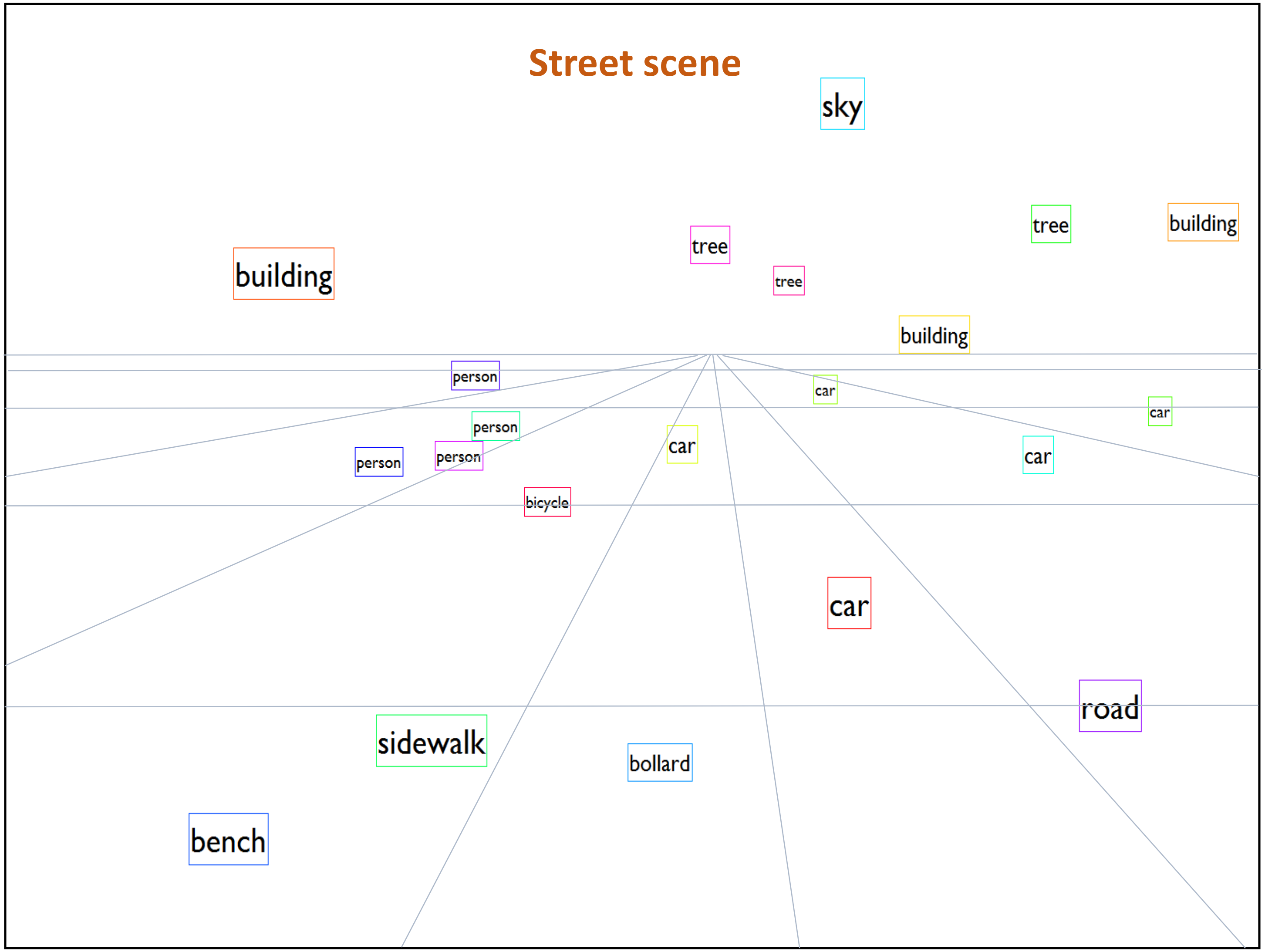
Scene geometry

Object classes

Object position

Object orientation

Object shape



[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

Scene geometry

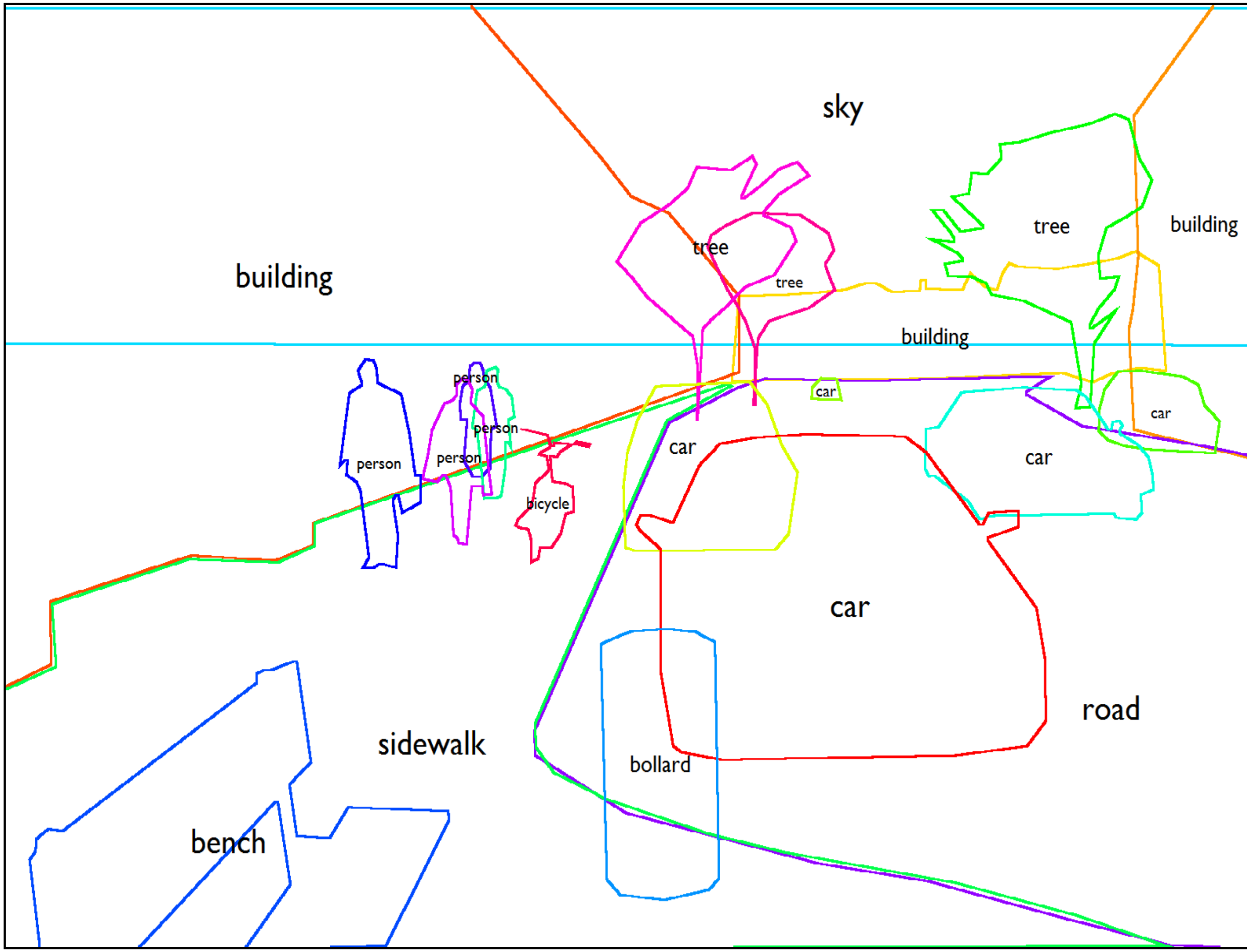
Object classes

Object position

Object orientation

Object shape

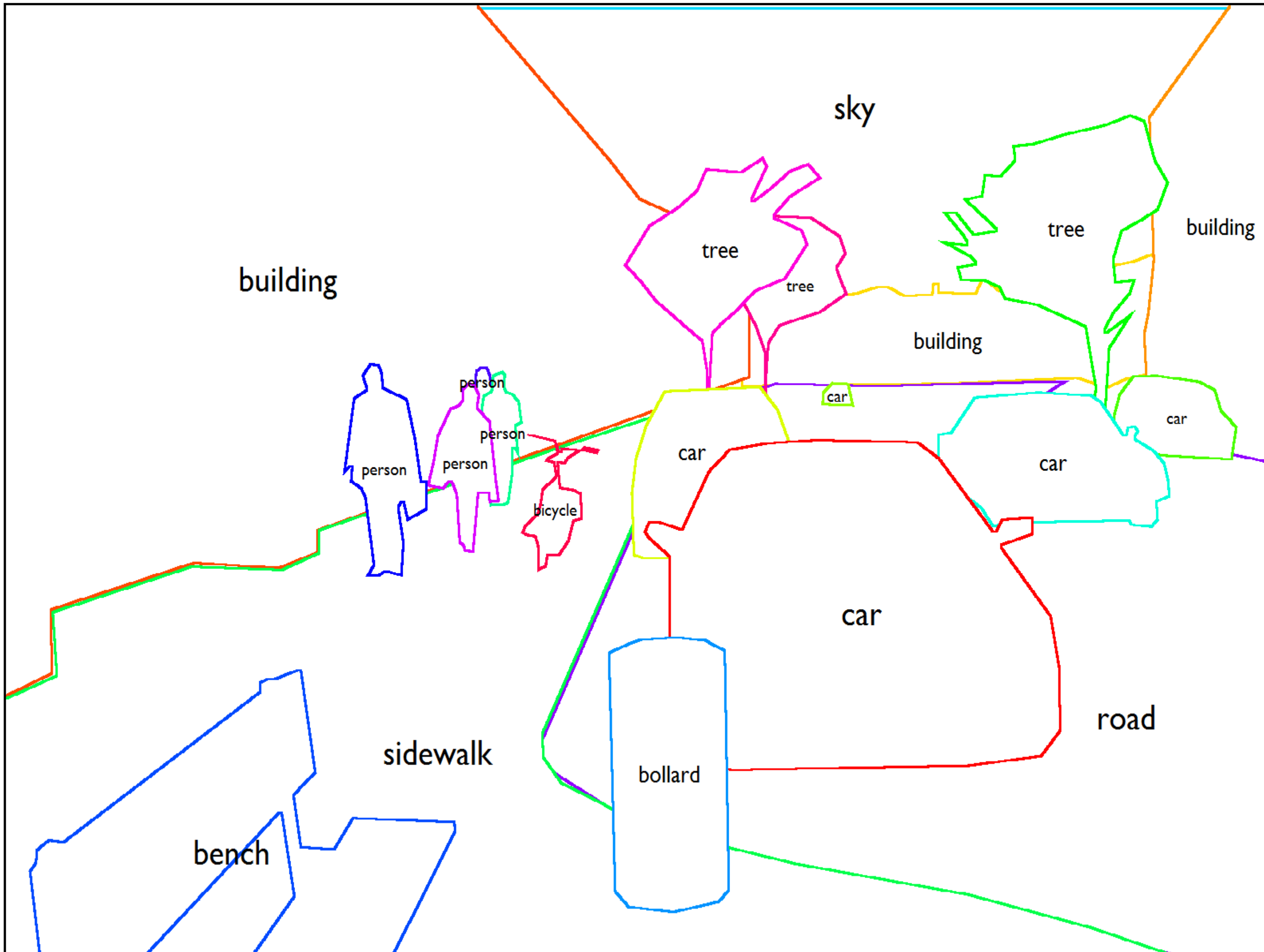
Depth/occlusions



[Slide Credits: John Winn, ICML 2008]

Image-formation model

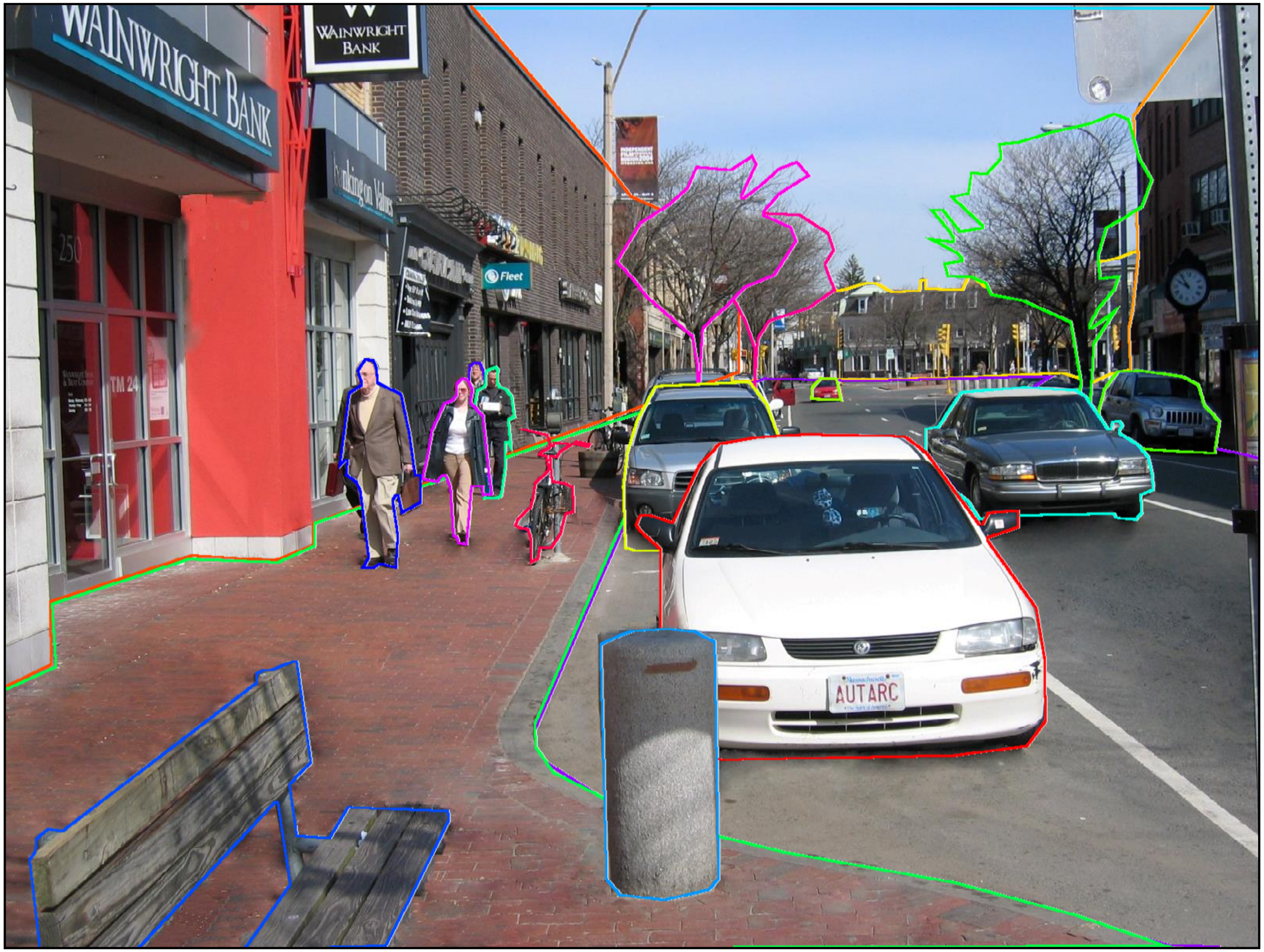
Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape
Depth/occlusions
Object appearance



[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape
Depth/occlusions
Object appearance
Illumination
Shadows



[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

Scene geometry

Object classes

Object position

Object orientation

Object shape

Depth/occlusions

Object appearance

Illumination

Shadows

[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

Scene geometry

Object classes

Object position

Object orientation

Object shape

Depth/occlusions

Object appearance

Illumination

Shadows

Motion blur

Camera effects

[Slide Credits: John Winn, ICML 2008]

Image-formation model

Scene type

Scene geometry

Object classes

Object position

Object orientation

Object shape

Depth/occlusions

Object appearance

Illumination

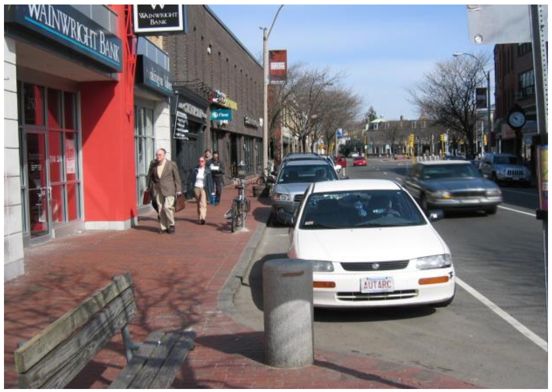
Shadows

Motion blur

Camera effects

[Slide Credits: John Winn, ICML 2008]

The "Scene Parsing" challenge --- a "grand challenge" of computer vision



Single image

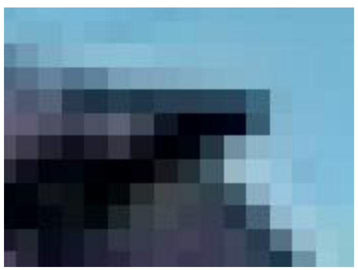
(Probabilistic) Script = {Camera, Light, Geometry, Material, Objects, Scene, Attributes, Others}

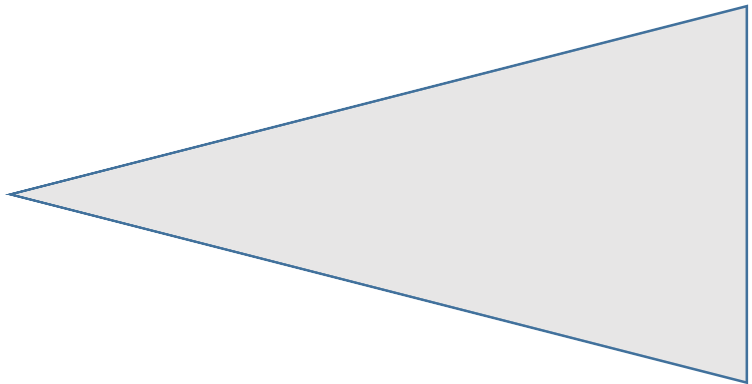
Many applications do not have to extract the full probabilistic script but only a subset, e.g. "does the image contain a car?"

… many examples to come later

Why is "scene parsing" hard?

Computer Graphics





3D Rich Representation,

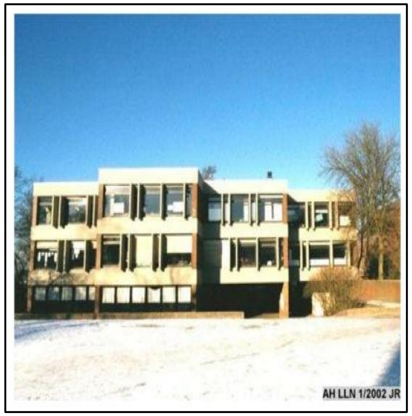
Script = {Camera, Light, Geometry, Material, Objects, Scene, Attributes, Others}

Computer Vision

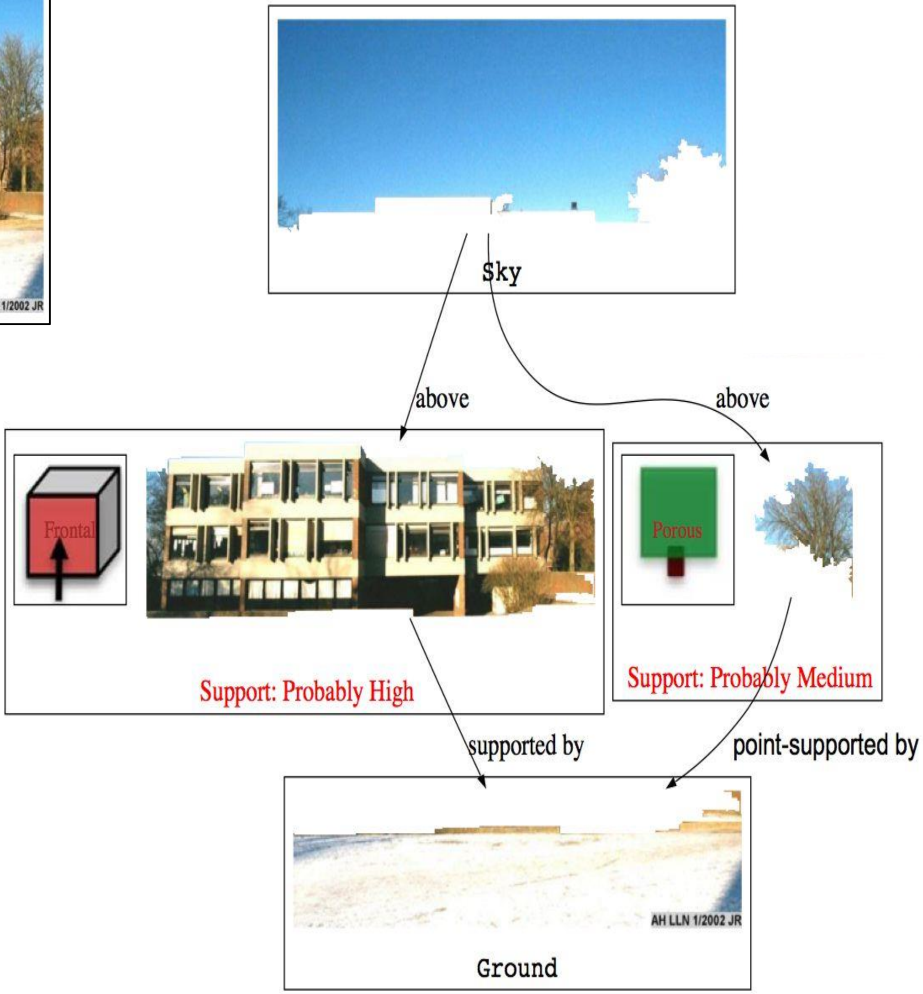
2D pixel representation

Computer Vision can be seen as "inverse graphics"

Example Result



Input



Scene graph

[Gupta, Efros, Herbert, ECCV '10]

Why is “scene parsing” hard?

Subgoal for July

Analysis of scenes consisting of non-overlapping objects from the

balls

bricks with faces of the same or different colors or textures

cylinders.

Each face will be of uniform and distinct color and/or texture.

Extensions for August

The first priority will be to handle objects of the same sort but with complex surfaces and backgrounds, e.g. cigarette pack with writing and bands of different color, or a cylindrical battery.

[Xiao et al. NIPS 2012]

[Sussman, Lamport, Guzman 1966]

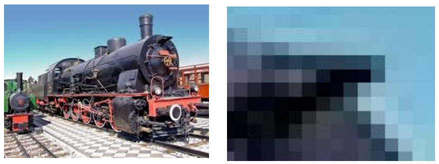
[Slide credits Andrew Blake]

Introduction to Computer Vision

What is computer Vision?

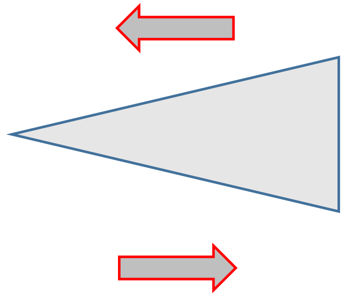
(Potential) Definition:
Developing computational models and algorithms to interpret digital images and visual data in order to understand the visual world we live in.

How can we interpret visual data?



2D pixel representation

Computer Graphics



Computer Vision

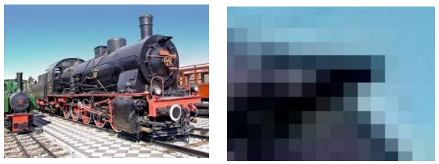
3D Rich Representation,

Script = {Camera, Light, Geometry, Material, Objects, Scene, Attributes, Others}

- What general (prior) knowledge of the world (not necessarily visual) can be exploit?
- What properties / cues from the image can be used?

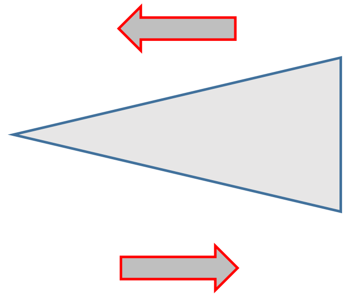
Both aspects are quite well understood (a lot is based on physics) … but how to use them optimally is an open question (see later)

How can we interpret visual data?



2D pixel representation

Computer Graphics



Computer Vision

3D Rich Representation,

Script = {Camera, Light, Geometry, Material, Objects, Scene, Attributes, Others}

- What general (prior) knowledge of the world (not necessarily visual) can be exploit?
- What properties / cues from the image can be used?

Both aspects are quite well understood (a lot is based on physics) … but how to use them optimally is an open question (see later)

Prior knowledge (examples)

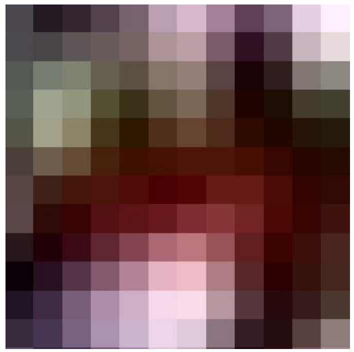
- "Hard" prior knowledge
  - Trains do not fly in the air
  - Objects are connected in 3D

- "Soft" prior knowledge:
  - The camera is more likely 1.70m above ground and not 0.1m.
  - Self-similarity: "all black pixels belong to the same object"

Prior knowledge … often harder to describe

- Describe Image Texture

Real Image

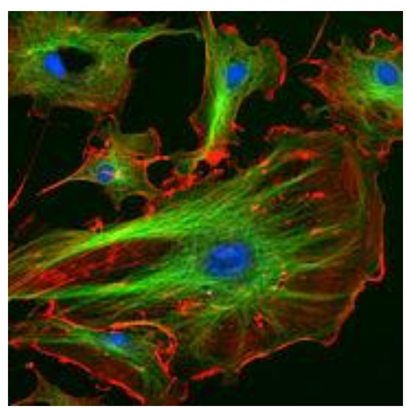


zoom

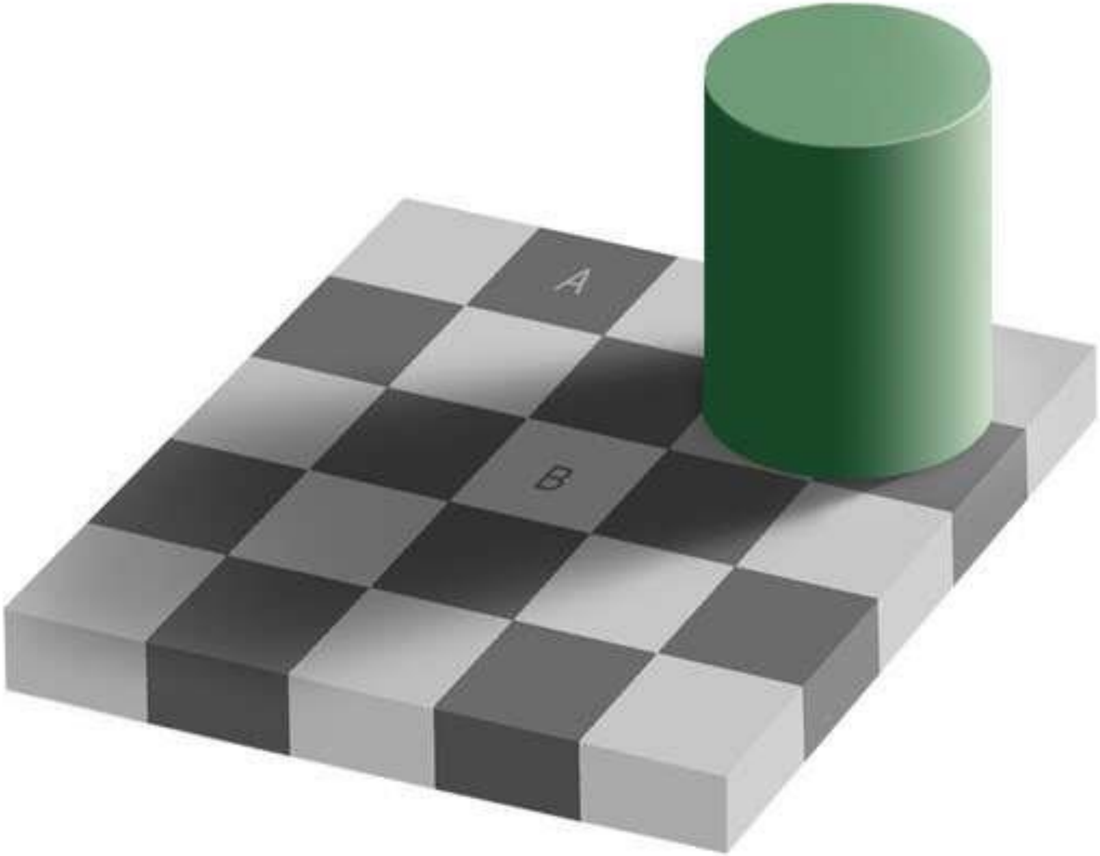
Not a real Image

zoom

- Microscopic Images. What is the true shape of these objects



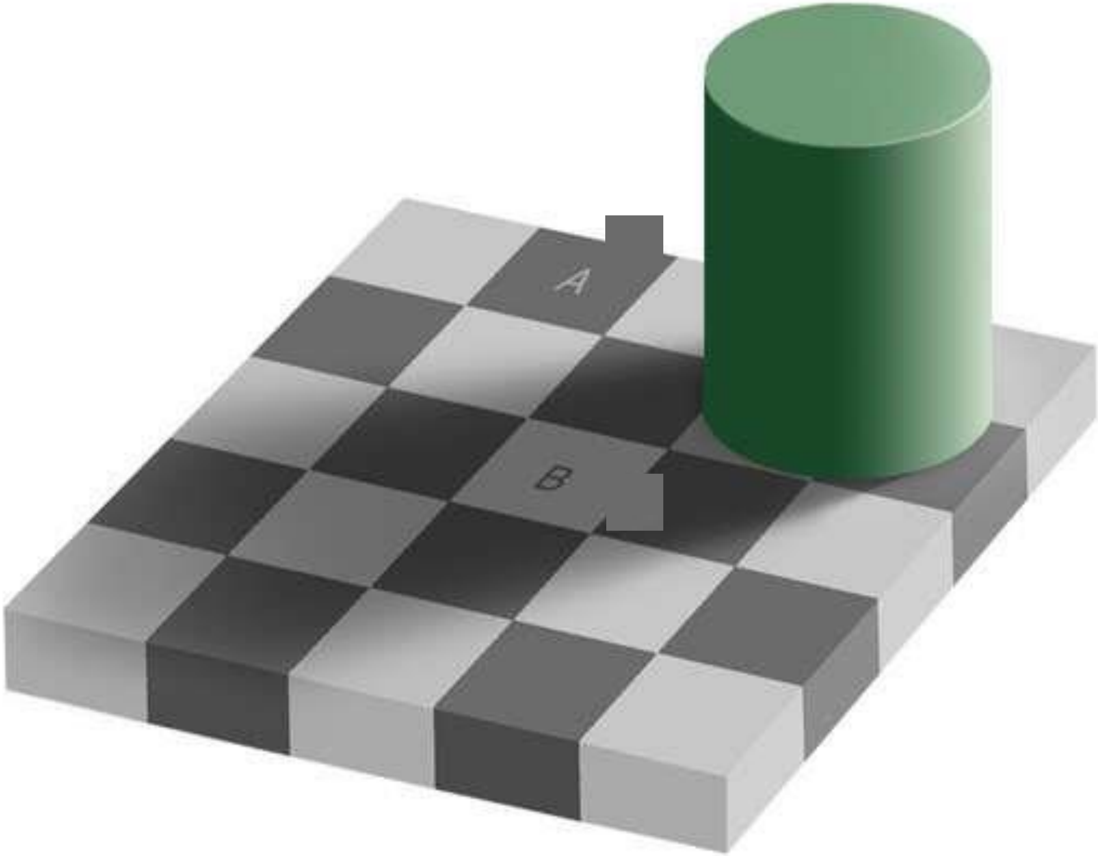
The importance of Prior knowledge



Which patch is brighter: A or B?

[Edward Adelson]

The importance of Prior knowledge

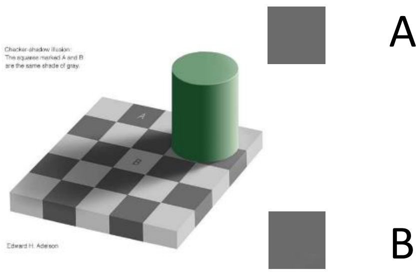


Which patch is brighter: A or B?

[Edward Adelson]

The importance of Prior knowledge

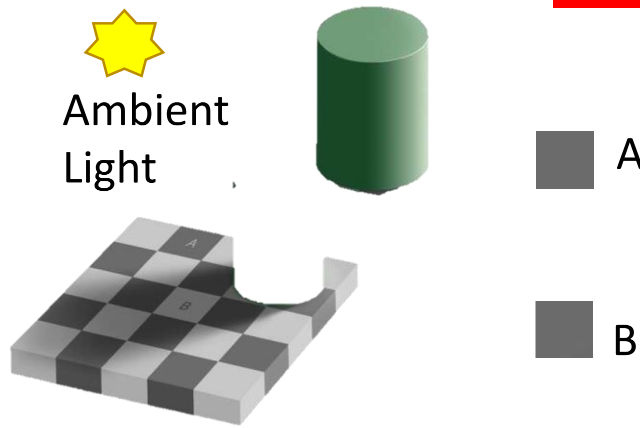
2D



2D Image - local

What the
computer sees

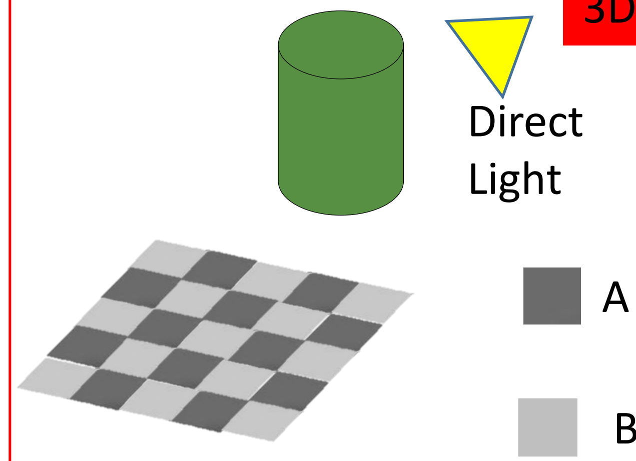
3D



True colors
in 3D world

An unlikely
3D representation
(hard to “see” it likes this,
for a human)

3D

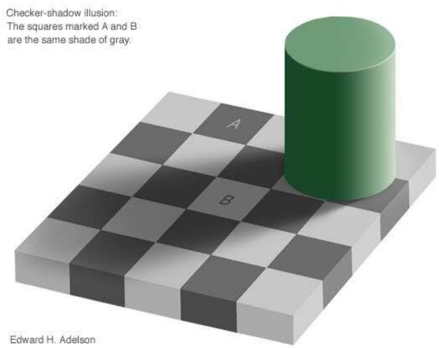


True colours
In 3D world

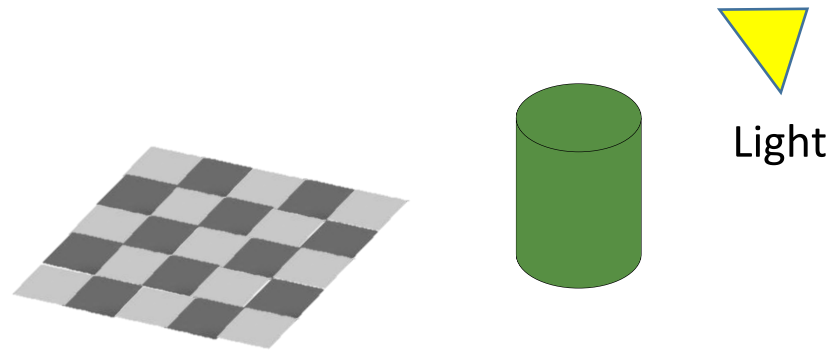
The most likely
3D representation

This is what humans “see” implicitly. Ideally the computer sees the same.

The importance of Prior knowledge



2D Image



3D representation

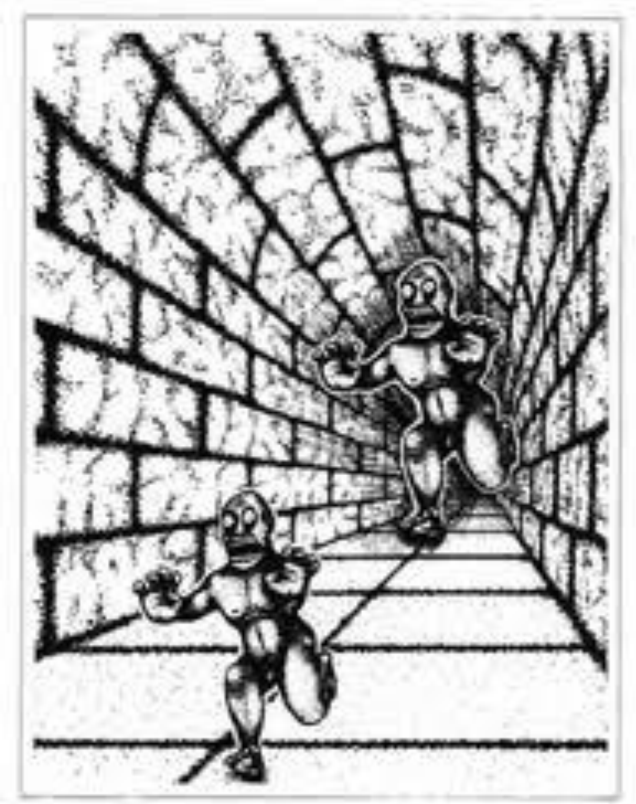
Humans see an image **not** as a set of 2D pixels. They understand an image as a projection of the 3D world we live in.

Humans have the prior knowledge about the world encoded, such as:

- Light cast shadows
- Objects do not fly in the air
- A car is likely to move but a table is unlikely to move

We have to teach the computer this prior knowledge to understand 2D images as picture of the 3D world !

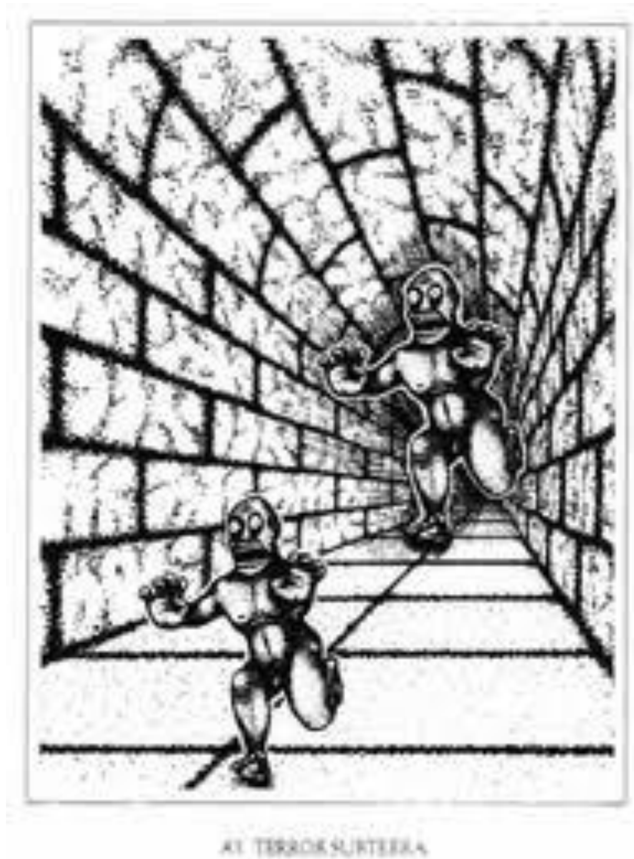
The importance of Prior knowledge



AT TERROR SUBTERRA

Which monster is bigger?

The importance of Prior knowledge



Which monster is bigger?

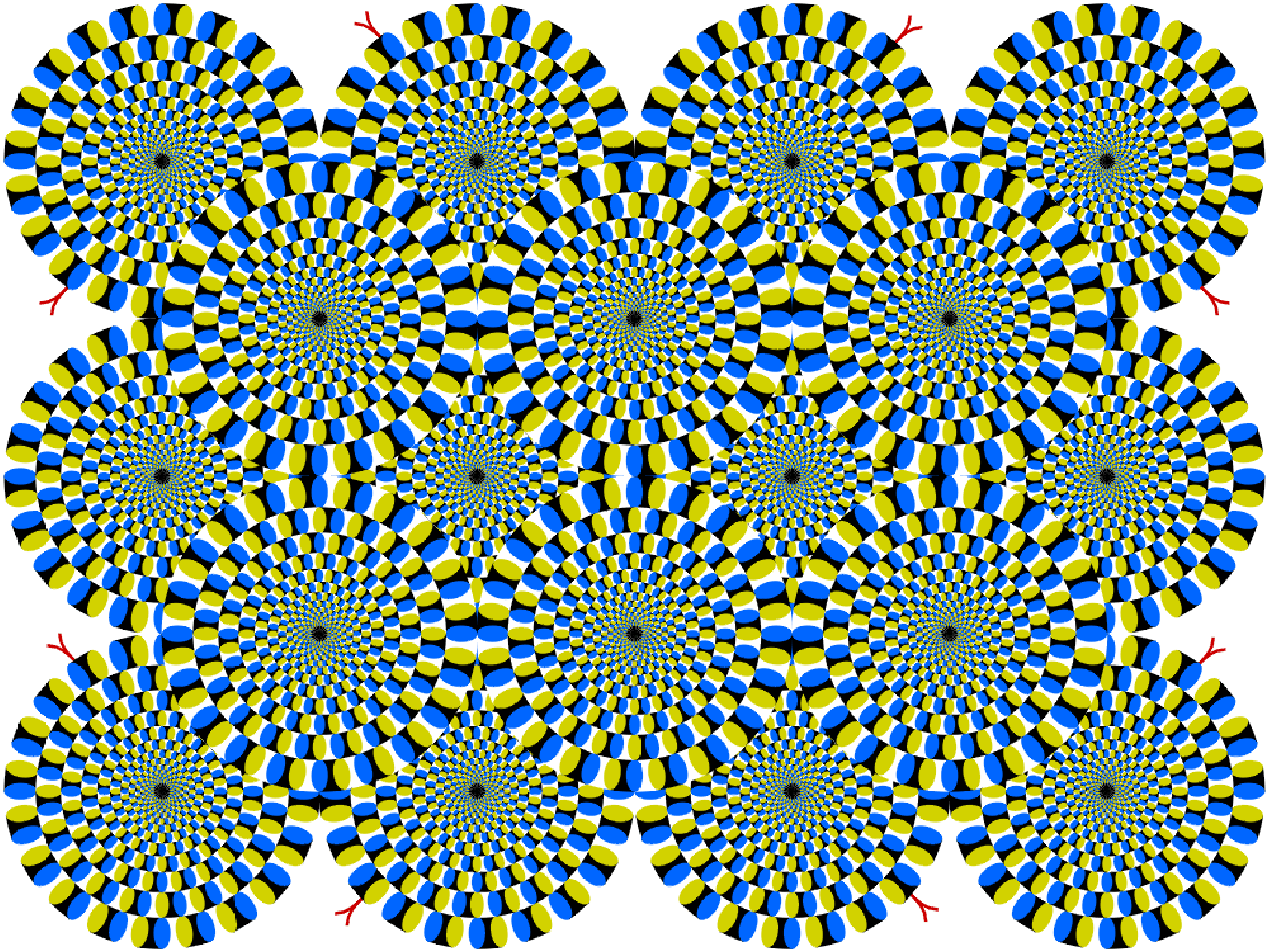
In the 2D Image

1meter

2meter

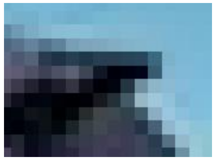
In the 3D world (truth)

Human Vision can be fooled

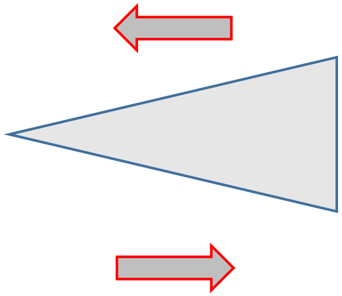


How can we interpret visual data?

Computer Graphics

2D pixel representation



3D Rich Representation,

Script = {Camera, Light, Geometry, Material, Objects, Scene, Attributes, Others}

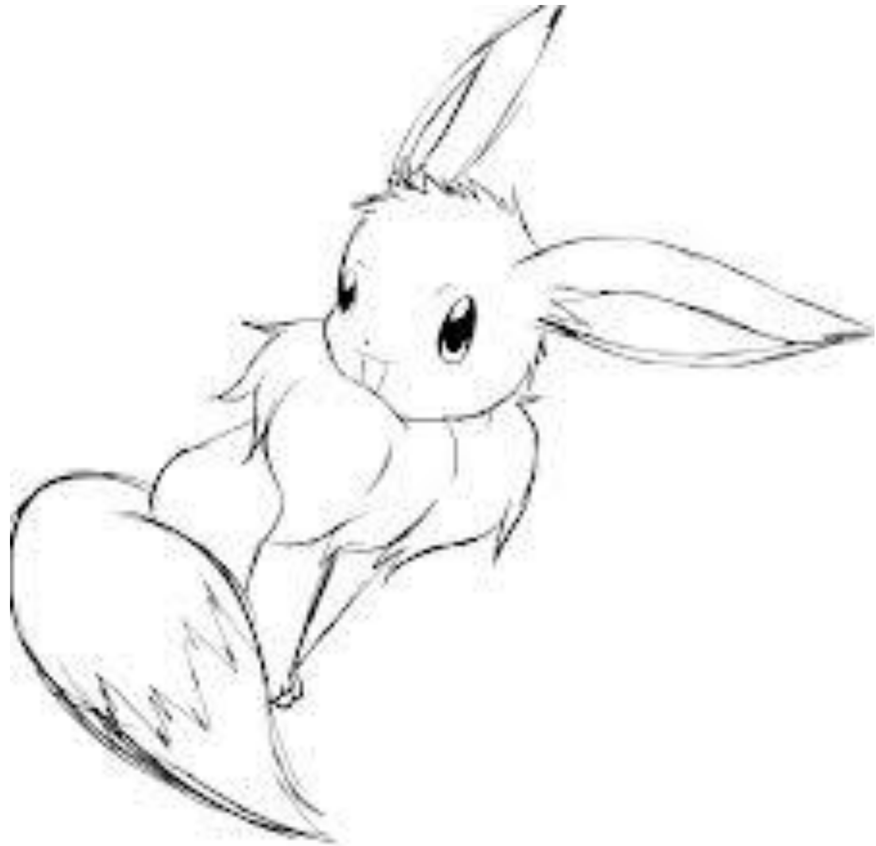
Computer Vision

- What general (prior) knowledge of the world (not necessarily visual) can be exploit?
- What properties / cues from the image can be used?

Both aspects are quite well understood (a lot is based on physics) … but how to use them optimally is an open question (see later)

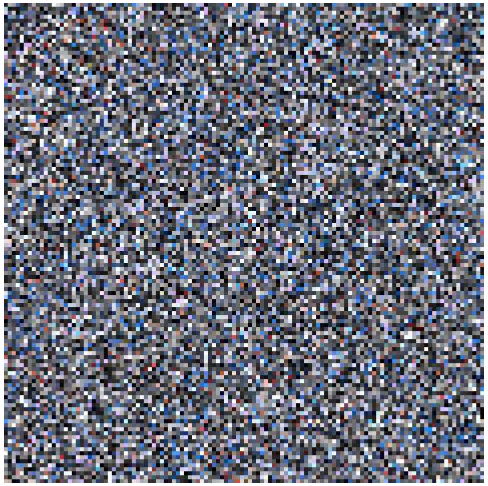
Cue: Appearance (Colour, Texture) for object recognition

Cue: Outlines (shape) for object recognition



Guess the Object

- Colour



- Texture

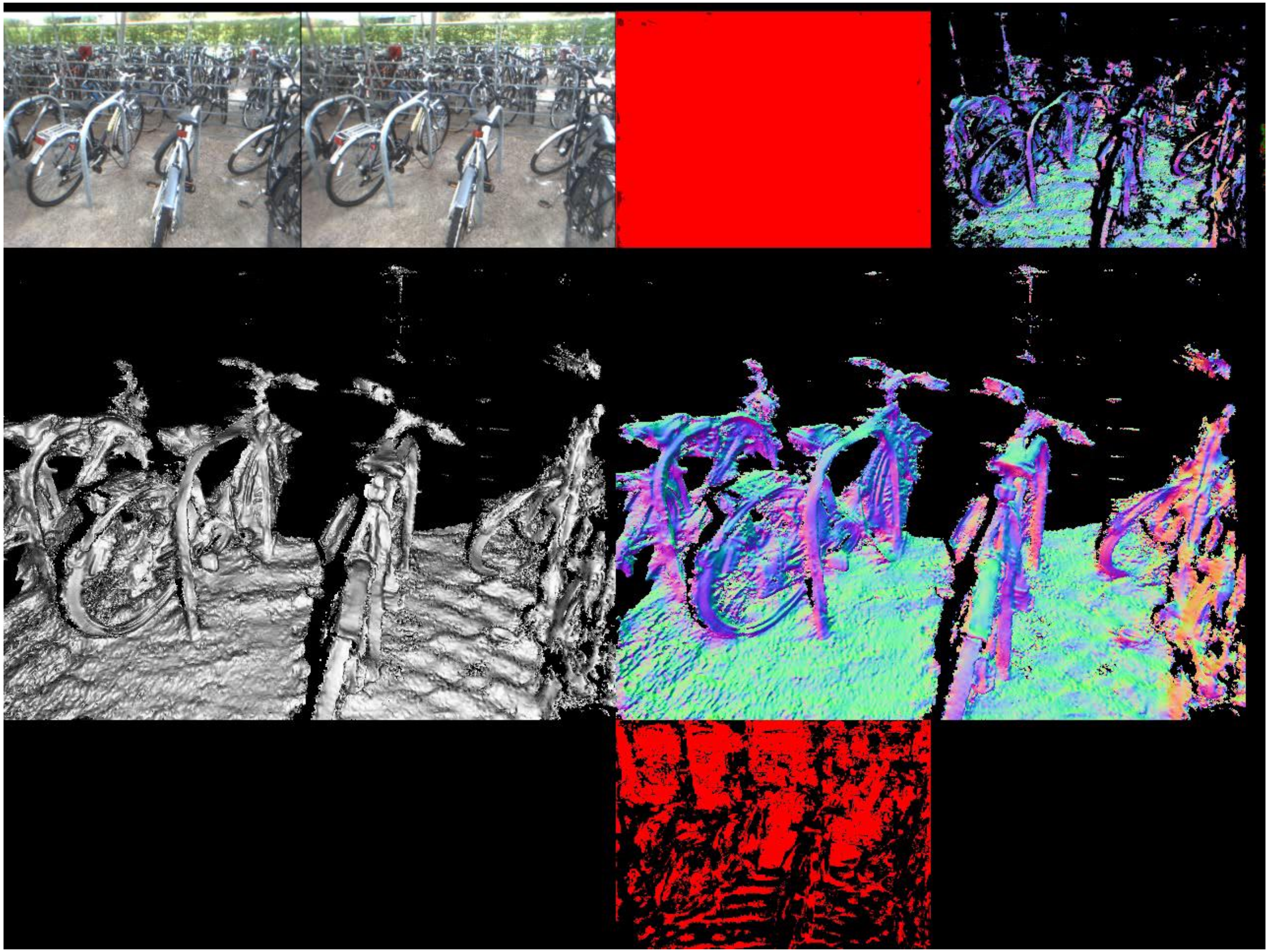
- Shape

[from JohnWinn ICML 2008]

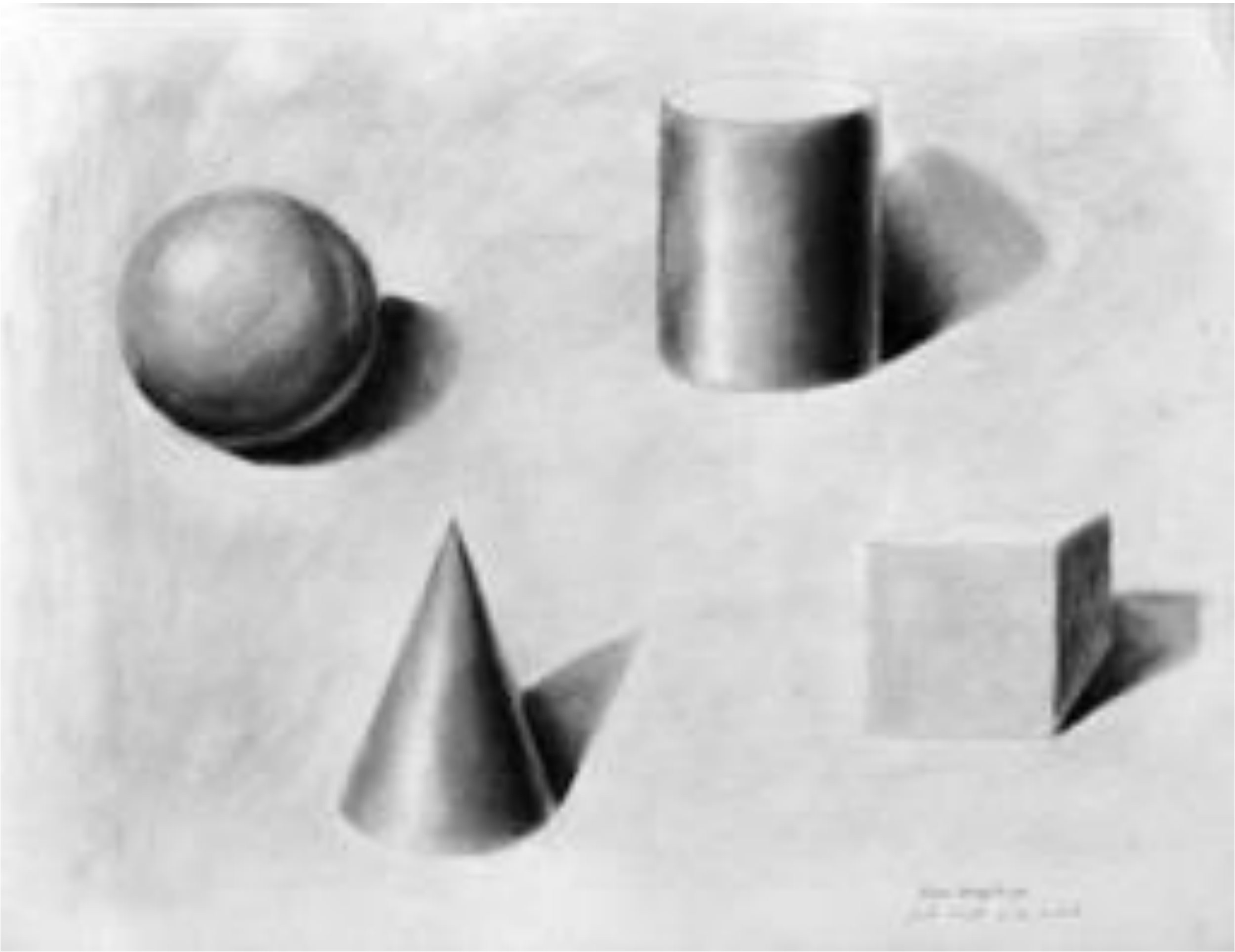
Cue: Context for object recognition

Cue: Context for object recognition

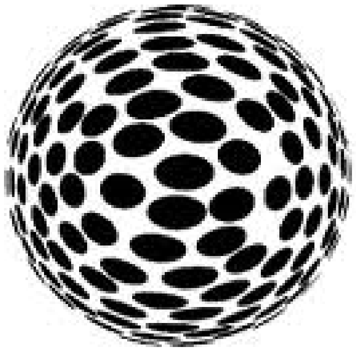
Cue: Multiple Frames for geometry estimation

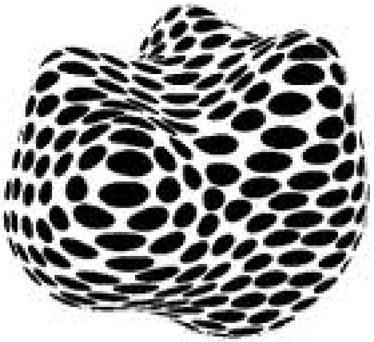


Cue: Shading & shadows for geometry and Light estimation



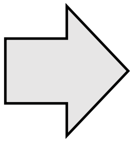
Texture gradient for geometry estimation





The "Scene Parsing" challenge --- a "grand challenge" of computer vision

Single image



(Probabilistic) Script = {Camera, Light, Geometry, Material, Objects, Scene, Attributes, Others}

Many applications do not have to extract the full probabilistic script but only a subset, e.g. "does the image contain a car?"

… many examples to come later

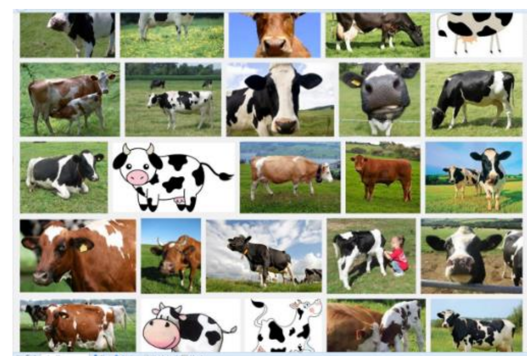
Many application scenarios are now in reach

To simplify/tackle the problem:

1) **Richer Input:**
Modern sensing technology; Video;
Cameras everywhere

2) **Rich Models**: Deep Learning

3) **Lots of Data** to learn from:
search engines; crowdsourcing;
graphics engines



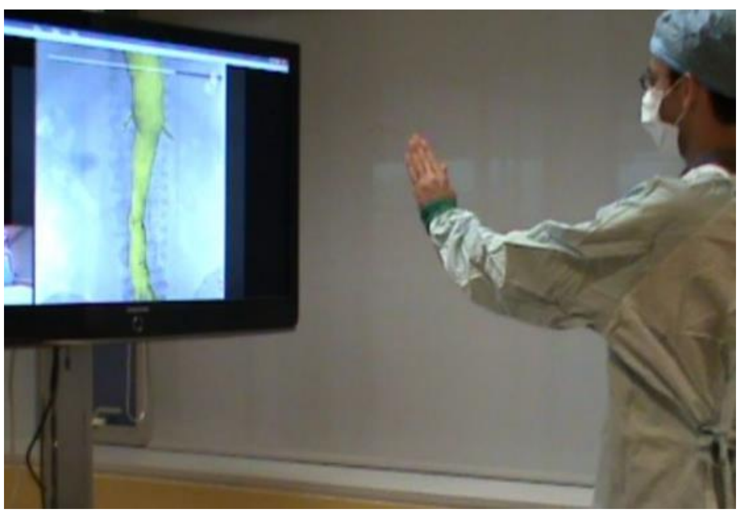
4) For **many practical** applications:
We do not have to infer the full probabilistic script

Hololens

Animate the world

[Chen et al. UIST '12]

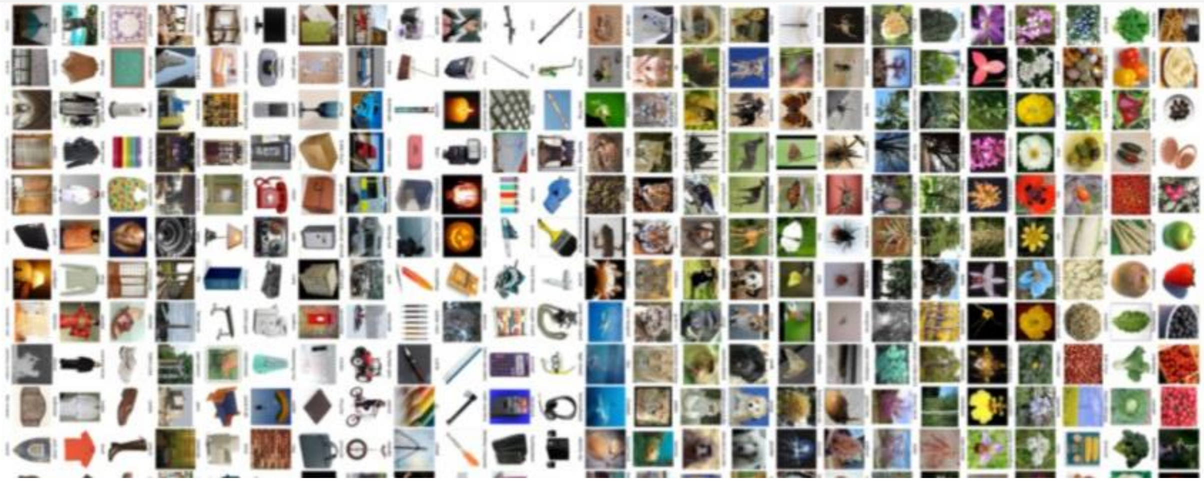
Kinect Body tracking and Gesture Recognition

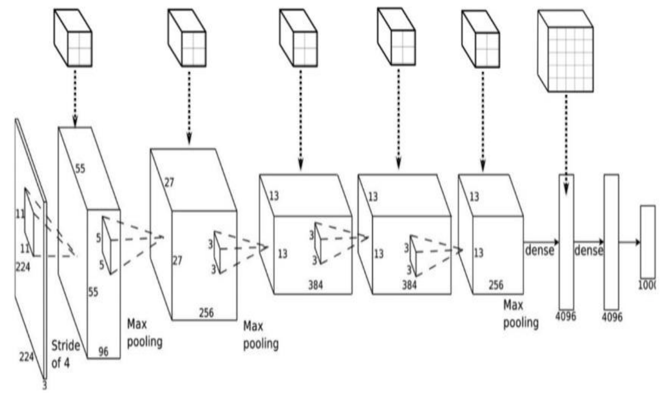


Start-Up 2012: Try Fashion online

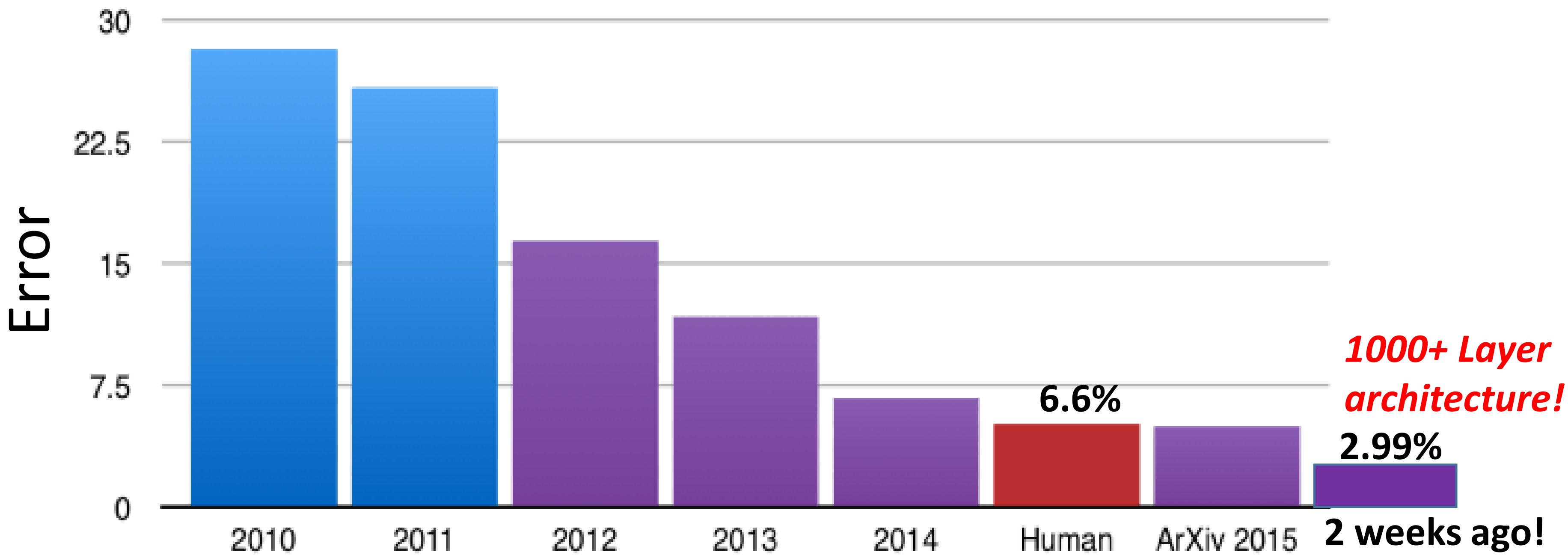
Very large impact in many field: Gaming, Robotics, HCI, Medicine, …

Object recognition – ImageNet

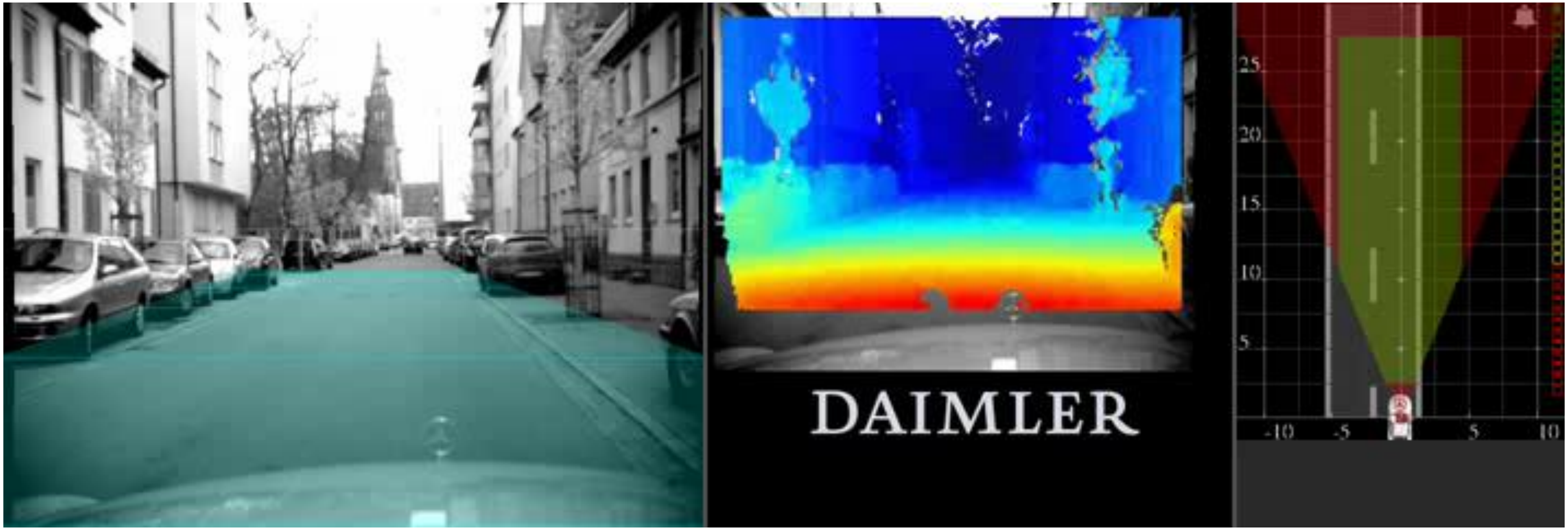




What's in the image (1000 classes)

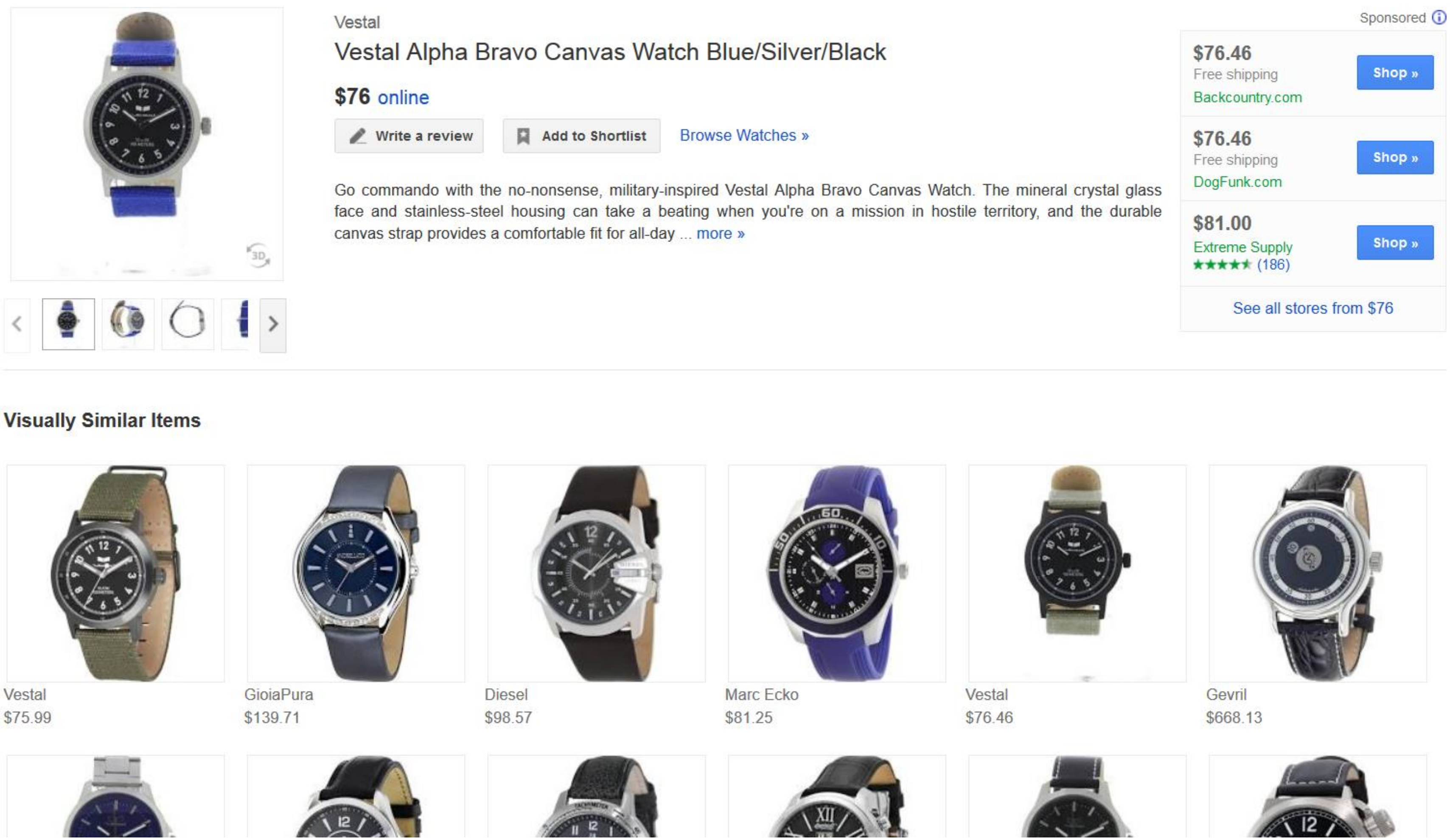


Real-time pedestrian detection



Daimler Research Lab

Start-up Company: Like.com



Interactive Image manipulation

[Agrawal et al '04]

AI for Art

The Shipwreck of the Minotaur
by J.M.W. Turner, 1805.

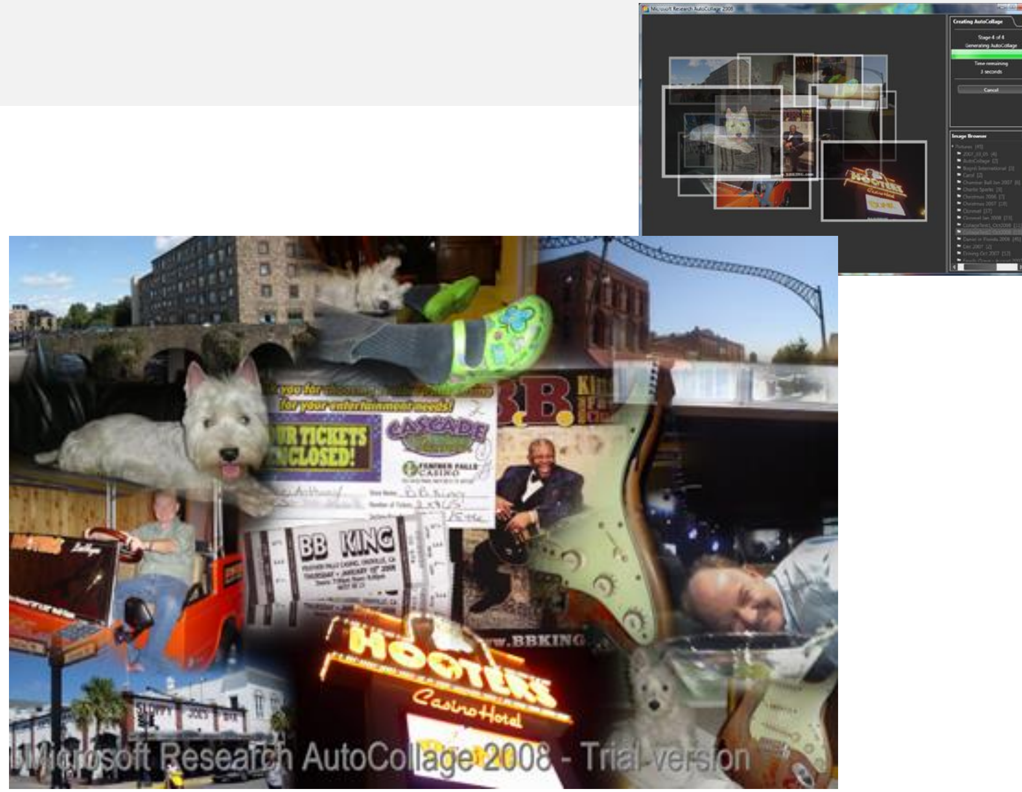
[A Neural Algorithm of Artistic Style, Gatys, Ecker, Bethge. ArXiv 2015]

Video Editing

[Rav-Acha et al. ‘08]

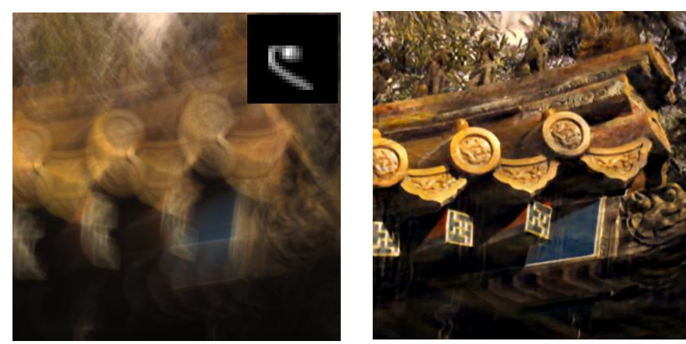
Other Industry

Pirates of the Caribbean,
Industrial Light and Magic



AutoCollage 2008 - Microsoft Research
[Rother et al. Siggraph 2006]

Robotics



Deconvolution (medical, astronomy)

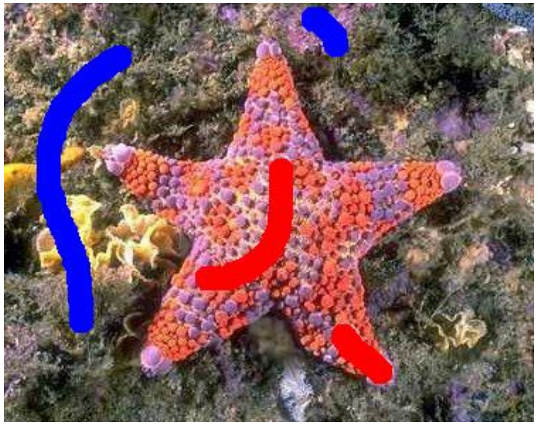
Introduction to Computer Vision

What is computer Vision?

(Potential) Definition:
Developing computational models and algorithms to interpret digital images and visual data in order to understand the visual world we live in.

Model versus Algorithm

Example: Interactive Segmentation



$z = (R, G, B)^n$

Goal

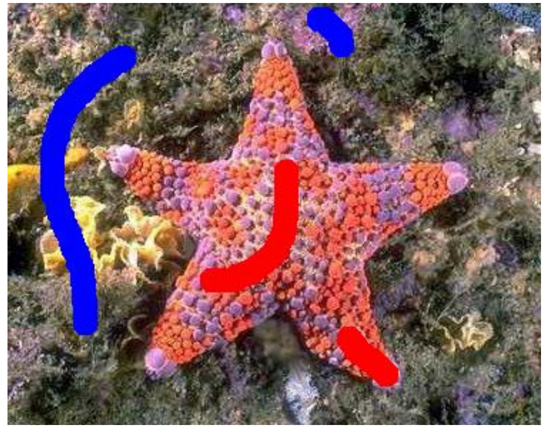
$\mathbf{x} = \{0,1\}^n$

Given **z**; derive binary **x**:

Model: Energy function $\boldsymbol{E}(\boldsymbol{x})$

Algorithm to minimization: $\boldsymbol{x}^* = argmin_x\, E(\boldsymbol{x})$

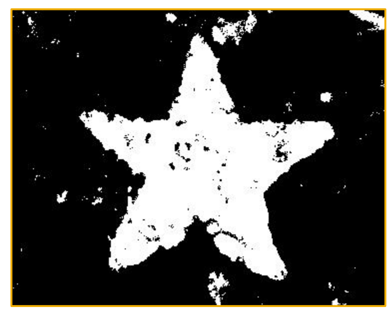
Model for a starfish



Goal: formulate $\boldsymbol{E}(\boldsymbol{x})$ such that

$\boldsymbol{E}(\boldsymbol{x}) = 0.01$

$\boldsymbol{E}(\boldsymbol{x}) = 0.05$



$\boldsymbol{E}(\boldsymbol{x}) = 0.05$

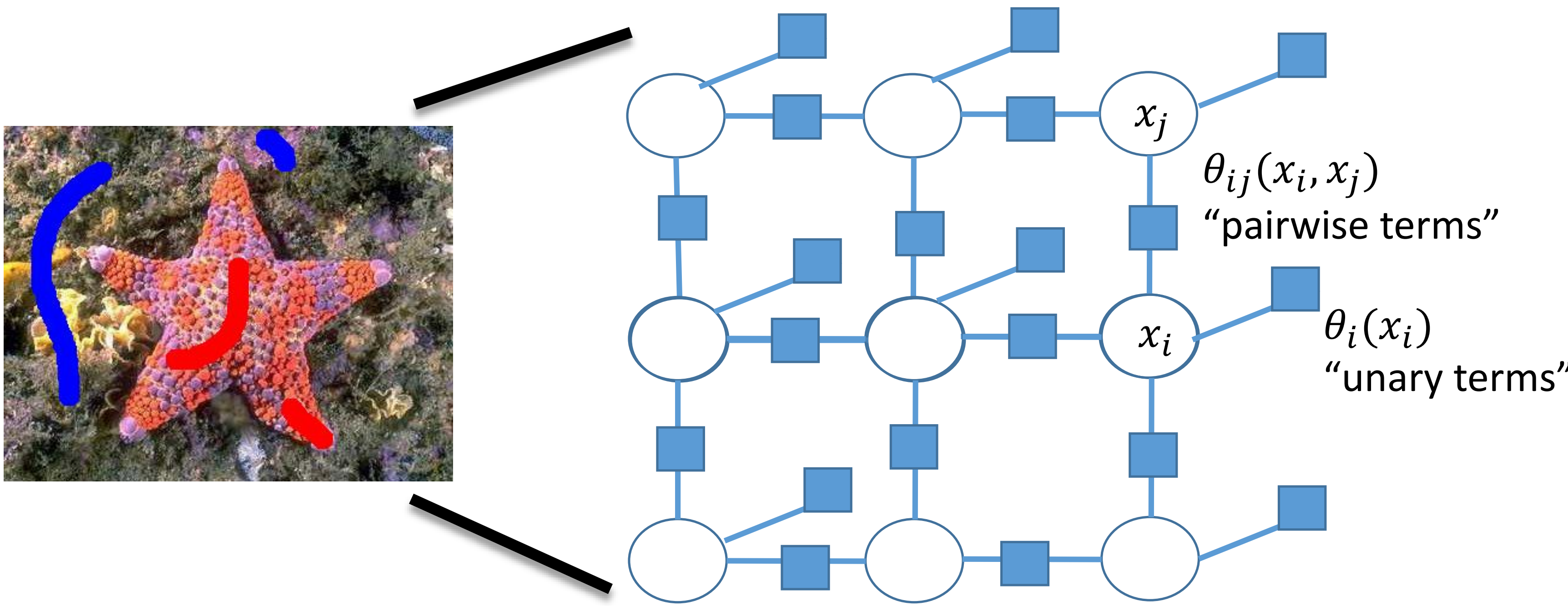
$\boldsymbol{E}(\boldsymbol{x}) = 0.1$

Optimal solution $\boldsymbol{x}^* = argmin_x \, E(\boldsymbol{x})$

How does the energy looks like?

Energy function (sum of local terms):

$$\boldsymbol{E}(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{i,j} \theta_{ij}(x_i, x_j)$$



Undirected graphical model

This is the focus of Machine Learning 2

Why is computer vision interesting (to you)?

- It is a challenging problem that is not yet solved

- It combines insights and tools from many fields and disciplines:
  - Mathematics and statistics
  - Cognition and perception
  - Engineering (signal processing)
  - Computer science

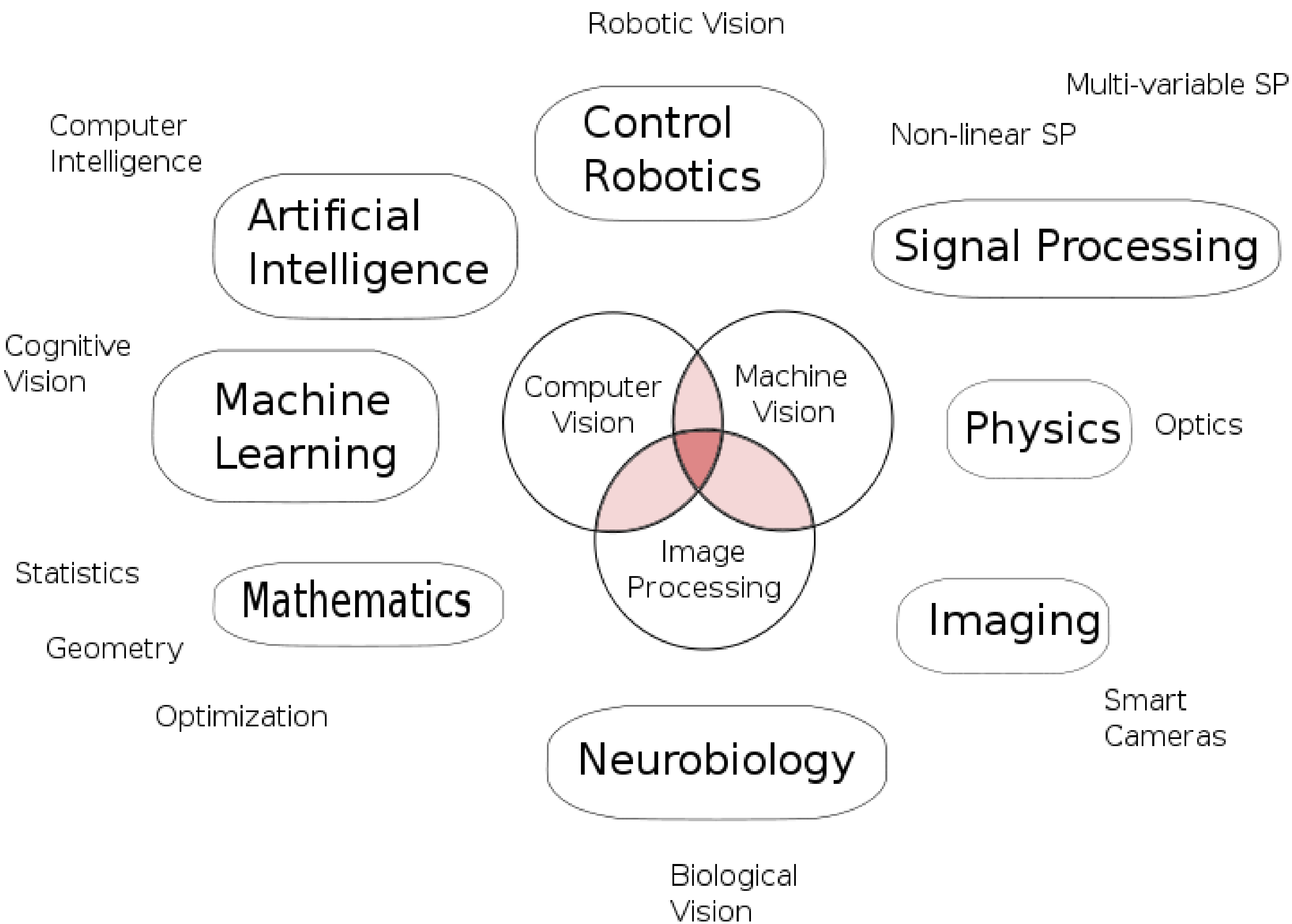
- Allows you to apply theoretical skills

  ... that you may otherwise only use rarely

- Quite rewarding:
  - Often visually intuitive and encouraging results

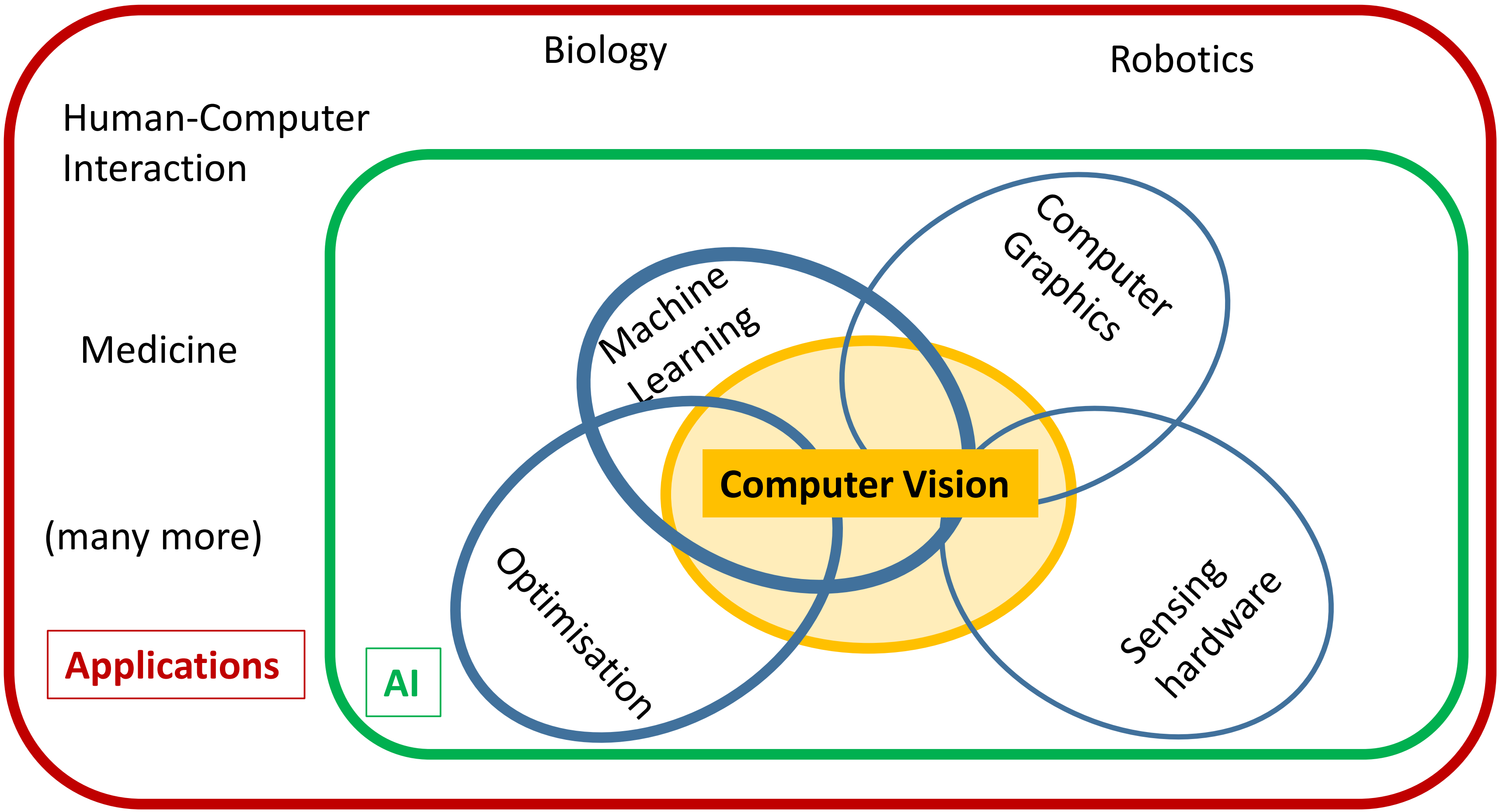
- It is a super-hot topic at the moment:
  - There are a lot of companies (big, small, start-up) working in computer vision and general AI
  - Size of Conferences are growing rapidly
  - Deep Learning has revolutionised the field

Relationship to other fields



[Wikipedia]

Relationship to other fields – my personal view



Roadmap: Basics of Digital Image Processing

- What is an Image?
- Point operators (ch. 3.1)
- Filtering: (ch. 3.2, ch 3.3, ch. 3.4)
  - Linear filtering
  - Non-linear filtering
- Multi-scale image representation (ch. 3.5)
- Edges detection and linking (ch. 4.2)
- Interest Point detection (ch. 4.1.1)

Computer Vision: Algorithms and Applications by Rick Szeliski; Springer 2011. An earlier version of the book is online: http://szeliski.org/Book/

What is an Image

- We can think of the image as a function:

$$I(x,y), \qquad I: \ \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

  - For every 2D point (pixel) it tells us the amount of light it receives
  - The size and range of the sensor is limited:

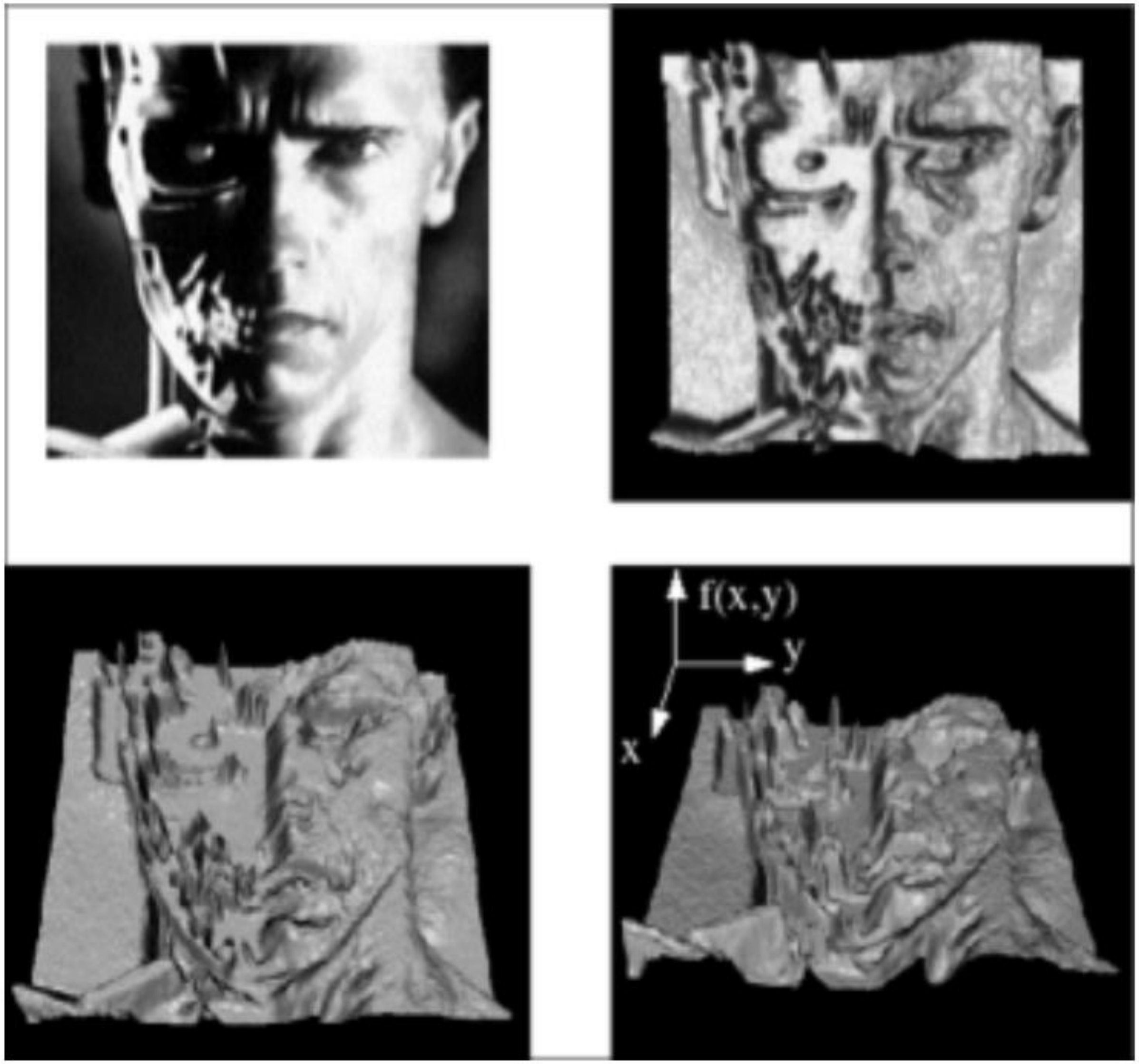
$$I(x,y), \qquad I: [a,b] \times [c,d] \to [0,m]$$

  - Colour image is then a vector-valued function:

$$I(x,y) = \begin{pmatrix} I_R(x,y) \\ I_G(x,y) \\ I_B(x,y) \end{pmatrix}, \qquad I: [a,b] \times [c,d] \to [0,m]^3$$

  - Comment, in most lectures we deal with grey-valued images and extension to colour is often "obvious"

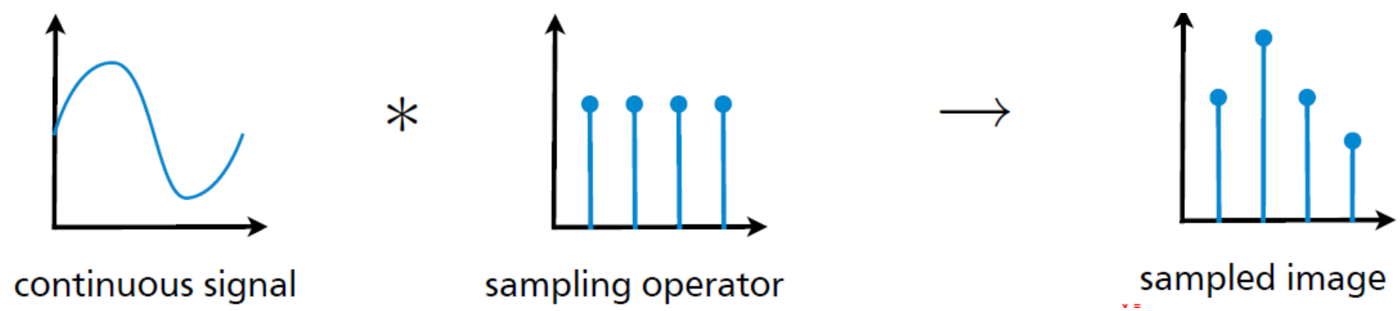
Images as functions



[from Steve Seitz]

Digital Images

- We usually do not work with spatially continuous functions, since our cameras do not sense in this way.
- Instead we use (spatially) discrete images
- Sample the 2D domain on a regular grid (1D version)



- Intensity/color values are usually also discrete
  Quantize the values per channel
  (e.g. 8 bit per channel)

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

Roadmap: Basics of Digital Image Processing

- What is an Image?
- Point operators (ch. 3.1)
- Filtering: (ch. 3.2, ch 3.3, ch. 3.4)
  - Linear filtering
  - Non-linear filtering
- Multi-scale image representation (ch. 3.5)
- Edges detection and linking (ch. 4.2)
- Interest Point detection (ch. 4.1.1)

Computer Vision: Algorithms and Applications by Rick Szeliski; Springer 2011. An earlier version of the book is online: http://szeliski.org/Book/

Point operators

- Point operators work on every pixel independently:

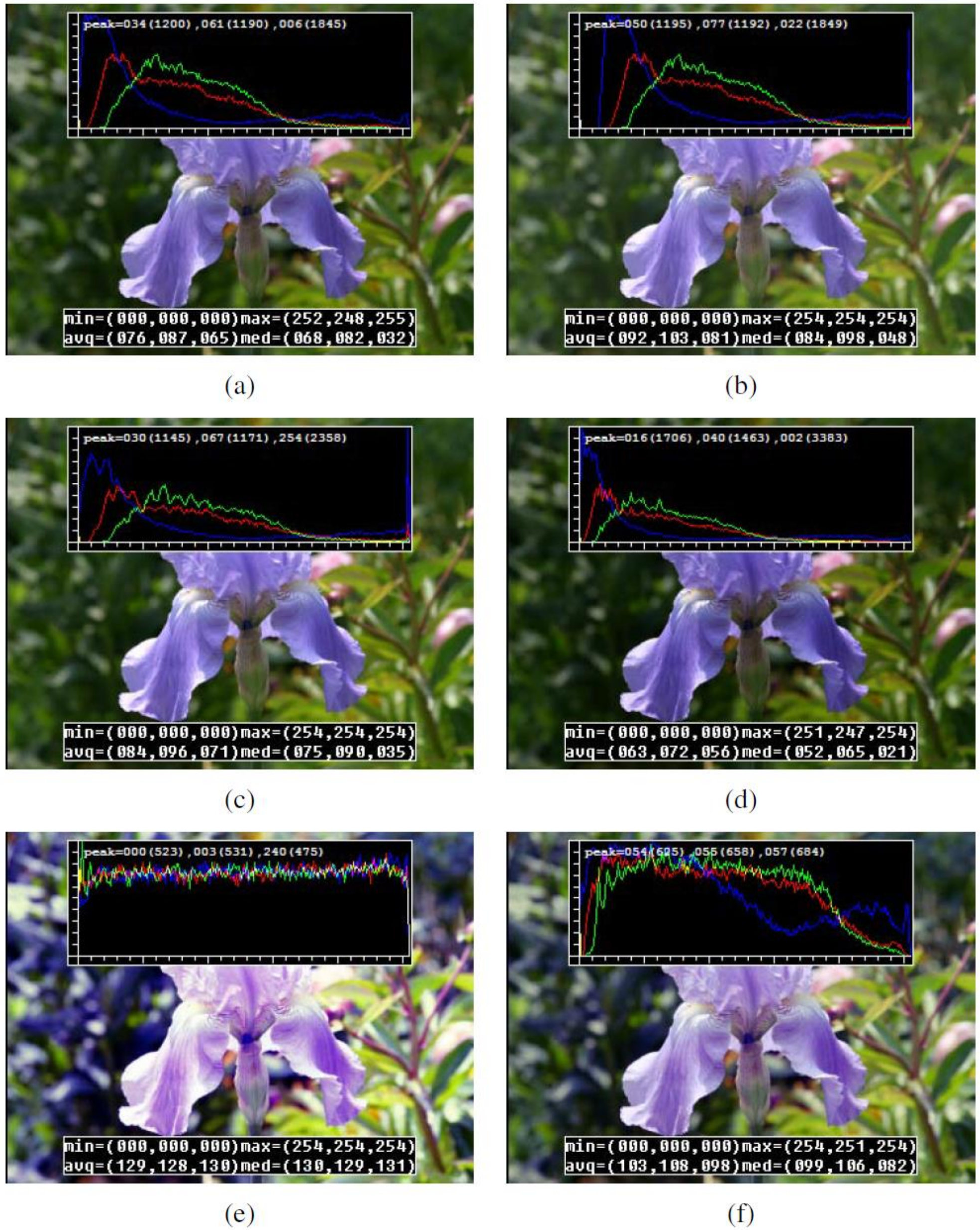
$$J(x, y) = h\big(I(x, y)\big)$$

- Examples for h:
  - Control contrast and brightness; $h(z) = az^b + c$

original

Contrast enhanced

Example



peak=034(1200),061(1190),006(1845)
min=(000,000,000)max=(252,248,255)
avg=(076,087,065)med=(068,082,032)
(a)
peak=050(1195),077(1192),022(1849)
min=(000,000,000)max=(254,254,254)
avg=(092,103,081)med=(084,098,048)
(b)
peak=030(1145),067(1171),254(2358)
min=(000,000,000)max=(254,254,254)
avg=(084,096,071)med=(075,090,035)
(c)
peak=016(1706),040(1463),002(3383)
min=(000,000,000)max=(251,247,254)
avg=(063,072,056)med=(052,065,021)
(d)
peak=000(523),003(531),240(475)
min=(000,000,000)max=(254,254,254)
avg=(129,128,130)med=(130,129,131)
(e)
min=(000,000,000)max=(254,251,254)
avg=(103,108,098)med=(099,106,082)
(f)