High Level Computer Vision

Introduction
@ April 10, 2019

Bernt Schiele & Mario Fritz

www.mpi-inf.mpg.de/hlcv/

Max Planck Institute for Informatics & Saarland University, Saarland Informatics Campus Saarbrücken
Computer Vision and Multimodal Computing Group
@ Max-Planck-Institute for Informatics

Gerard Pons-Moll
Real Virtual Humans

Paul Swoboda
Combinatorial Vision Group

Bernt Schiele
Computer Vision

Zeynep Akata
Multimodal Deep Learning
U Amsterdam

Mario Fritz
Scalable Learning & Perception
CISPA Helmholtz Center i.G.

Max Planck Institute for Informatics & Saarland University,
Saarland Informatics Campus Saarbrücken
Computer Vision

• Lecturer:
  ‣ Bernt Schiele (schiele@mpi-inf.mpg.de)
  ‣ Mario Fritz (mfritz@mpi-inf.mpg.de)

• Assistants:
  ‣ Yang He (yang@mpi-inf.mpg.de)
  ‣ Rakshith Shetty (rshetty@mpi-inf.mpg.de)

• Language:
  ‣ English

• mailing list for announcements etc.
  ‣ send email (see instructions on the web)
    Rakshith Shetty <rshetty@mpi-inf.mpg.de>
Lecture & Exercise

- Officially: 2V (lecture) + 2Ü (exercise)
  - Lecture: Wed: 10:15am - 12pm (room 024)
  - Exercise: Mon: 10:15am - 12pm (room 024)

- typically 1 exercise sheet every 1-2 weeks
  - part of the final grade
  - some pencil and paper, mostly practical including a project
  - larger project in second half of lecture
    - we/you propose projects, mentoring, final presentation

- 1. exercise is Python tutorial

- Exam
  - oral exam (grading 50% oral exam and 50% exercises)
  - after the SS - there will be proposed dates
Material

• For "non-deep-learning" parts of the lecture:
  ‣ available online
    http://szeliski.org/Book

• Background on deep learning:
  Deep Learning Book
  ‣ available online
    http://deeplearning.org
Why Study Computer Vision

• Science
  ‣ Foundations of perception. How do WE as humans 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://www.cclondon.com/imagingandcameras.html)

- **Surveillance**
  - Face Recognition
  - Airport Security (People Tracking)

- **Medical Imaging**
  - (Semi-)automatic segmentation and measurements

- **Autonomous Driving & Robotics**
More Applications

- Vision on Cellphones:
  - e.g. Google Goggles
- Vision for Interfaces:
  - e.g. Microsoft Kinect
- Reconstruction
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

• Role of Deep Learning
  - with several slides taken from Fei-Fei Li, Justin Johnson, Serena Yeung @ Stanford

• 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

- high intra-class variation
- low inter-class variation
- high intra-class variation
Sample Category: Motorbikes
Basic Idea

I know where the Eiffel Tower is.

global

local

I know where the Eiffel Tower is.
Video...
Articulation Model

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

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

400 annotated training images
Modeling Body Dynamics

- Visualization of the hierarchical Gaussian process latent variable model (hGPLVM)
Our Subgraph Multicut Tracking Results

- Detection Hypotheses
- Tracklet Hypotheses
- Hypotheses Decomposition
- Final Tracks

Dotted rectangles are interpolated tracks.
More Results

**Decompositions (clusters)**

**Tracks**

Dotted rectangles are interpolated tracks.
More Results

Decompositions (clusters)

Tracks

Dotted rectangles are interpolated tracks.
Deep Learning have become an important tool for object recognition (and other computer vision tasks)

Let's briefly discuss CNNs (Convolutional Neural Networks)
Ingredients for Deep Learning

Algorithms

Data

Computation

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
GigaFLOPs per Dollar

- **CPU**
- **GPU**

Deep Learning Explosion

- GTX 1080 Ti
- GeForce GTX 580
  - (AlexNet)
- GeForce 8800 GTX

Time:
- 1/2004
- 10/2006
- 7/2009
- 4/2012
- 12/2014
- 9/2017

Slide credit: Fei-Fei, Justin Johnson, Serena Yeung
GigaFLOPs per Dollar

- CPU
- GPU
- TPU

TITAN V (Tensor Cores)

Deep Learning Explosion

GTX 1080 Ti

GeForce GTX 580 (AlexNet)

GeForce 8800 GTX

Time:
- 1/2004
- 10/2006
- 7/2009
- 4/2012
- 12/2014
- 9/2017

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
IMAGENET

22K categories and 14M images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate

- Plants
  - Tree
  - Flower

- Food

- Materials

- Structures
  - Artifact
  - Tools
  - Appliances
  - Structures

- Person
  - Scenes
    - Indoor
    - Geological Formations
    - Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
The Image Classification Challenge:
1,000 object classes
1,431,167 images

Output:
- Scale
- T-shirt
- Steel drum
- Drumstick
- Mud turtle

Output:
- Scale
- T-shirt
- Giant panda
- Drumstick
- Mud turtle

Russakovsky et al. IJCV 2015

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
Validation classification
Validation classification

- lens cap
  - reflex camera
  - Polaroid camera
  - pencil sharpener
  - switch
  - combination lock

- abacus
  - abacus
  - typewriter keyboard
  - space bar
  - computer keyboard
  - accordion

- slug
  - slug
  - zucchini
  - ground beetle
  - common newt
  - water snake

- hen
  - cocker spaniel
  - partridge
  - English setter

- tiger
  - tiger cat
  - tabby
  - boxer
  - Saint Bernard

- chambered nautilus
  - lampshade
  - throne
  - goblet
  - table lamp
  - hamper

- tape player
  - cellular telephone
  - slot
  - reflex camera
  - dial telephone
  - iPod

- planetarium
  - dome
  - mosque
  - radio telescope
  - steel arch bridge
The Image Classification Challenge:
1,000 object classes
1,431,167 images

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
IMAGENET Large Scale Visual Recognition Challenge

Year 2010
NEC-UIUC

- Dense descriptor grid: HOG, LBP
- Coding: local coordinate, super-vector
- Pooling, SPM
- Linear SVM

[Lin CVPR 2011]

Year 2012
SuperVision

Year 2014
GoogLeNet

- Pooling
- Convolution
- Other

Year 2015
MSRA

[He ICCV 2015]

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
AlexNet (2012)

- 5 convolutional layers
- 3 fully-connected layers
How deep is enough?

How deep is enough?

VGG-M (2013)
AlexNet (2012)

16 convolutional layers
50 convolutional layers
152 convolutional layers


Convolutional Neural Networks (CNNs) were not invented overnight...
1998
LeCun et al.

K

Input

Image Maps

Convolutions

Subsampling

Output

Fully Connected

# of transistors

10^6

# of pixels used in training

10^7

2012
Krizhevsky et al.

# of transistors

10^9

GPUs

# of pixels used in training

10^{14}

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
Try it out yourself

- Caffe ist an open implementation from the Berkeley Vision Group
  - http://caffe.berkeleyvision.org
  - http://demo.caffe.berkeleyvision.org
Deep Learning
have become an important tool
for object recognition / image classification

but there exist many other computer vision tasks
where Deep Learning is also an essential ingredient

a few examples...
Human Pose Estimation

- **Single Person Pose Estimation** - two “phases”

  - **Phase 1**: *pictorial structures models* e.g.
    [Felzenszwalb&Huttenlocher@ijcv05],
    [Andriluka&al@ijcv11], [Yang&Ramanan@pami13],
    [Pishchulin&al@iccv13], …

  - **Phase 2**: using *deep learning* e.g.
    [Thoshev,Szegedy@cvpr14], [Thompson&al@nips14],
    [Chen&Yuille@nips14], [Carreira&al@cvpr16],
    [Hu&Ramanan@cvpr16], [Wei&al@cvpr16],
    [Newell&al@cvpr16], …
MPII Human Pose Dataset: Dataset demo

- 410 human activities (after merging similar activities)
- over 40,000 annotated poses
- over 1.5M video frames

http://human-pose.mpi-inf.mpg.de/
Analysis - overall performance

Best Methods today:
deep learning “takes” over

✓ since CVPR’14, dataset has become de-facto standard benchmark
✓ large training set facilitated development of deep learning methods
Cityscapes: Large-Scale Datasets for Semantic Labeling of Street Scenes

- Joint effort of:

[Cordts, Omran, Ramos, Rehfeld, Enzweiler, Benenson, Franke, Roth, Schiele@cvpr16]

Classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>road</td>
<td>ground</td>
</tr>
<tr>
<td>sidewalk</td>
<td></td>
</tr>
<tr>
<td>car(^1)</td>
<td></td>
</tr>
<tr>
<td>truck(^1)</td>
<td></td>
</tr>
<tr>
<td>bus(^1)</td>
<td></td>
</tr>
<tr>
<td>on rails(^1)</td>
<td>vehicle</td>
</tr>
<tr>
<td>motorcycle(^1)</td>
<td></td>
</tr>
<tr>
<td>bicycle(^1)</td>
<td></td>
</tr>
<tr>
<td>license plate(^2)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>building</td>
<td></td>
</tr>
<tr>
<td>wall</td>
<td></td>
</tr>
<tr>
<td>fence</td>
<td></td>
</tr>
<tr>
<td>traffic sign</td>
<td>infra-structure</td>
</tr>
<tr>
<td>traffic light</td>
<td></td>
</tr>
<tr>
<td>pole</td>
<td></td>
</tr>
<tr>
<td>bridge(^2)</td>
<td></td>
</tr>
<tr>
<td>tunnel(^2)</td>
<td></td>
</tr>
<tr>
<td>sky</td>
<td>sky</td>
</tr>
</tbody>
</table>

\(^1\) Single instance annotation available
\(^2\) Not included in fine label set challenges
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

Ours: a person on skis jumping over a ramp
Ours: a skier is making a turn on a course
Ours: a cross country skier makes his way through the snow
Ours: a skier is headed down a steep slope

Baseline: a man riding skis down a snow covered slope

[Rakshith’17]
Towards a Visual Turing Challenge

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

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

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is on the refrigerator?
Vision + Language: magnet, paper
Language Only: magnet, paper

What is the colour of the table?
Vision + Language: blue, white
Language Only: blue, green, red, yellow

How many towels are hanged?
Vision + Language: 3
Language Only: 4

What is hanged on the chair?
Vision + Language: clothes
Language Only: jacket

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

How many chairs are there?
Vision + Language: 1
Language Only: 4

What is fixed on the window?
Vision + Language: curtain
Language Only: curtain

What is the object close to the sink?
Vision + Language: faucet
Language Only: faucet

What are the things on the cabinet?
Vision + Language: photo
Language Only: photo

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

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the sink?
Vision + Language: sink
Language Only: faucet

What is the object close to the table?
Vision + Language: lamp
Language Only: plant

What is the colour of the table and chair?
Vision + Language: brown
Language Only: brown

What is the object close to the counter?
Vision + Language: sink
Language Only: stove

What is the colour of the table and chair?
Video Object Segmentation

Goal: Separating a specific foreground object from background in a video given its 1st frame mask annotation.

DAVIS 2016
[Perazzi et al.'16]
MaskTrack - Proposed Approach

—we process video per-frame, using guidance from previous frame

Frame \( t-1 \)
output mask

Frame \( t \)
input

MaskTrack

DeepLab
[Chen et al., ICLR’15]

Frame \( t \)
output mask

—we want to train from static images only
Qualitative Results

https://www.mpi-inf.mpg.de/masktrack
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)

  “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”
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’ …?
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
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
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
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 brightness
  - Foreground/Background Boundaries
  - Object-Object-Boundaries
  - Shadow Edges
  - Changes in Albedo or Texture
  - Changes in Surface Normals
Line Drawings: Good Starting Point for Recognition?
The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".
VISION

David Marr

David Marr, 1970s

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
Stages of Visual Representation, David Marr, 1970s

- Input Image
- Perceived intensities
- Primal sketch: Zero crossings, high edges, lines, curves, boundaries
- 2½-D sketch: Local surface orientation
- 3-D model: 3-D models hierarchically organized in terms of surface and volumetric primitives

slide credit: Fei-Fei, Justin Johnson, Serena Yeung
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
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 & Segmentation

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

- 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

- 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)
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

• Role of Deep Learning
  - with several slides taken from Fei-Fei Li, Justin Johnson, Serena Yeung @ Stanford

• 2 case studies:
  ‣ Recovery of 3D structure
    - slides taken from Michael Black @ Brown University / MPI Intelligent Systems
  ‣ Object Recognition
    - intuition from human vision...