Computer Vision 1

THEO GEVERS

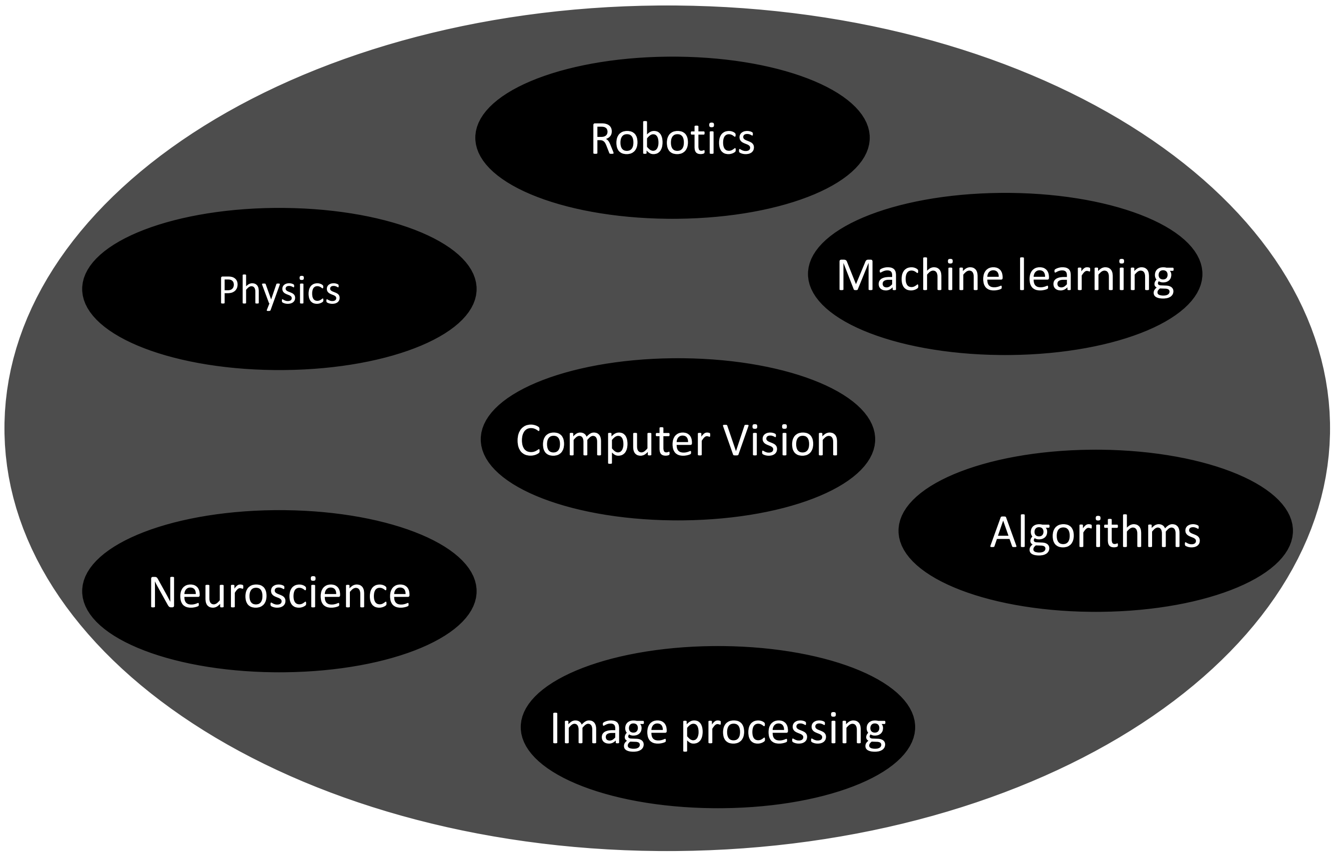
MASTER AI

UNIVERSITY OF AMSTERDAM

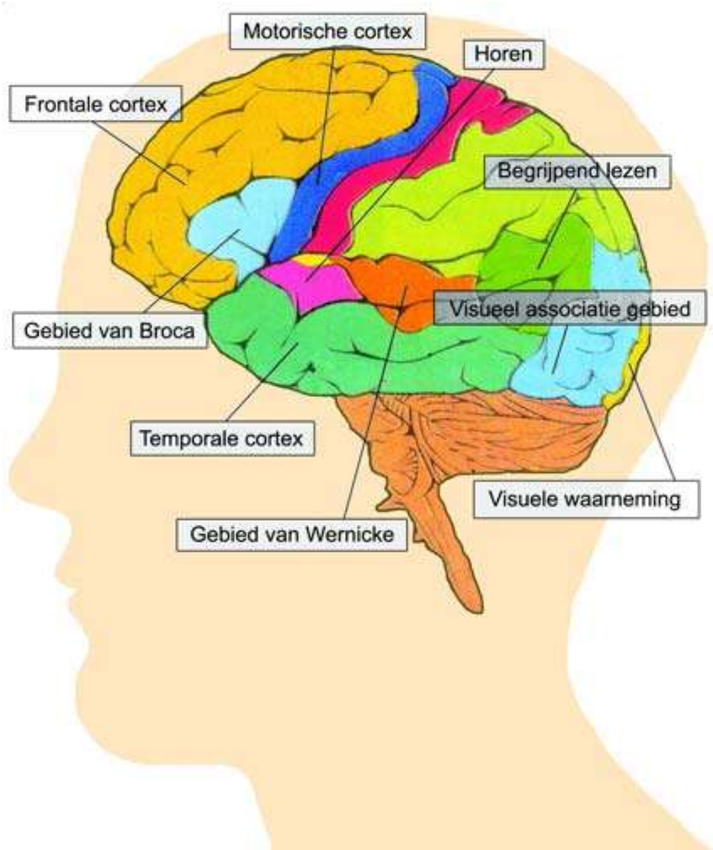
What is Computer Vision?

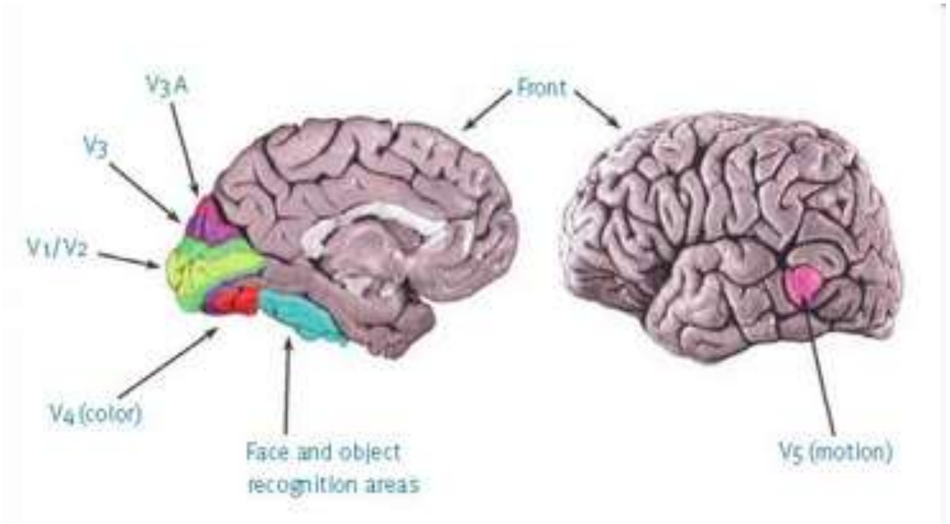
- Automatic understanding of images and video
  - Computing properties of the 3D world from visual data (**measurement**)
    E.g., Image processing, Physical modeling, Stereo, 3D reconstruction
  - Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. (**perception and interpretation**)
    E.g., Segmentation, Image Classification, Object Detection,
  - And of course many cross-overs

What is it related to?



Visual Cortex





Average number of images seen by a person:

3 images/second * 60 * 60 * 16 * 365 * 60

3.7 billion images

Images

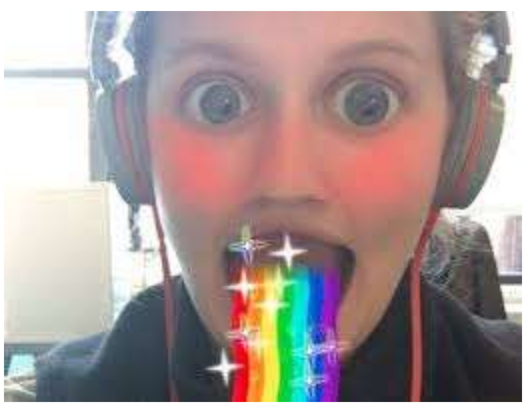
YouTube has more than 1 billion videos

Google has more than 1.5 billion images

Facebook: more than 400 million image uploads per day

Instagram: more than 150 milliion image uploads per day

CV: Applications



CV: Applications

Hololens Lenovo phab 2 pro

GO and Tesla

Course Organization

(more details on **blackboard**)

This is a 6 credit course in 7 weeks. The course has lectures of basic **theory** based on the freely downloadable books http://szeliski.org/Book/ and http://www.deeplearningbook.org/, and related papers (all readings provided at *Blackboard*). **Seminars** are given to practice the basic theory. Further, the course contains hands-on experience in the **practical** lab sessions.

Prerequisites

- Linear algebra, calculus, and probability
- Machine learning
- Matlab, data structures

Course Organization

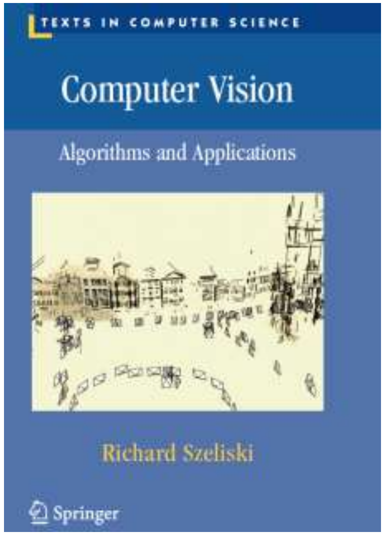
- Grades run from 1 to 10 (highest). Each part of the course should be scored **at least a 5.5**. The final grade is based on the following weighted parts:
- Exam: 50%
- Practical (in groups of 2): 50%

Course Organization

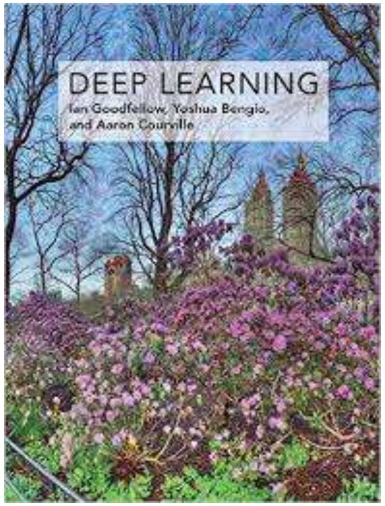
- **Exam, 50%**
- The questions will cover the basic theory on computer vision. The slides and exercises (seminars) will form the basis for the (closed book) exam. Read the book and papers for background knowledge but the slides and exercises are most important.
- **Focus on chapters**:
  - **Szeliski:** 1, 2.1.1 + 2.1.2 + 2.2 + 2.3.2 + 2.3.3, 3.1 + 3.2 + 3.3, 4 (all), 5.1.1 + 5.1.4 + 5.2 + 5.3 + 5.4 + 5.5, 6.1 + 6.3, 7.1, 8.1 + 8.2 + 8.4, 10.3, 11.1 + 11.2 + 11.3 + 11.4, 12.1 + 12.2 + 12.3, 14.1 + 14.3 + 14.4 + 14.5 (http://szeliski.org/Book/)
  - **Bengio:** 4 + 5.1 + 5.2 + 5.3 + 5.7 + 5.8 + 5.9, 7.2 + 7.4 + 9.1 + 9.2 + 9.3, 12.1 + 12.2 (http://www.deeplearningbook.org/)

Textbooks

- CV1 is based on "*Computer Vision: Algorithms and Applications*" by Richard Szeliski
  - Freely available for download from http://szeliski.org/Book/



- and "*Deep Learning*" by Ian Goodfellow, Yoshua Bengio and Aaron Courville
  - Freely available for download from http://www.deeplearningbook.org/

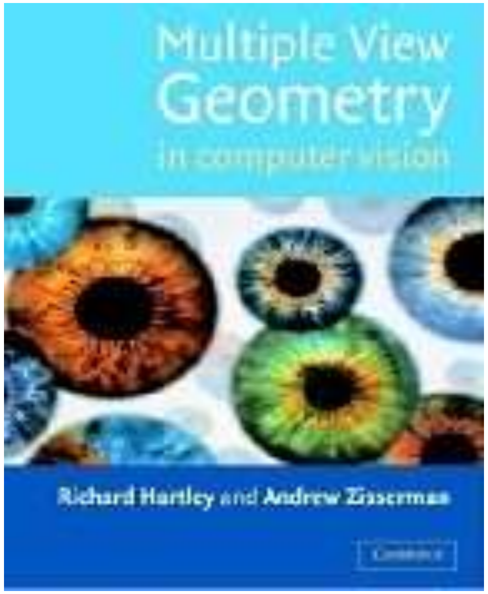


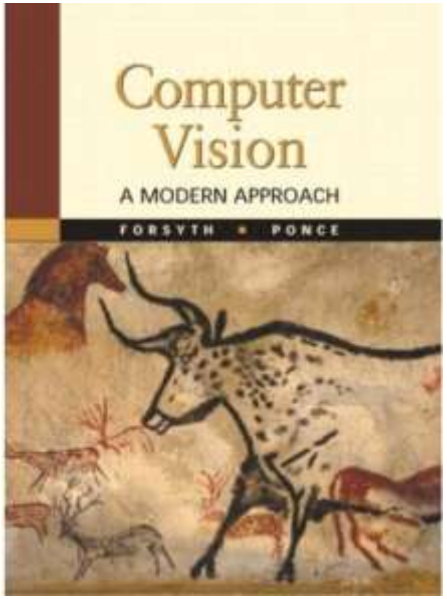
Background

- Two other useful books

  Forsyth, David A., and Ponce, J. Computer Vision: A Modern Approach, Prentice Hall, 2003.

  Hartley, R. and Zisserman, A. Multiple View Geometry in Computer Vision, Academic Press, 2002.



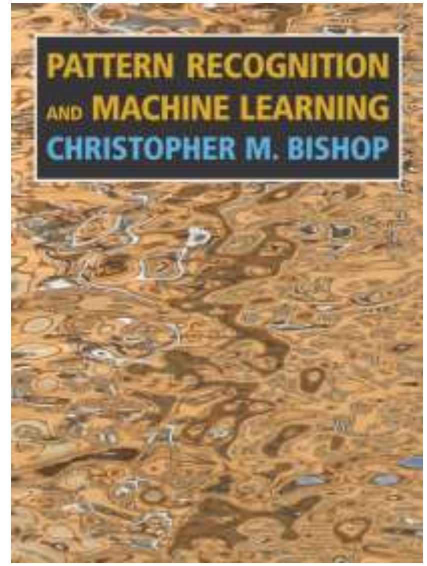


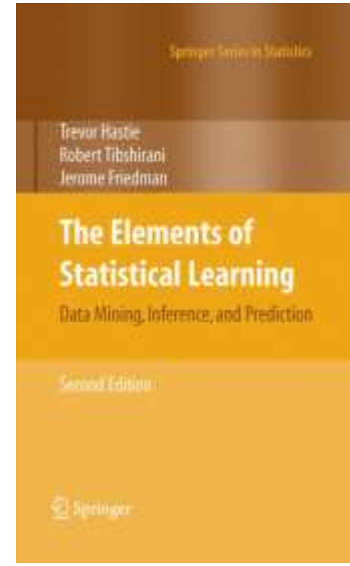
Even More Background and Tools

- Machine Learning books:

  Bishop, C. Pattern Recognition and Machine Learning, 2006.

  Hastie, T., Tribshirani R., and Friedman J, The Elements of Statistical Learning, Freely available for download: http://www-stat.stanford.edu/~tibs/ElemStatLearn/





- Software, programming tools

  Matlab, image processing toolbox,
  C++, openCV
  Python, PIL, numpy, scipy
  PRTools, Shogun, (machine learning toolboxes)

- Scientific papers:

  Top Journals: Transactions of Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, TIP, (Computer Vision and Image Understanding)
  Top Conferences: ICCV, ECCV, CVPR, NIPS, SIGGRAPH, BMVC, (ACM-Multimedia)
  http://www.cvpapers.com/

Tools and Tutorials

Tools

• VLFeat, open source implementations of Computer Vision algorithms Link
• OpenCV, open source Computer Vision framework Link
• Theano, Python Math Expression Library (Neural Network Optimization) Link
• Caffe, Neural Network Framework Link

Tutorials

• Neural Network Tutorial Link
• Matlab Tutorials
  - David Griffiths' Matlab notes Link
  - UCSD Computer Vision course Matlab introduction Link
• Camera Model Visualization Link

Lectures/Theory

- 06-02-2018, 17:00-19:00, C0.05, **Introduction** (*Szeliski 1*)
- 13-02-2018, 17:00-19:00, C0.05, **Image Formation** (*Szeliski: 2.1.1 + 2.1.2 + 2.2 + 2.3.2 + 2.3.3*)
- 20-02-2018, 17:00-19:00, C0.05, **Color and Image Processing** (*Szeliski: 3.1 + 3.2 + 3.3*)
- 27-02-2018, 17:00-19:00, C0.05, **Feature Detection, Motion and Classification** (*Szeliski: 4, 8.1.1 + 8.1.3 + 8.2.1 + 8.4; Bengio: 4 + 5.1 + 5.2 + 5.3 + 5.7 + 5.8 + 5.9*)
- 06-03-2018, 17:00-19:00, C0.05, **Object Recognition: BoW and ConvNets** (*Szeliski: 5.1.1 + 5.1.4 + 5.1.5 + 5.2 + 5.3 + 5.4, 6.1 + 6.3, 14.1 + 14.2.1 + 14.3 + 14.4.1; Bengio: 7.2 + 7.4 + 9.1 + 9.2 + 9.3*)
- 13-03-2018, 17:00-19:00, C0.05, **Deep Learning, Stereo and 3D Reconstruction** (*Szeliski: 11.1 + 11.2 + 11.3 + 11.4, 12.1 + 12.2: Bengio: 12.1 + 12.2*)
- 20-03-2018, 17:00-19:00, C0.05, **Applications** (*Szeliski: 12.6.2 + 12.6.3 + 12.2.4)*
- 26-03-2018, Monday, 9:00-12:00, **Written Exam**

Seminars/Exercises

- 14-02-2017, 15:00-17:00, CWI-Turingzaal
- 21-02-2017, 15:00-17:00, CWI-Turingzaal
- 07-03-2017, 15:00-17:00, CWI-Turingzaal
- 14-03-2017, 15:00-17:00, CWI-Turingzaal

Practical Assignments

- **Practical, 50%**
- You will implement 2 methods for object recognition. Each week you do a part. The idea is to combine computer vision modules to have a real working system in the end. In the first weeks you will do separate modules. Then, you have to combine the modules to get the final system and write a small report about.

Lab Session

7 TA's: Berkay Kicanaoglu, Mert Kilickaya, Anil Baslamisli, Hanan ElNaghy, Hoang An Le, William Thong and Minh Ngo

- 08-02-2017, 17:00-19:00 **Introduction to MatLab (optional)**
- 15-02-2017, 17:00-19:00 **Photometric Stereo & Color**
- 22-02-2017, 17:00-19:00 **Neighborhood Processing: Gabor & Gaussian Filters**
- 01-03-2017, 17:00-19:00 **Harris Corner Detector, Optical Flow and Feature Tracking**
- 08-03-2017, 17:00-19:00 **Image Alignment and Stitching / Final Project**
- 15-03-2017, 17:00-19:00 **Final Project**
- 22-03-2017, 17:00-19:00 **Final Project**

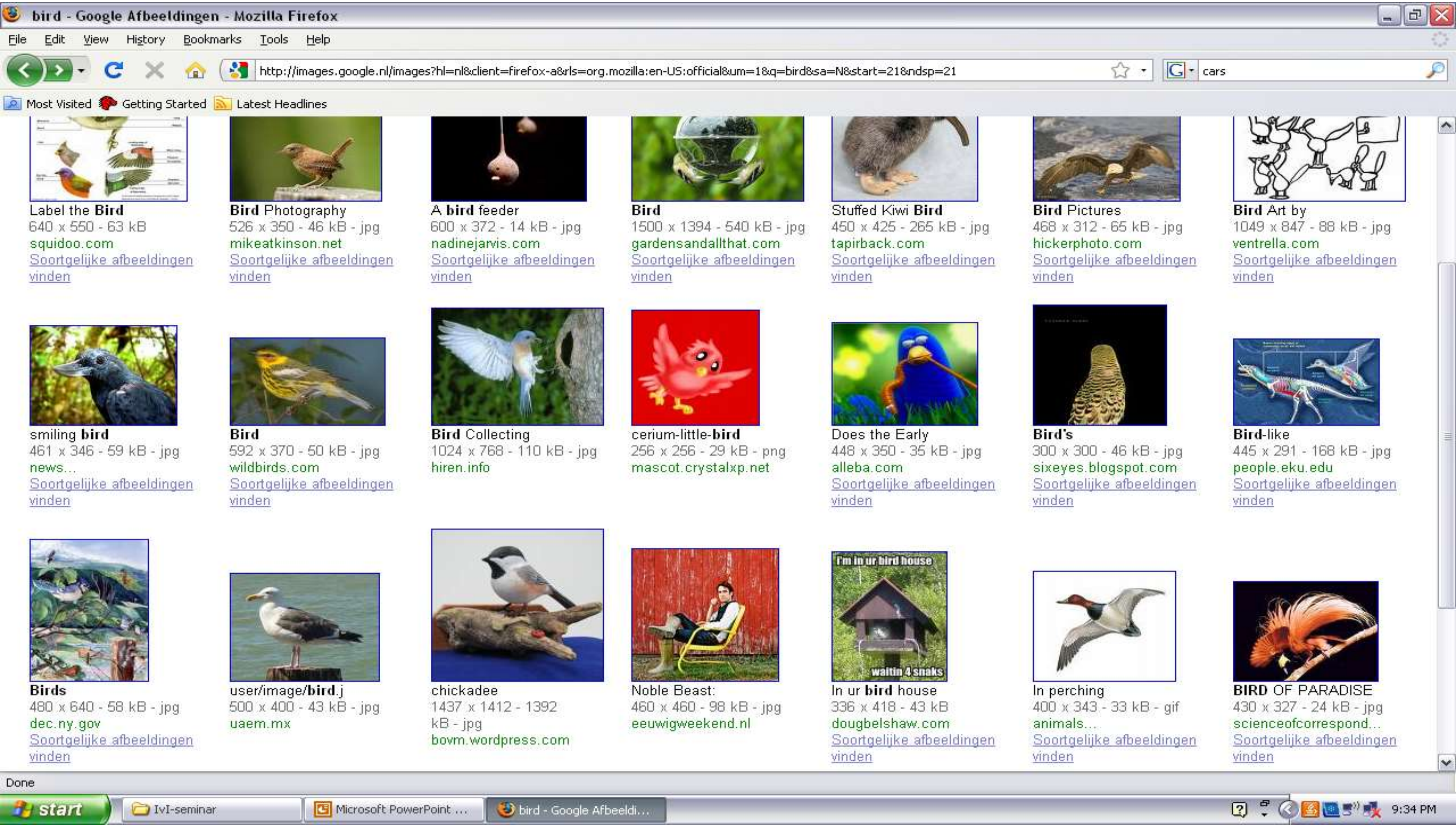
Today's Class (Szeliski Chapter 1)

- Specifics of this course
- Introduction to Computer Vision
  - What are Computer Vision Systems
  - Computer vision and machine learning
- Image search

Image and Video Access

- Today, there are billions of images on the Internet and in collections such as FaceBook and Flickr.
- Suppose I want to find pictures of birds, humans, cars, boats or videos of explosion, violence etc

Google Image Search – Bird(2)



Flickr Search for tag *monkey*








- Even with humans doing the labeling, the data is extremely noisy -- context, polysemy, photo sets
- Tags are not enough either!

Video Retrieval

Given a shot from a video…
… is some semantic *concept* present in that shot?

Example concepts:

- Airplane
- Building
- **Car**
- Crowd
- **Desert**
- **Explosion**
- **Outdoor**
- People
- **Vehicle**
- **Violence**

Object/Scene Categories

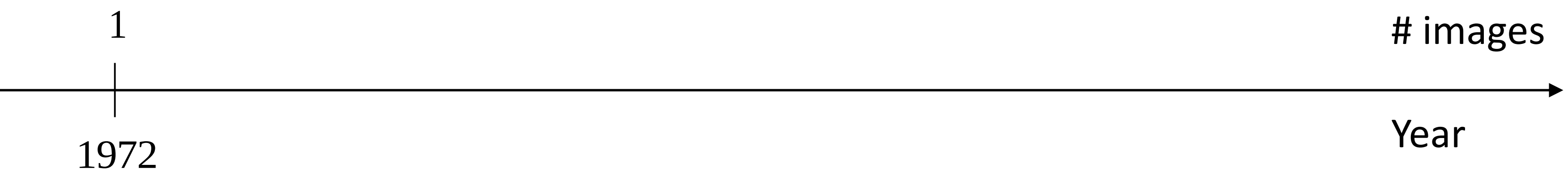
Aircraft | Animal | Boat | Building | Bus | Car | Chart | Corp. leader | Court

Crowd | Desert | Entertainment | Explosion | Face | Flag USA | Gov. leader | Map | Meeting

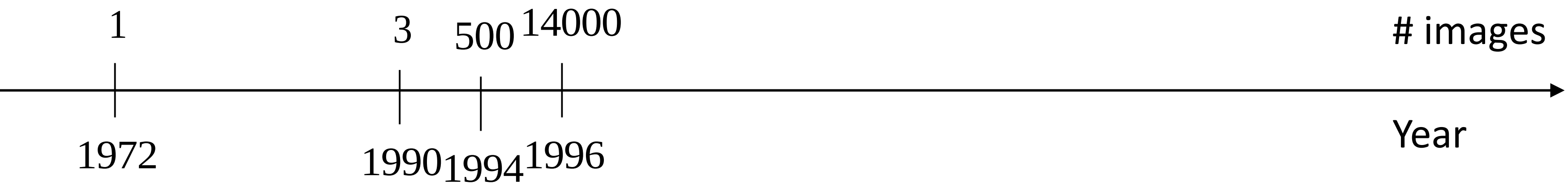
Military | Mountain | Natural disaster | Office | Outdoor | People | People marching | Police / security | Prisoner

Screen | Sky | Sports | Studio | Truck | Urban | Vegetation | Vehicle | Violence

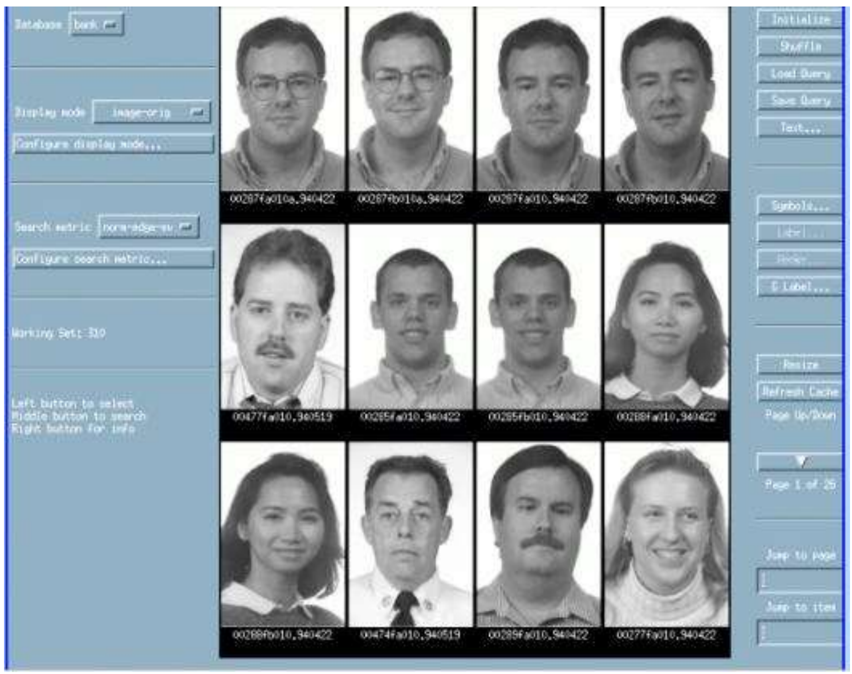
Datasets in perspective



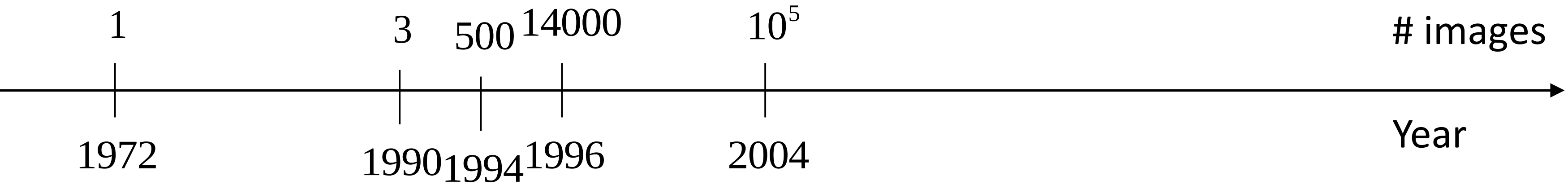
Lena (1972)



DARPA Faces (1996)



In 1996 DARPA released
14000 images,
from over 1000 individuals.

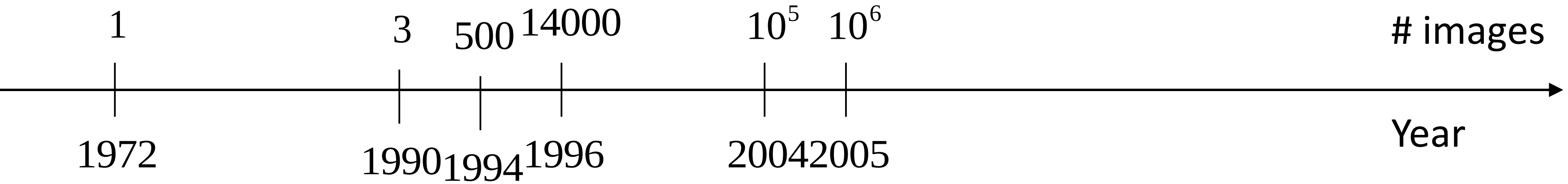


Caltech 101 and 256 (100,000)

Griffin, Holub, Perona, 2007

Fei-Fei, Fergus, Perona, 2004

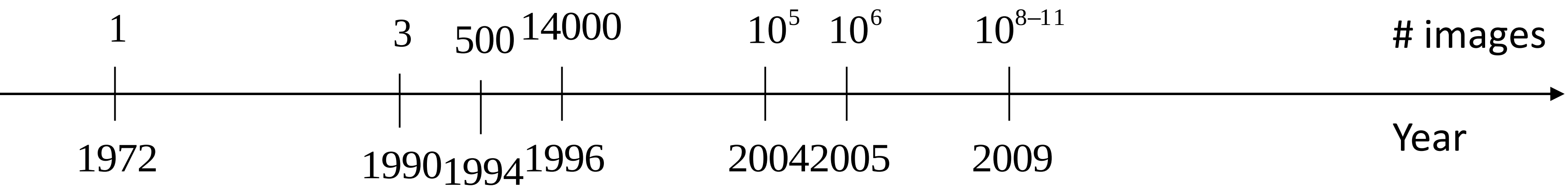
101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003



TRECVID and PASCAL VOC competition (2005-2009)

- 86 hours of video from TRECVID 2005
- Shot segmentation available: 43.907 shots
- Ground truth available from Mediamill Challenge

- The goal of VOC challenge is to recognize objects from a number of visual object classes in realistic scenes
- The twenty object classes are:
  - *Person:* person
  - *Animal:* bird, cat, cow, dog, horse, sheep
  - *Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train
  - *Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor.



> 1 billion images!

Number of images seen by all humanity:

1020

106,456,367,669 humans1 * 60 years * 3 images/second * 60 * 60 * 16 * 365 =

1 from http://www.prb.org/Articles/2002/HowManyPeopleHaveEverLivedonEarth.aspx

Number of images seen during the first 10 years:

(3 images/second * 60 * 60 * 16 * 365 * 10 = 630,720,000)

Labeling

Labeling to get a Ph.D.

Just labeling

LabelMe

Labeling for fun

Labeling for money

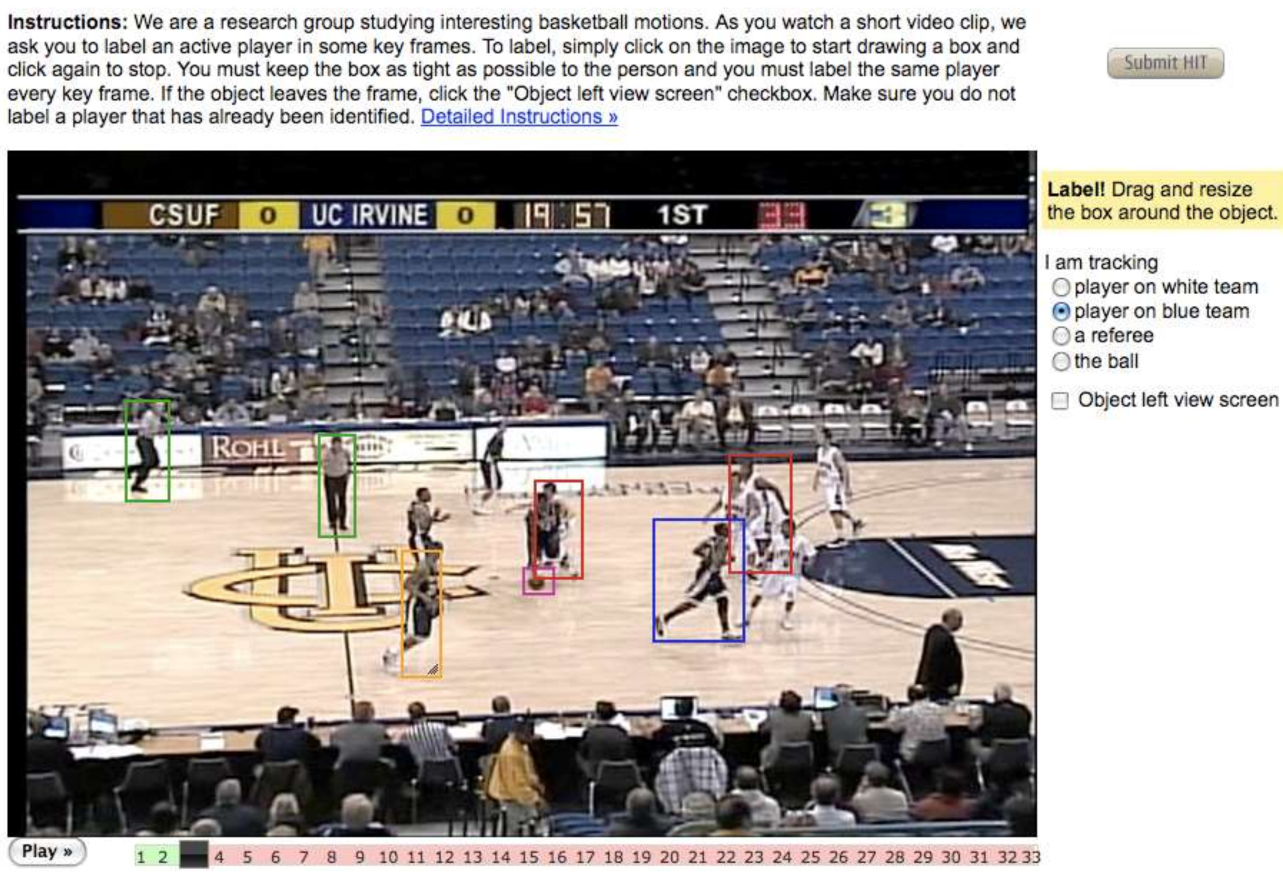
amazon mechanical turk
Artificial Artificial Intelligence

1 cent per image

Task: Label one object in this image

Labeling human activities

Carl Vondrick, Deva Ramanan, Don Patterson



https://workersandbox.mturk.com/mturk/preview?groupId=0YNZVTYH13MZP2ZVKS30

Image and Video Formation

Challenges

- Viewpoint variation

- Illumination change

- Orientation and scale

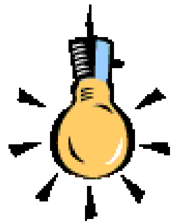
- Occlusion

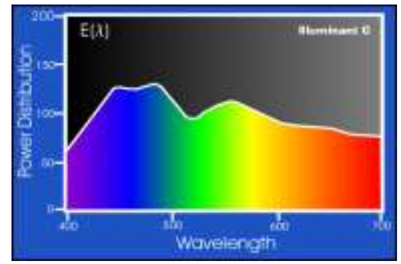
- Clutter

- Appearance change

What makes an image

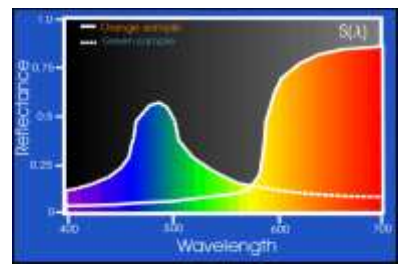
Light source





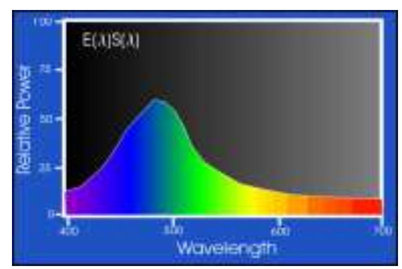
$$e(\lambda)$$

Object



$$\rho(\lambda)$$

Sensor



$$e(\lambda)\rho(\lambda)$$

$$R = \int_{\lambda} e(\lambda)\rho(\lambda) f_R(\lambda) d\lambda, \quad G = \int_{\lambda} e(\lambda)\rho(\lambda) f_G(\lambda) d\lambda, \quad B = \int_{\lambda} e(\lambda)\rho(\lambda) f_B(\lambda) d\lambda$$

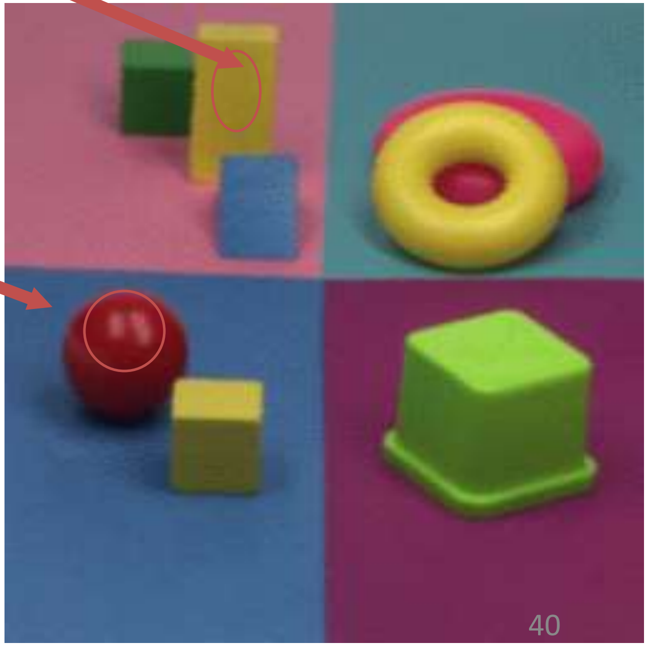
Image Formation

Reflection model:

body = $m_b(\vec{n},\vec{s}) \int_{\lambda} f_C(\lambda) e(\lambda) c_b(\lambda) d\lambda \quad +$

surface = $m_s(\vec{n},\vec{s},\vec{v}) \int_{\lambda} f_C(\lambda) e(\lambda) c_s(\lambda) d\lambda$

for {R,G,B} giving an R-, B-, G-sensor response

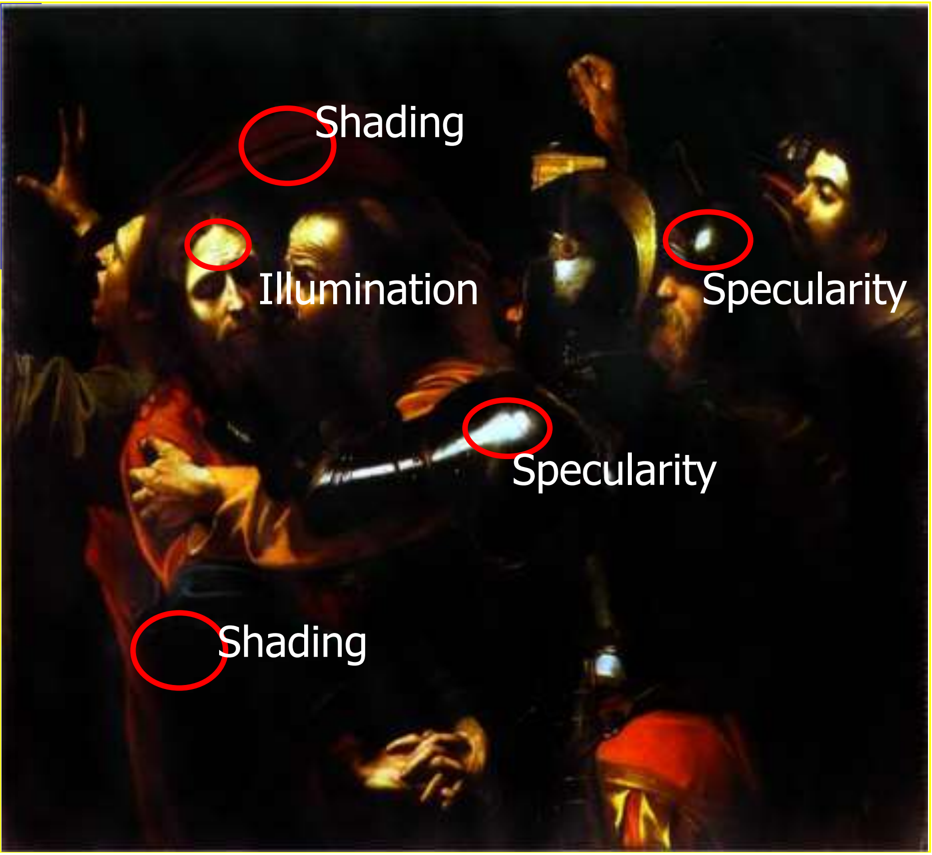


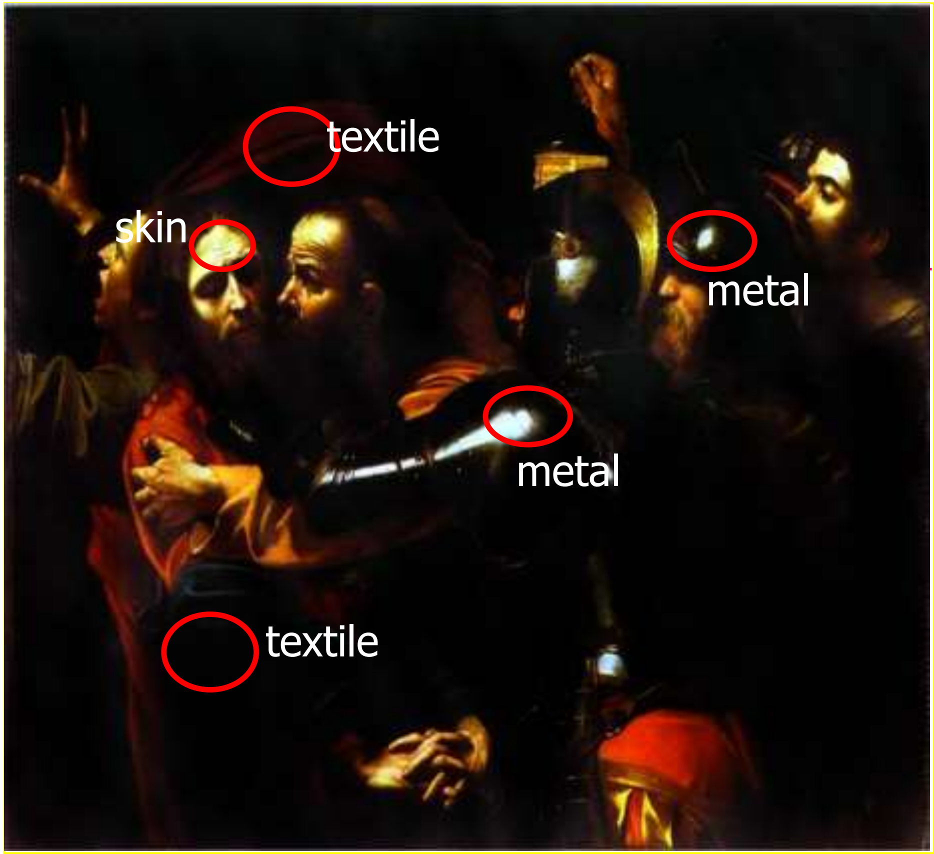
Reflection Model

C = $m_b(\vec{n},\vec{s}) \int_\lambda f_C(\lambda)e(\lambda)c_b(\lambda)d\lambda$ + $m_s(\vec{n},\vec{s},\vec{v}) \int_\lambda f_C(\lambda)e(\lambda)c_s(\lambda)d\lambda$

$c_b(\lambda)$	surface albedo	viewpoint invariant
$e(\lambda)$	illumination	scene dependent
$\vec{n}$	object surface normal	object shape variant
$\vec{s}$	illumination direction	scene dependent
$\vec{v}$	viewer's direction	viewpoint variant
$f_C(\lambda)$	sensor sensitivity	scene dependent

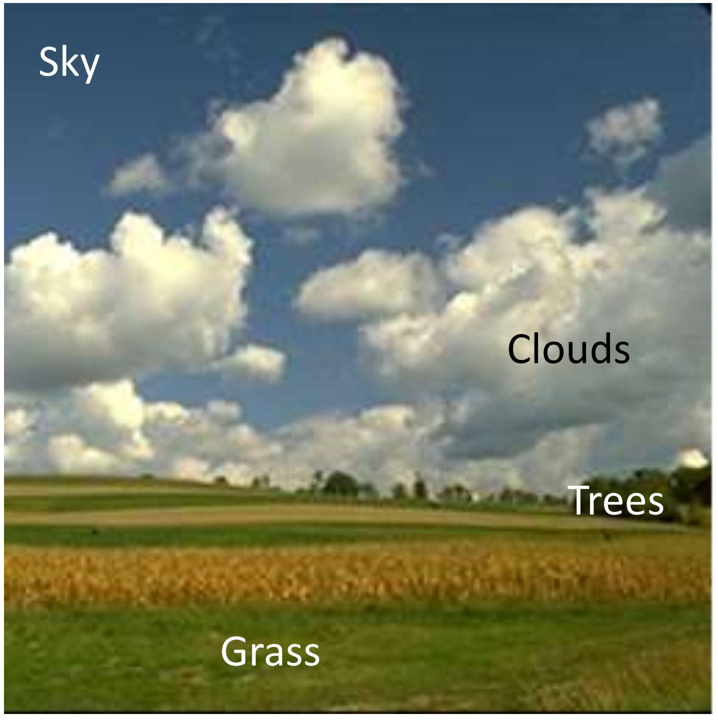
Image semantics (low-level)





Categorization (high-level)

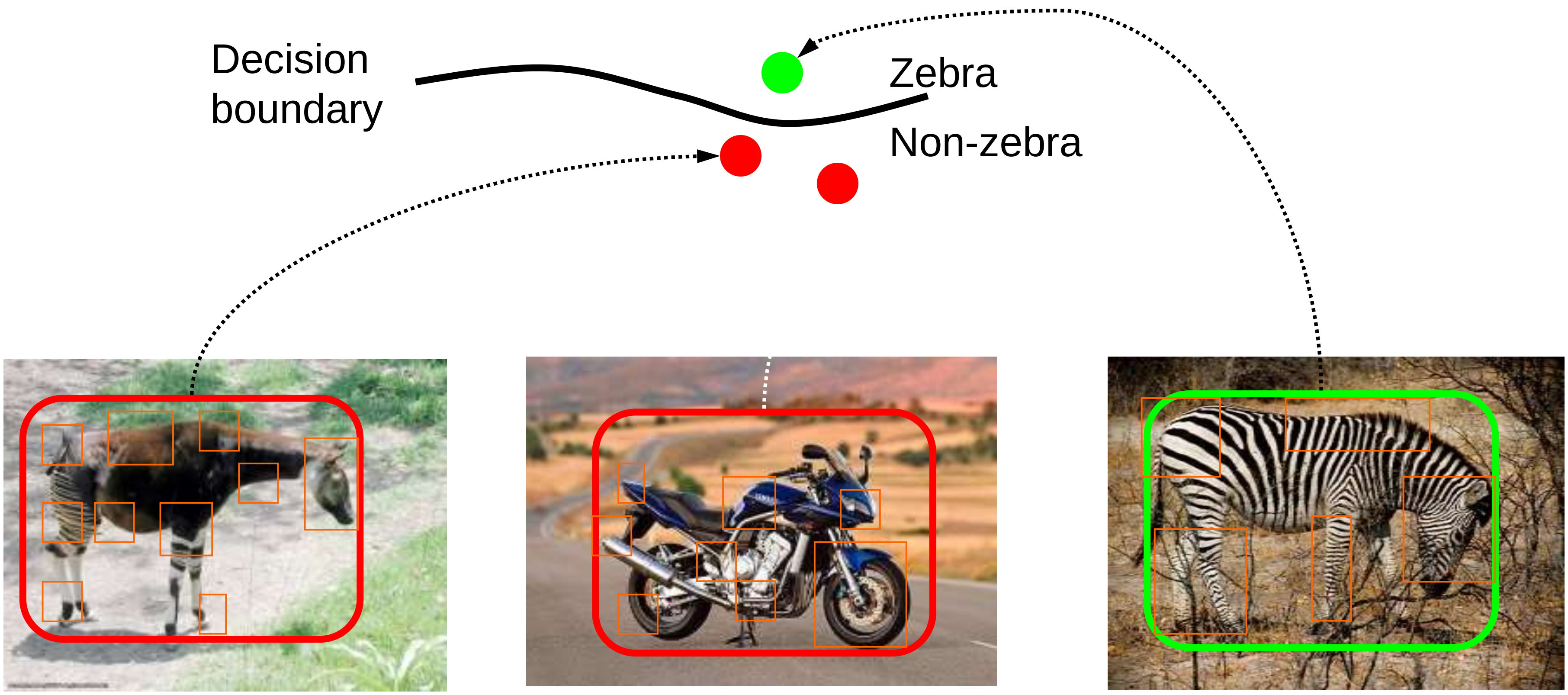
Outdoor/Landscape/vegatation…



Outdoor/city/street…

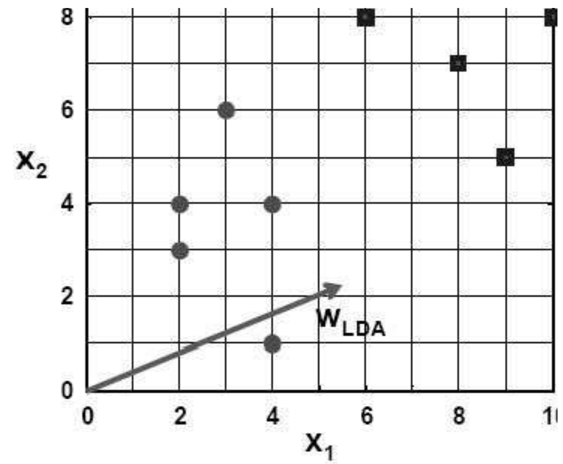
Machine learning

- Direct modeling of $\dfrac{p(zebra \mid image)}{p(no\ zebra \mid image)}$



LDA – 2 Classes: Example

- $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
- $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$



- **SOLUTION (by hand)**
  - The class statistics are:

$$S_1=\begin{bmatrix}0.80 & -0.40\\ -0.40 & 2.60\end{bmatrix};\ S_2=\begin{bmatrix}1.84 & -0.04\\ -0.04 & 2.64\end{bmatrix}$$

$$\mu_1=\begin{bmatrix}3.00 & 3.60\end{bmatrix};\qquad \mu_2=\begin{bmatrix}8.40 & 7.60\end{bmatrix}$$

  - The within- and between-class scatter are

$$S_B=\begin{bmatrix}29.16 & 21.60\\ 21.60 & 16.00\end{bmatrix};\ S_W=\begin{bmatrix}2.64 & -0.44\\ -0.44 & 5.28\end{bmatrix}$$

  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$S_W^{-1}S_Bv=\lambda v\Rightarrow\left|S_W^{-1}S_B-\lambda I\right|=0\Rightarrow\begin{vmatrix}11.89-\lambda & 8.81\\ 5.08 & 3.76-\lambda\end{vmatrix}=0\Rightarrow\lambda=15.65$$

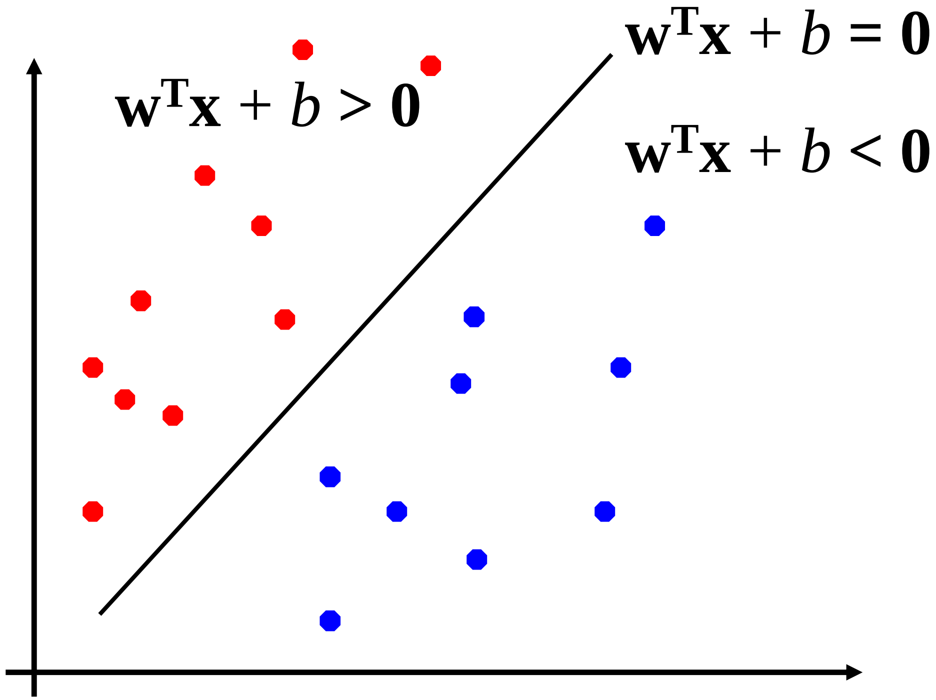
$$\begin{bmatrix}11.89 & 8.81\\ 5.08 & 3.76\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix}=15.65\begin{bmatrix}v_1\\ v_2\end{bmatrix}\Rightarrow\begin{bmatrix}v_1\\ v_2\end{bmatrix}=\begin{bmatrix}0.91\\ 0.39\end{bmatrix}$$

  - Or directly by

$$w^*=S_W^{-1}(\mu_1-\mu_2)=\begin{bmatrix}-0.91 & -0.39\end{bmatrix}^T$$

Linear Separators

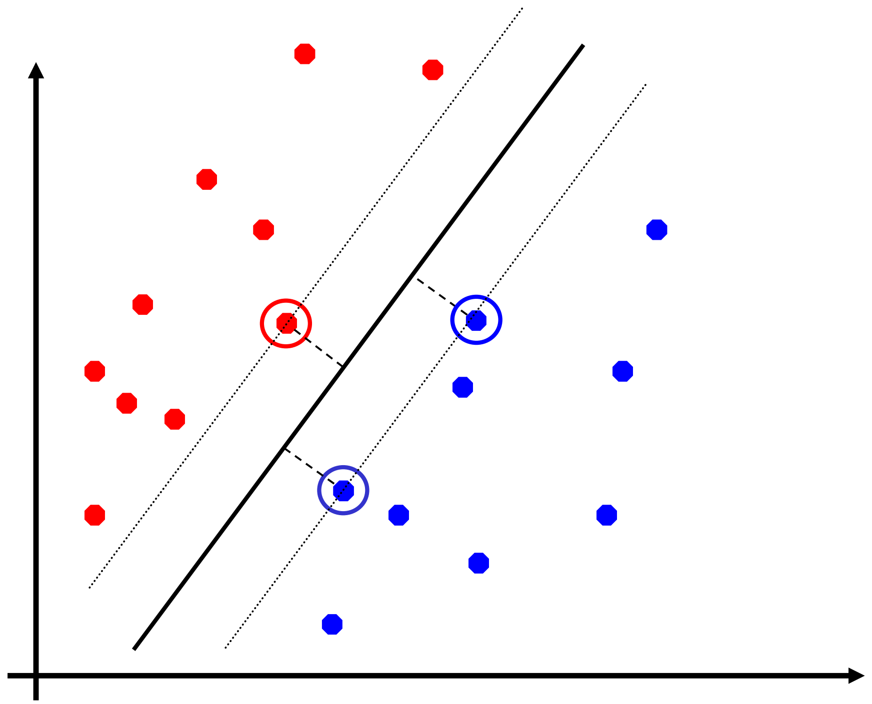
Binary classification can be viewed as the task of separating classes in feature space:



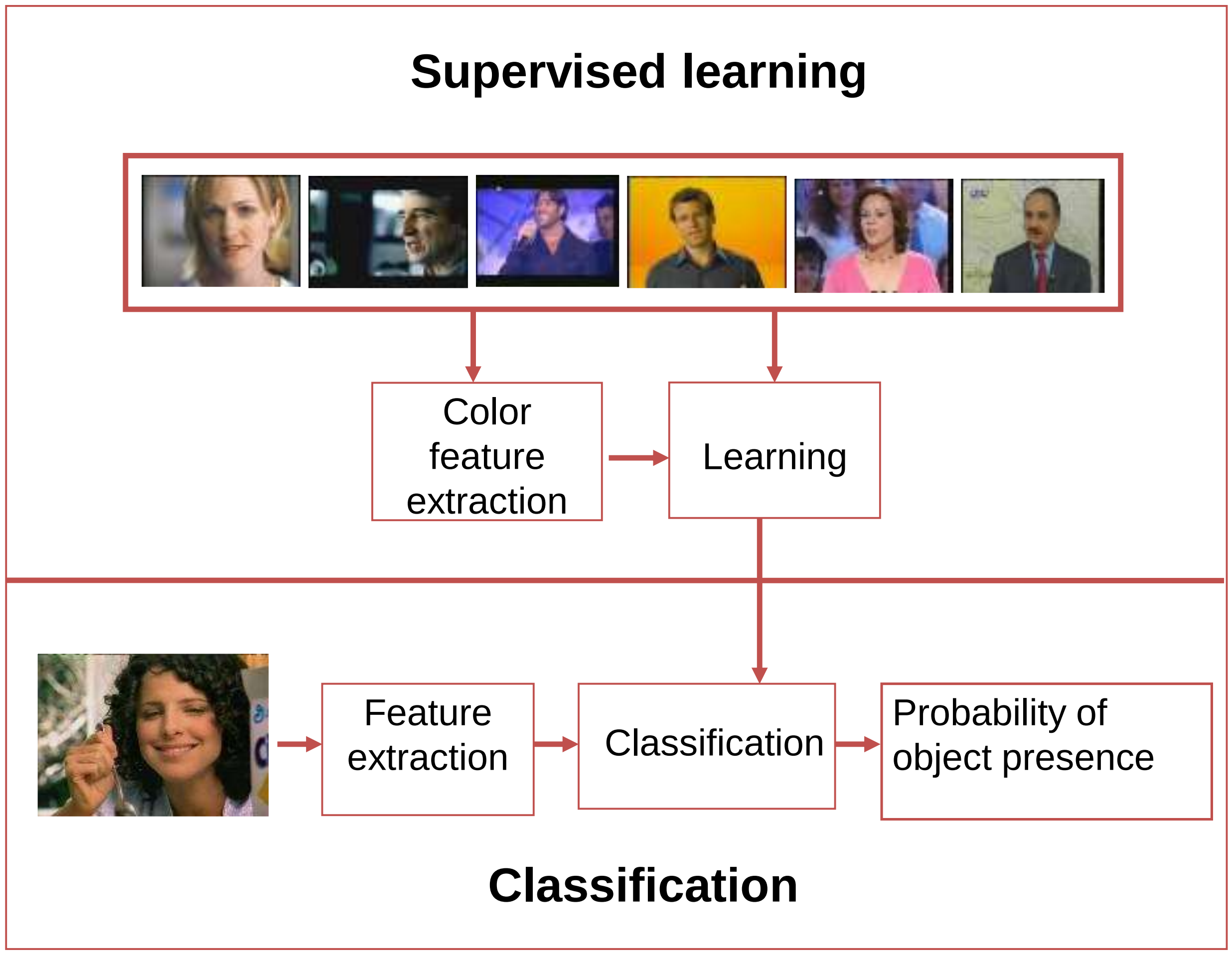
$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$

Maximum Margin Classification: SVM

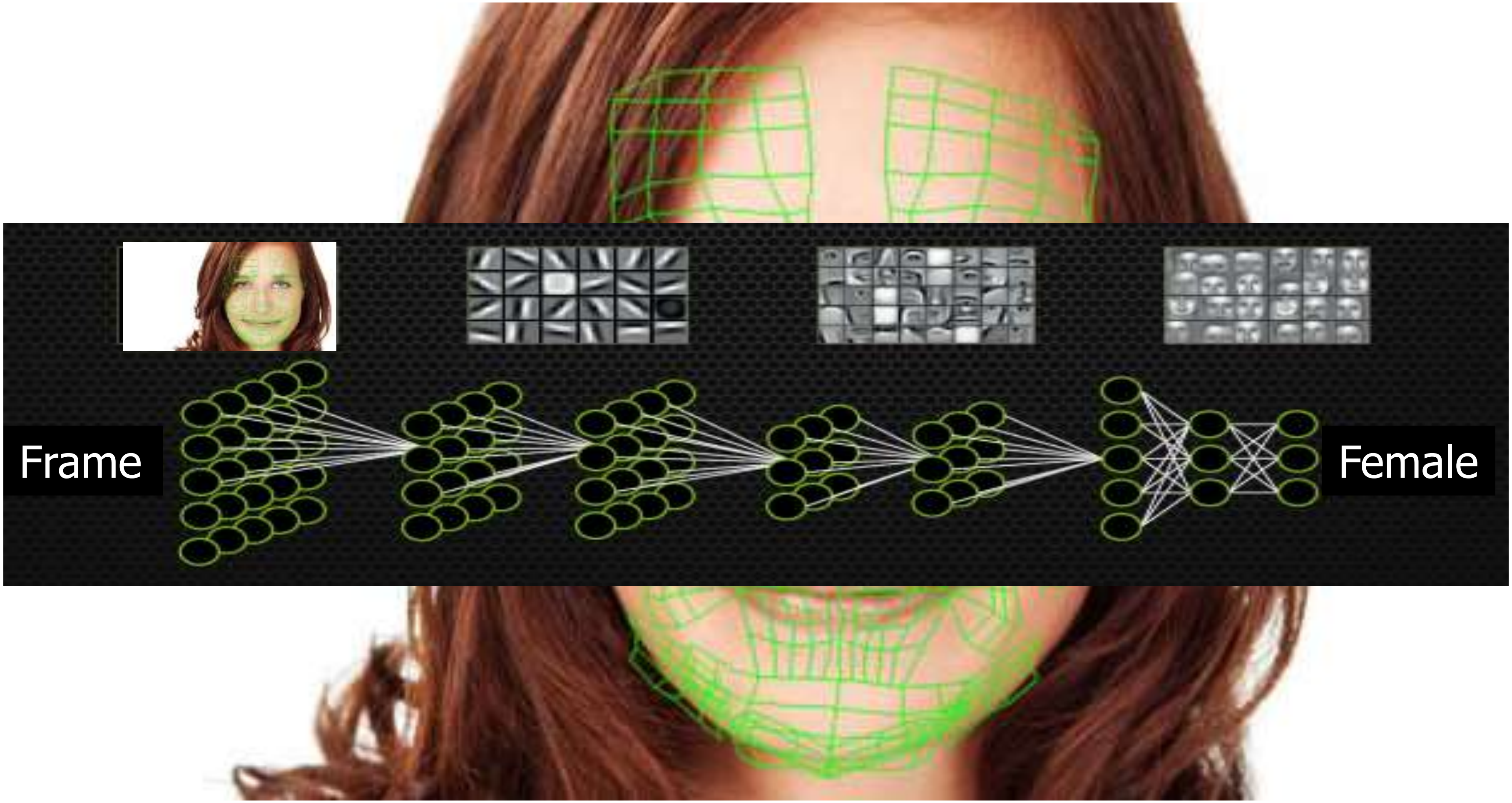
- Maximizing the margin is good.
- Implies that only support vectors matter; other training examples are ignorable.



Recognition scheme



Deep Learning : Convolutional Neural Networks



Real-world Datasets

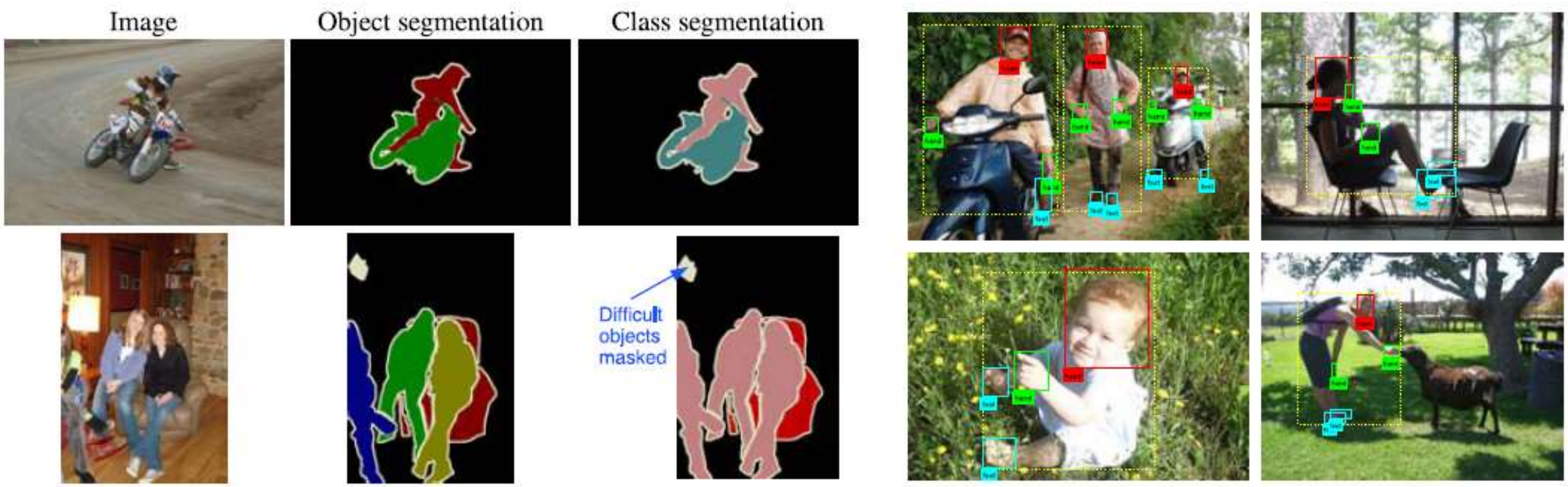
PASCAL VOC

Year	Statistics	New developments	Notes
2005	Only 4 classes: bicycles, cars, motorbikes, people. Train/validation/test: 1578 images containing 2209 annotated objects.	Two competitions: classification and detection	Images were largely taken from exising public datasets, and were not as challenging as the flickr images subsequently used. This dataset is obsolete.
2006	10 classes: bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep. Train/validation/test: 2618 images containing 4754 annotated objects.	Images from flickr and from Microsoft Research Cambridge (MSRC) dataset	The MSRC images were easier than flickr as the photos often concentrated on the object of interest. This dataset is obsolete.
2007	20 classes: • *Person:* person • *Animal:* bird, cat, cow, dog, horse, sheep • *Vehicle:* aeroplane, bicycle, boat, bus, car, motorbike, train • *Indoor:* bottle, chair, dining table, potted plant, sofa, tv/monitor Train/validation/test: 9,963 images containing 24,640 annotated objects.	• Number of classes increased from 10 to 20 • Segmentation taster introduced • Person layout taster introduced • Truncation flag added to annotations • Evaluation measure for the classification challenge changed to Average Precision. Previously it had been ROC-AUC.	This year established the 20 classes, and these have been fixed since then. This was the final year that annotation was released for the testing data.
2008	20 classes. The data is split (as usual) around 50% train/val and 50% test. The train/val data has 4,340 images containing 10,363 annotated objects.	• Occlusion flag added to annotations • Test data annotation no longer made public.	

PASCAL VOC

Year	Statistics	New developments	Notes
2009	20 classes. The train/val data has 7,054 images containing 17,218 ROI annotated objects and 3,211 segmentations.	• From now on the data consists of the previous years' images augmented with new images. In earlier years an entirely new data set was released each year. • Augmenting allows the number of images to grow each year, and means that test results can be compared on the previous years' images. • Segmentation becomes a standard challenge (promoted from a taster)	• No difficult flags were provided for the additional images (an omission). • Test data annotation not made public.
2010	20 classes. The train/val data has 10,103 images containing 23,374 ROI annotated objects and 4,203 segmentations.	• Action Classification taster introduced. • Associated challenge on large scale classification introduced based on ImageNet. • Amazon Mechanical Turk used for early stages of the annotation.	• Method of computing AP changed. Now uses all data points rather than TREC style sampling. • Test data annotation not made public.

PASCAL VOC



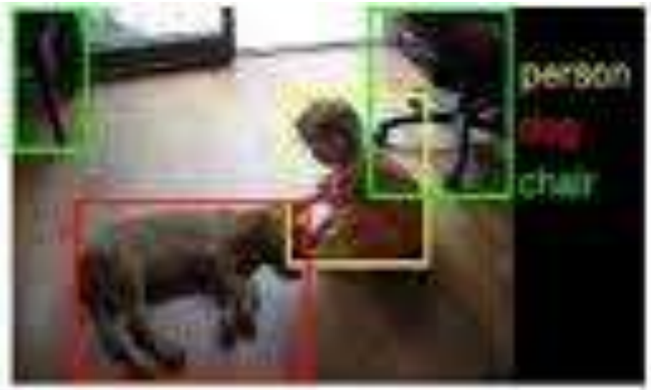
(a) Segmentation taster

(b) Person layout taster

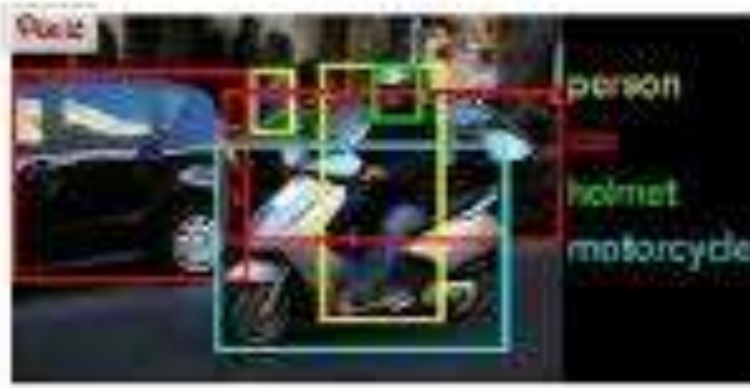
ImageNet

ImageNet


bird


person
chair


person
flower pot

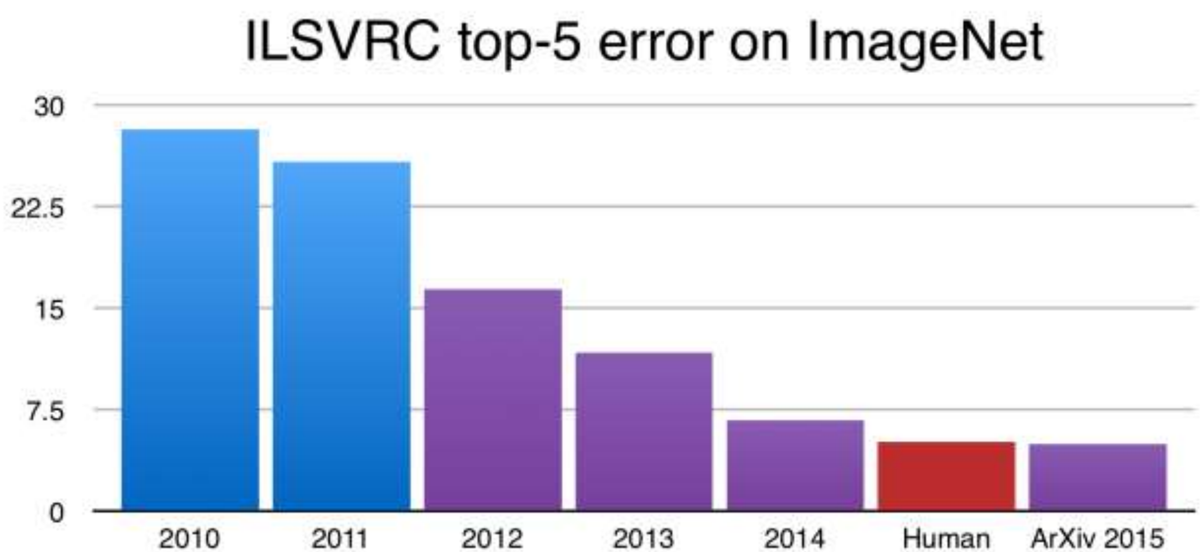

person
helmet
motorcycle

ImageNet

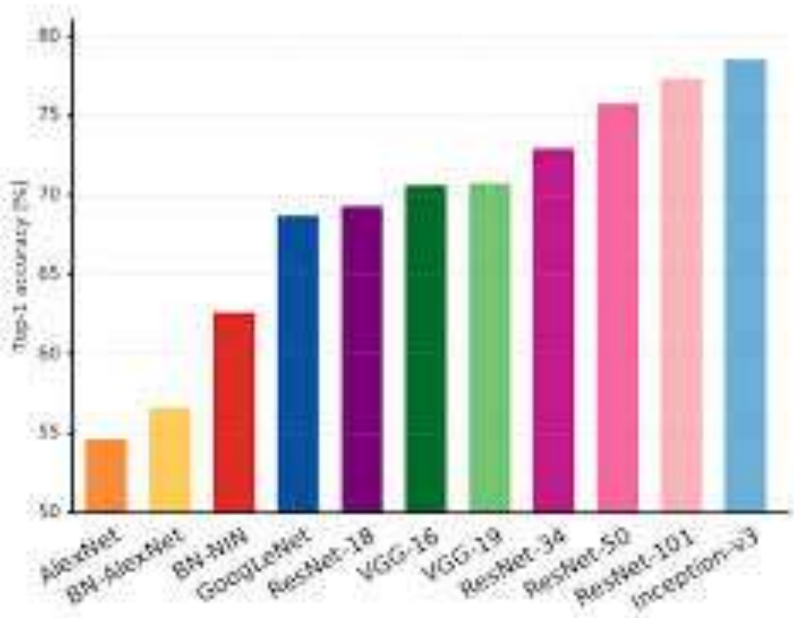
ImageNet Challenge

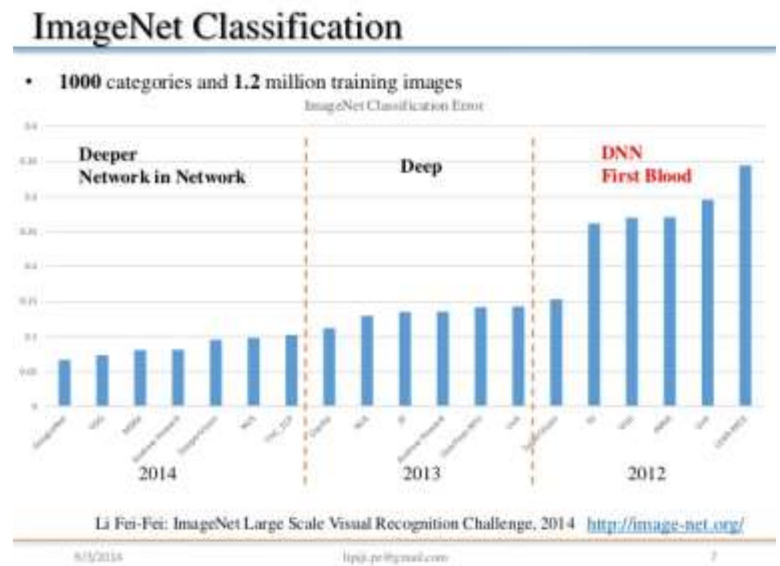
IMAGENET

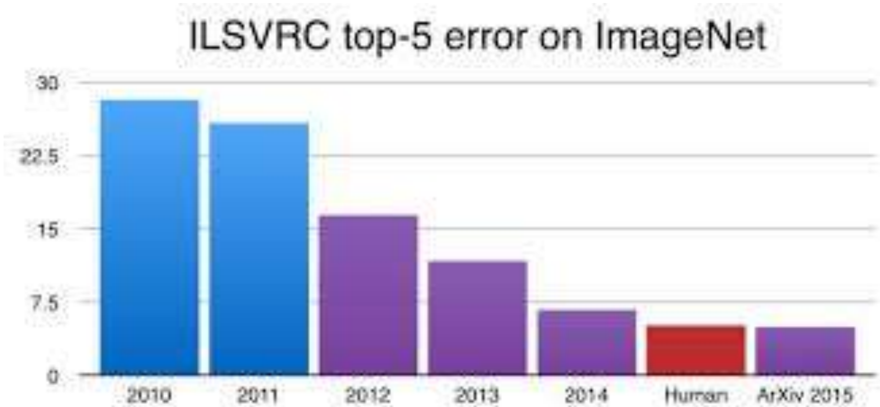
- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

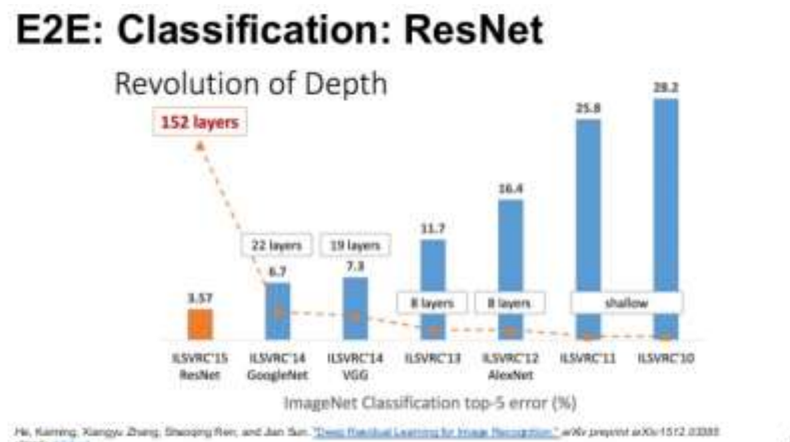


ImageNet Challenge 2012-2016



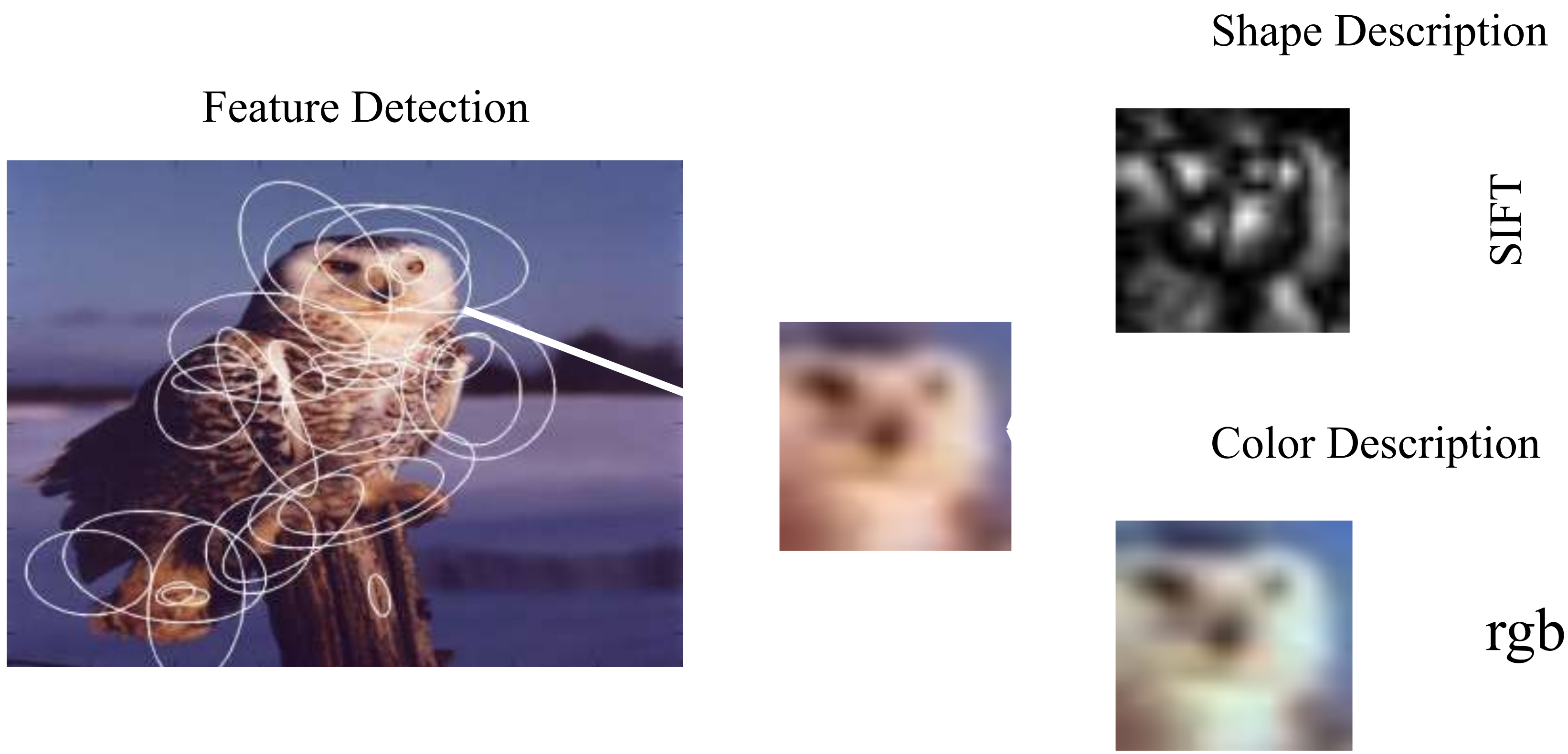






Feature Extraction and Object Recognition

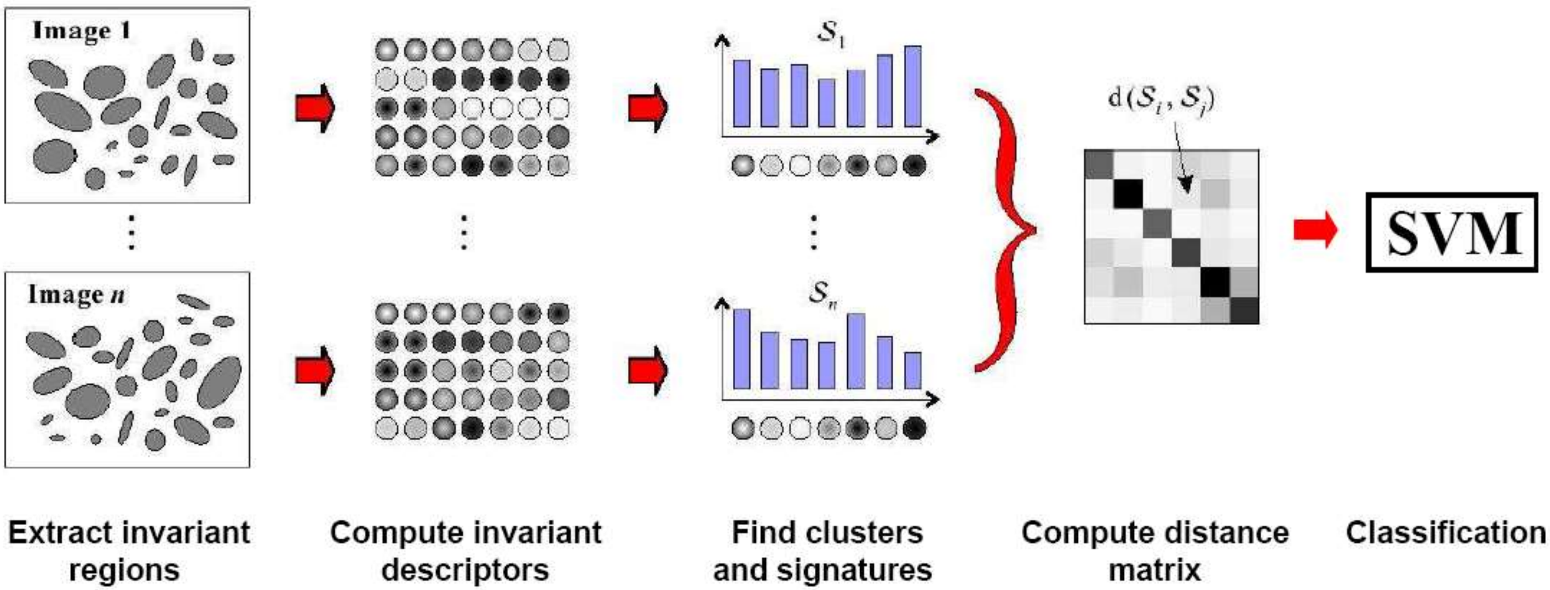
SIFT


Shape Description
Feature Detection
SIFT
Color Description
rgb

Invariance and Learning

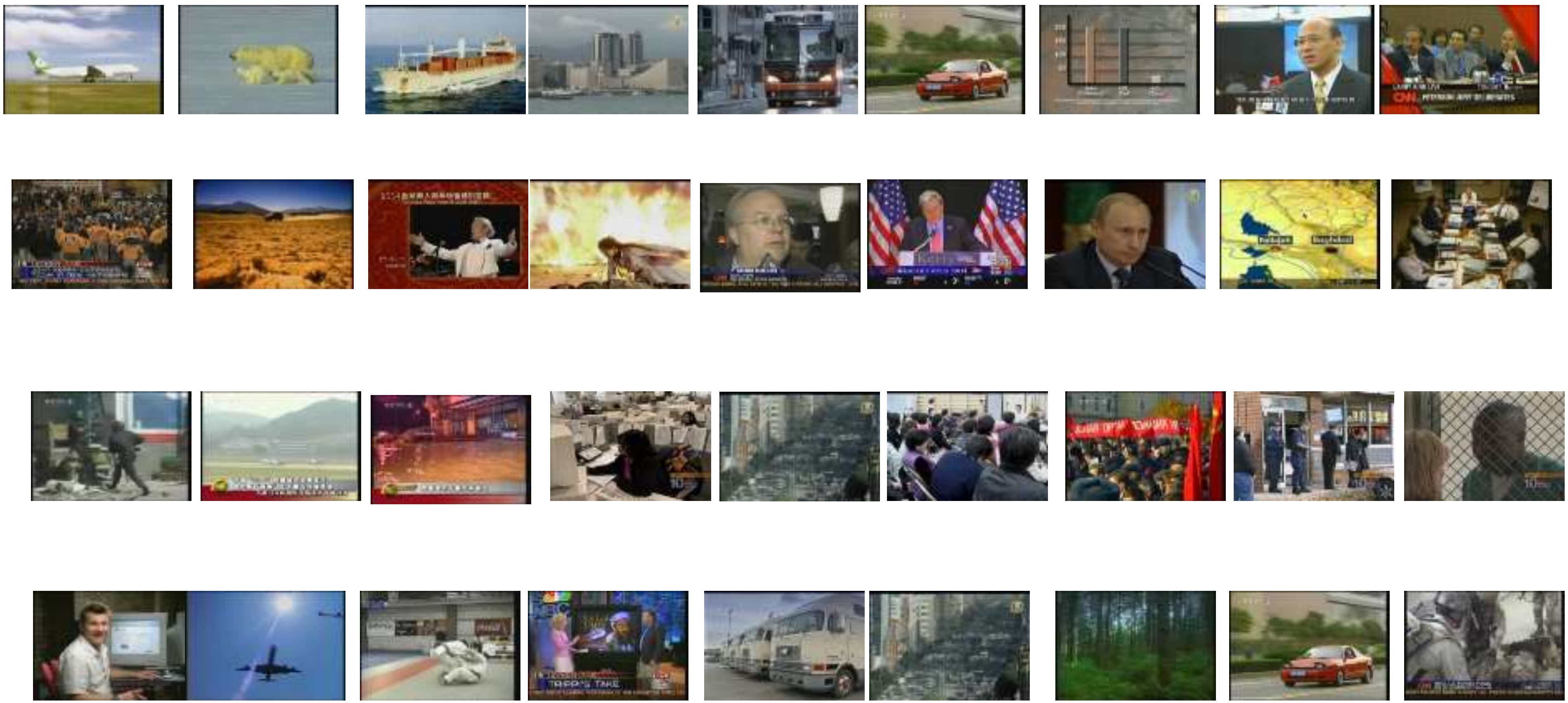
Learn from the content-based image retrieval field: Zhang (2005) achieved state of the art performance using local features:

- invariant region detectors
- SIFT descriptor



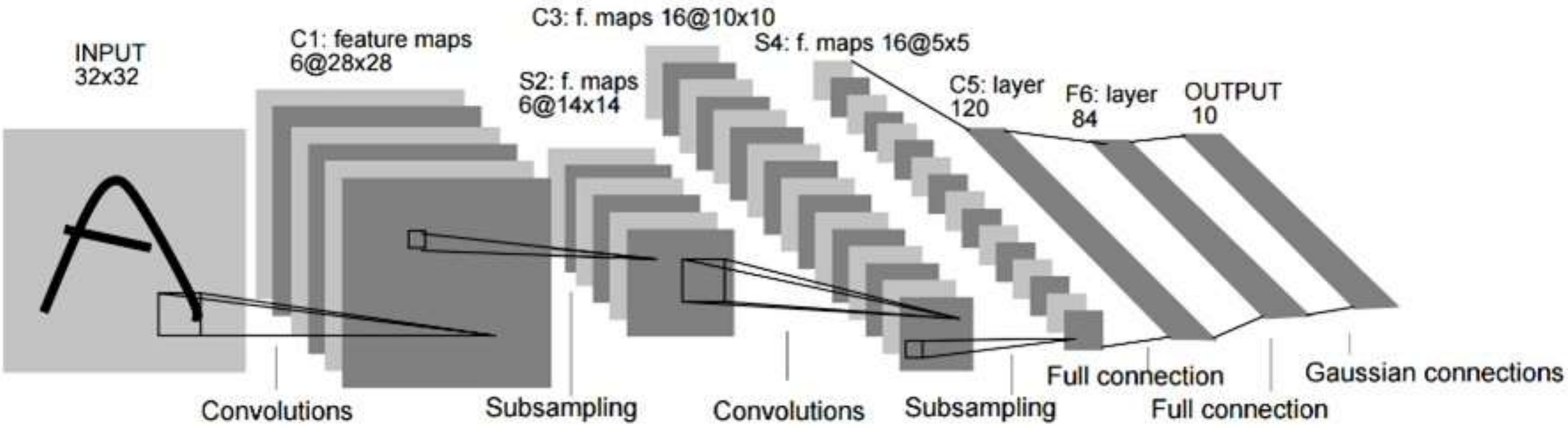
TrecVid Dataset: News Broadcasts

- 85 hours of data of TRECVID 2005
- Shot segmentation available: 43.907 shots
- Concept annotations available



39 concepts

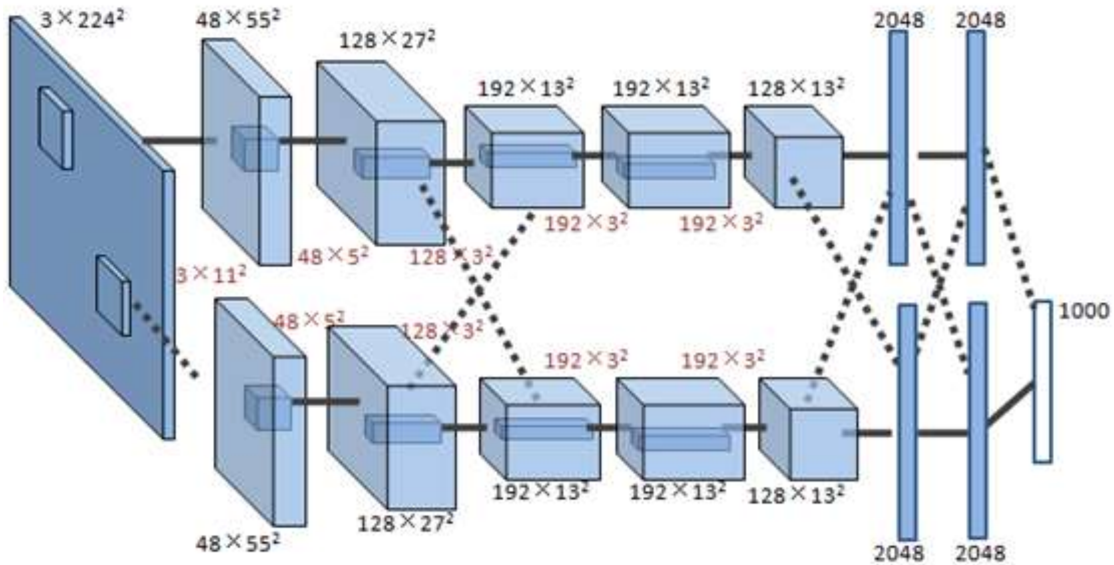
LeNet [LeCun et al. 1998]



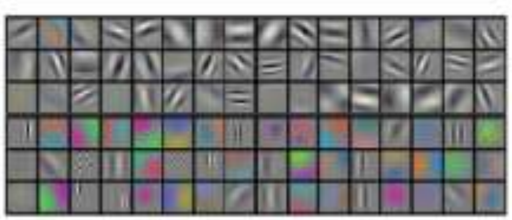
Gradient-based learning applied to document recognition [LeCun, Bottou, Bengio, Haffner 1998]

LeNet-1 from 1993

AlexNet

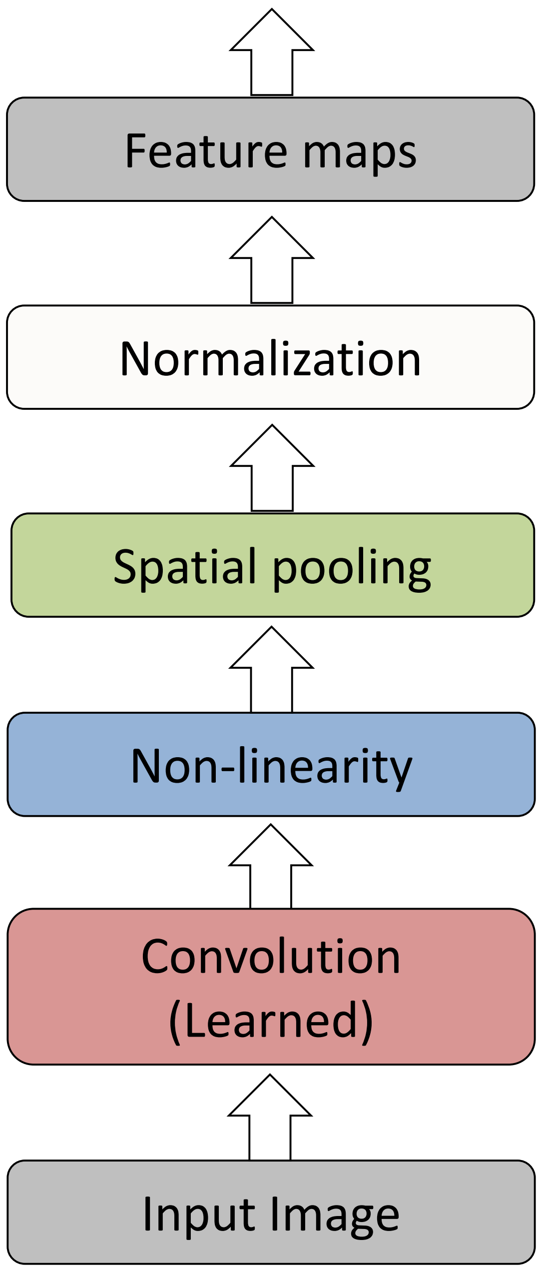


3 × 224²
48 × 55²
128 × 27²
192 × 13²
192 × 13²
128 × 13²
2048
2048
3 × 11²
48 × 5²
128 × 3²
192 × 3²
192 × 3²
1000
48 × 5²
128 × 3²
192 × 3²
192 × 3²
48 × 55²
128 × 27²
192 × 13²
192 × 13²
128 × 13²
2048
2048



mite
container ship
motor scooter
leopard
grille
mushroom
cherry
Madagascar cat

Convolutional Neural Networks



slide credit: S. Lazebnik

Learning and tracking

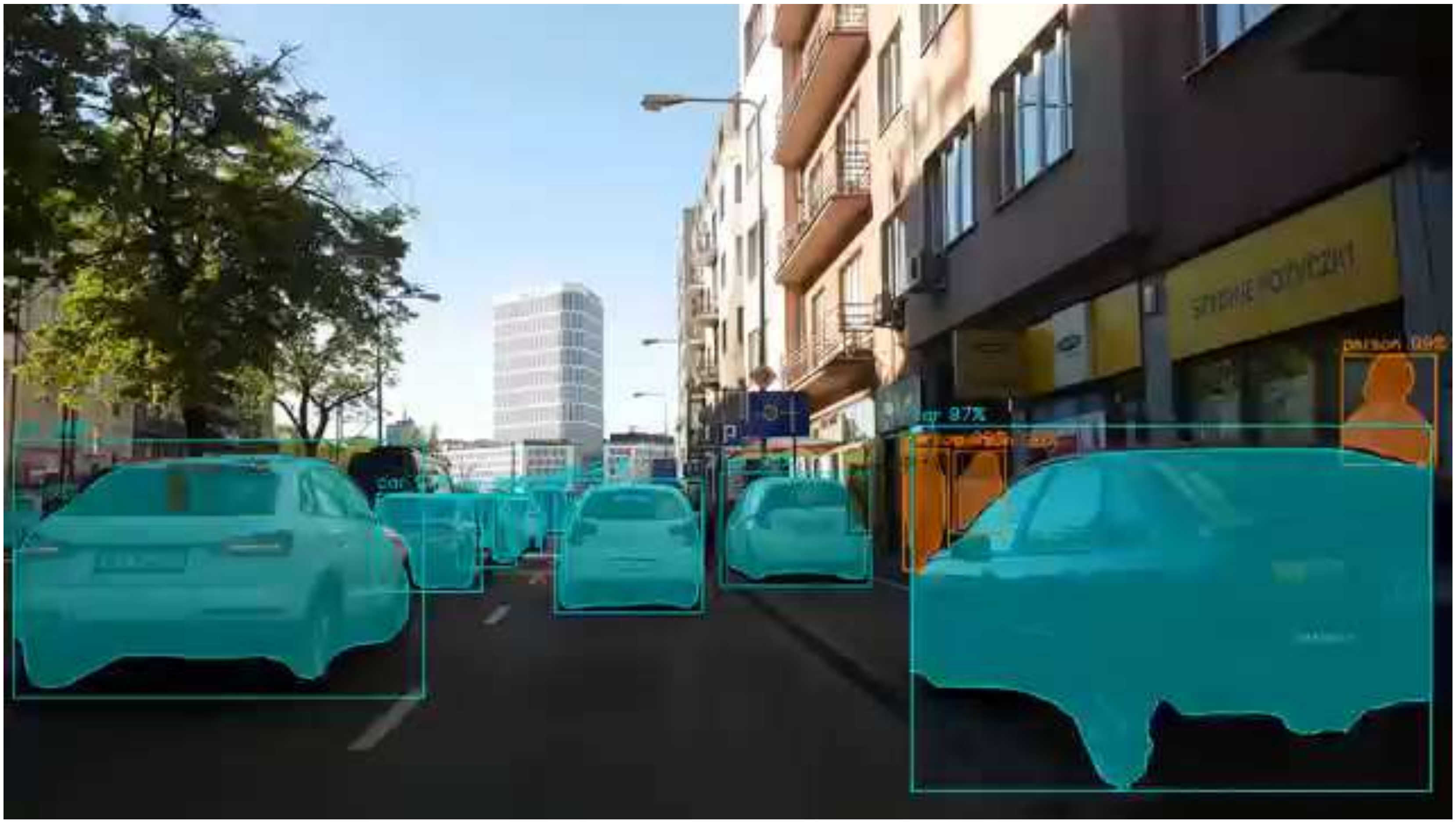
Tracking

Object Tracking

Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh
Carnegie Mellon University

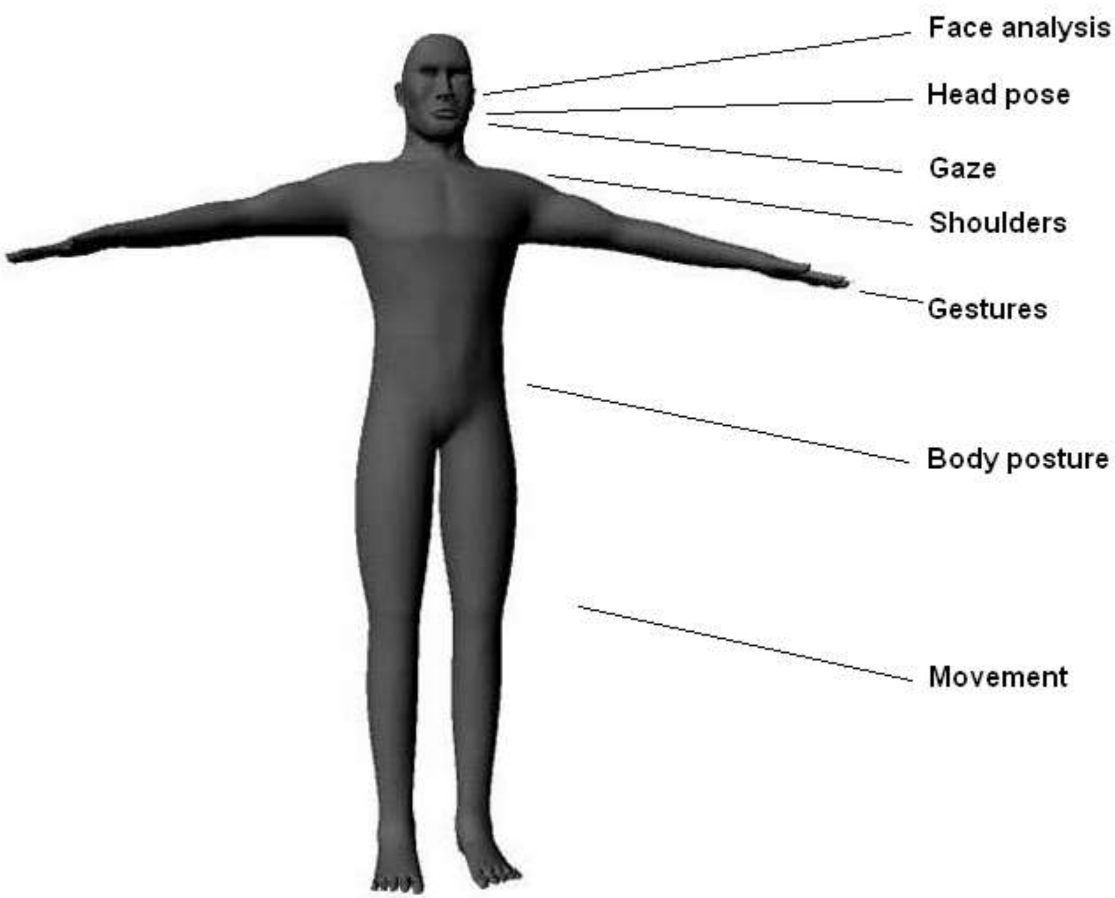
Semantic Segmentation and Tracking



Human behaviour analysis

Activity Recognition

Visual analysis of the human body




Buffy
Joyce
Dawn
Buffy
Buffy
Giles
Other
Tara
Willow
Willow
Buffy
Dawn
Willow
Tara
Willow
Willow
Buffy
Dawn
Tara
Buffy
Willow
Tara
Xander
Anya
Buffy
Riley
Spike

A Morphable Model For The Synthesis Of 3D Faces

Blanz and Thomas Vetter

Max-Planck-Institut fur biologische Kybernetik,

Tubingen, Germany

Face Analysis

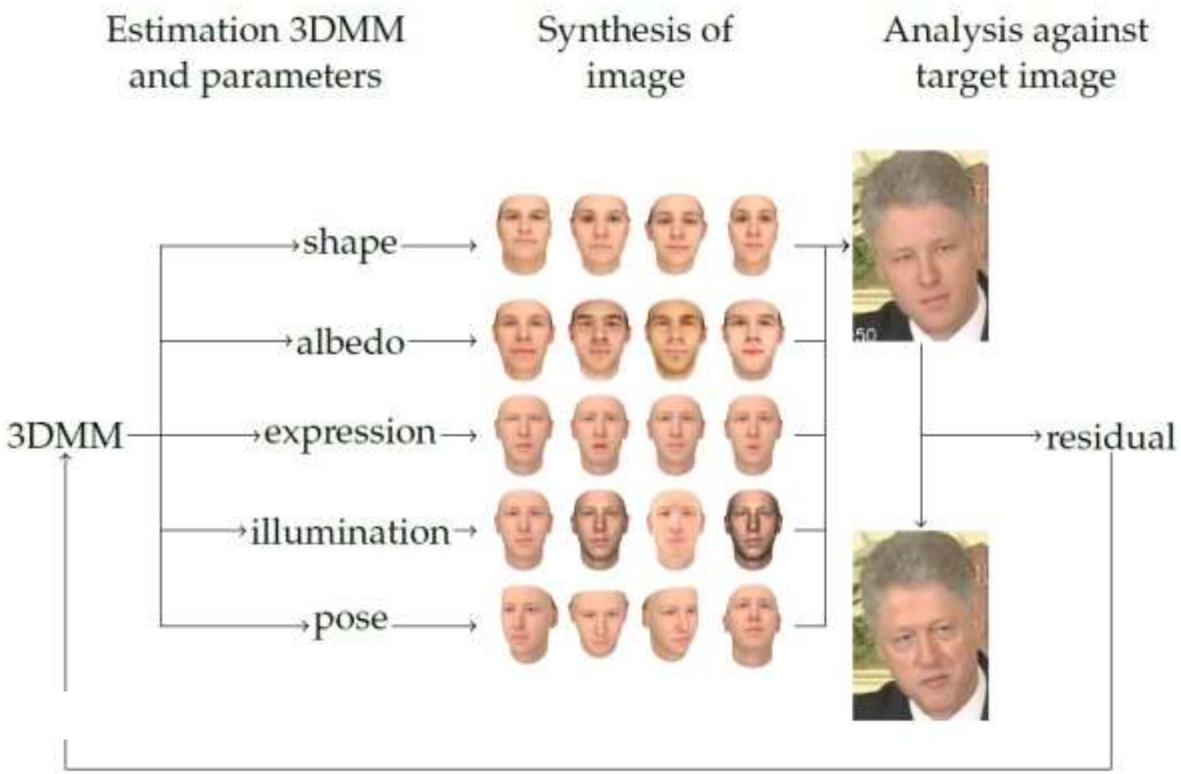
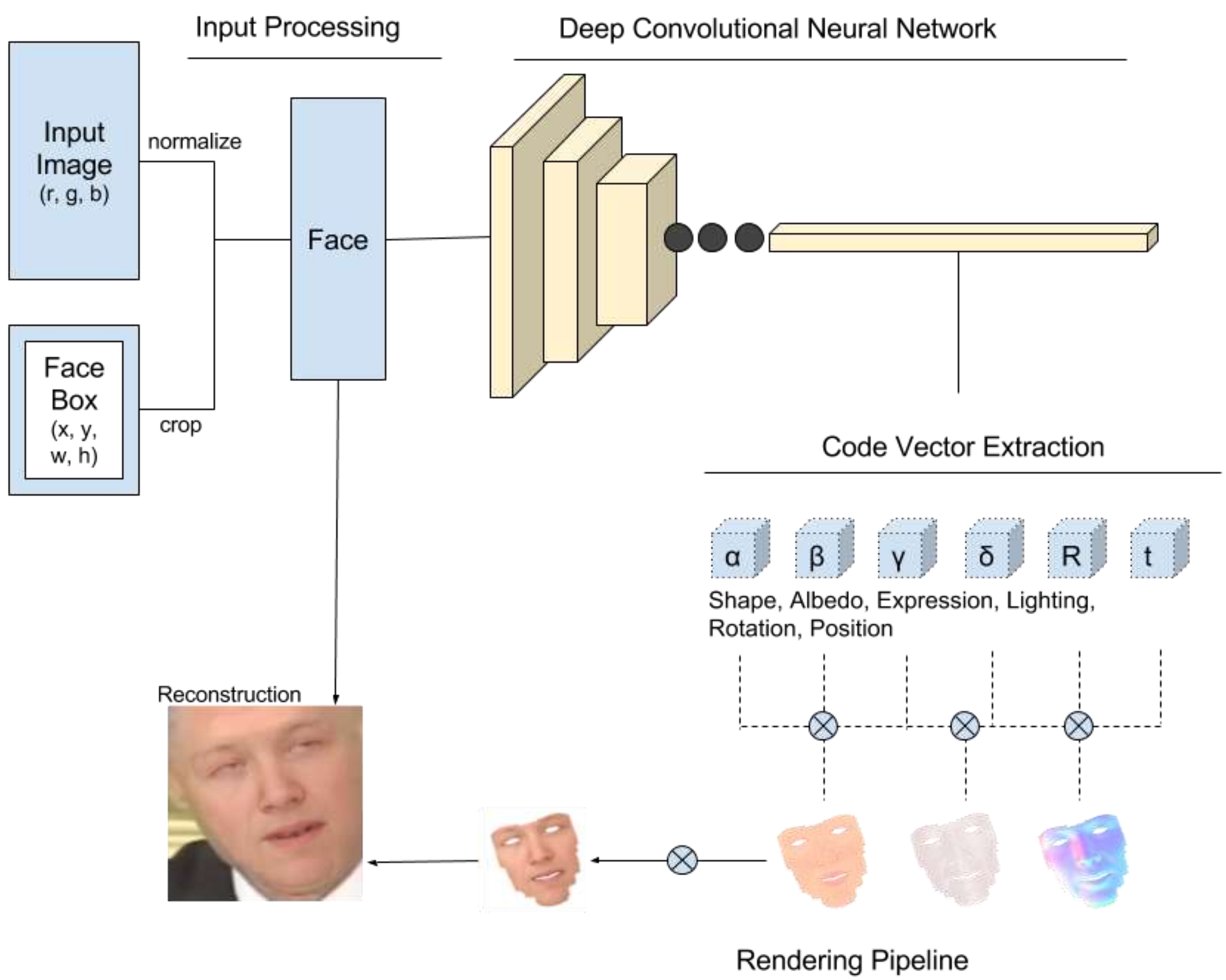


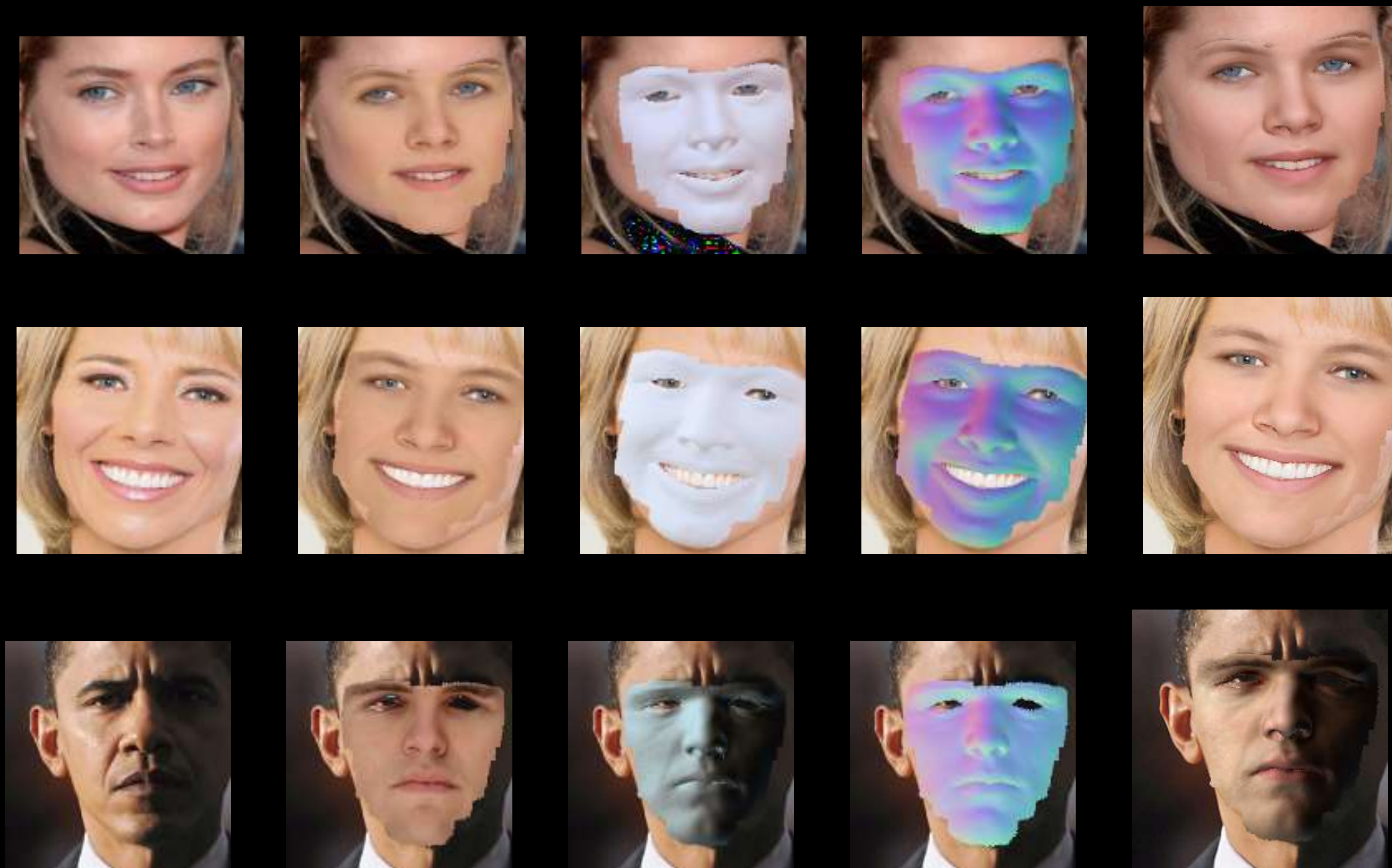
FIGURE 1.1: Example of 3DMM fitting following an analysis-by-synthesis approach. From left to right: The 3D Morphable Model, prediction of parameters for the different components of the optimization problem and rendered results, analyses against target image, and feedback loop.

CNN Model: Multi-Task Supervised



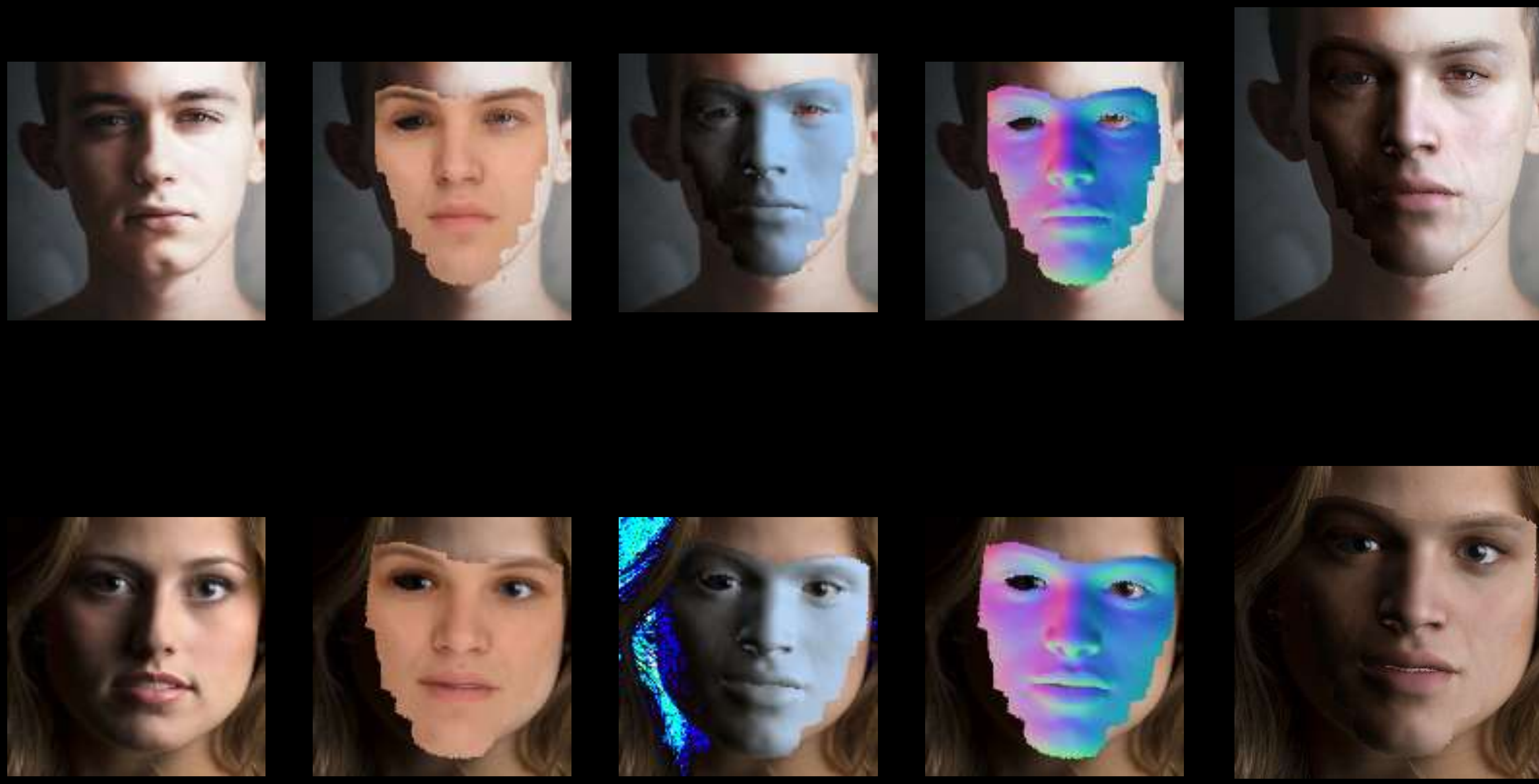
CNN Model: Results

Input Image | Albedo | Lighting | Surface Normals | Reconstruction



CNN Model: Results

Input Image | Albedo | Lighting | Surface Normals | Reconstruction



Consumer Applications

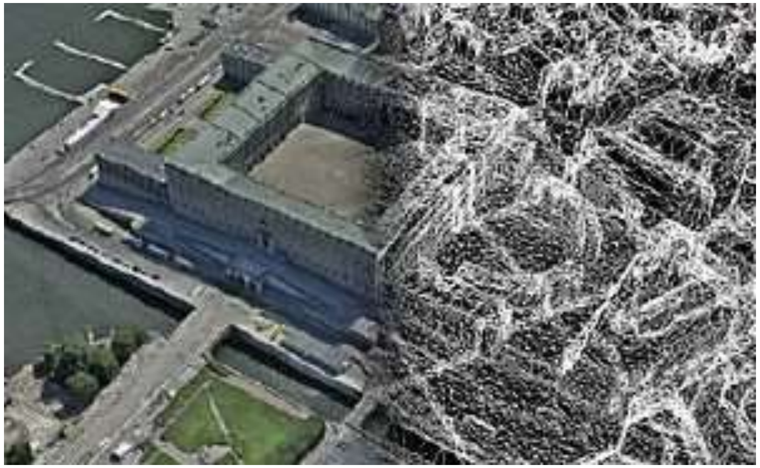
Face2Face: Real-time Face Capture and Reenactment of RGB Videos

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt
Matthias Nießner (University of Erlangen-Nuremberg, Max-Planck-Institute for Informatics, Stanford University)

Computer Vision:
Many more Applications

Style Transfer

3D from Images



Building Rome in a Day: Agarwal et al. 2009

Human shape capture

Human Shape Capture

Interactive Games: Kinect

- Object Recognition: http://www.youtube.com/watch?feature=iv&v=fQ59dXOo63o
- Mario: http://www.youtube.com/watch?v=8CTJL5lUjHg
- 3D: http://www.youtube.com/watch?v=7QrnwoO1-8A
- Robot: http://www.youtube.com/watch?v=w8BmgtMKFbY

Sports

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

http://www.sportvision.com/

Vision in space

NASA'S Mars Exploration Rover Spirit captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

Vision systems (JPL) used for several tasks

- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read "Computer Vision on Mars" by Matthies et al.

Computer Vision 1 (total # slides 88)

End of Lecture 1 | Summary

1. Specifics of this course

2. Introduction to Computer Vision
   What are Computer Vision Systems
   Computer vision and machine learning

3. Image search