Computer Vision I - Algorithms and Applications: Introduction

Carsten Rother

22/10/2013
Admin Stuff

- Language: German/English; Slides: English (all the terminology and books are in English)
- Lecturer: Carsten Rother
- Exercises: Dmitri Schlesinger
- Staff Email: carsten.rother@tu-dresden.de
- Announcements: online (to be set up)

Course Books:
- Also pointers to conference and journal articles
Course Overview (total 14 lectures)

VL1 (21.10): Introduction
Ex1: Intro to OpenCV (H.Heidrich)

Ex2: Fast Methods for Filtering; explain home work

VL3 (4.11): Image Formation Models
Ex3: homework

VL4 (11.11): Single View Geometry and Camera Calibration
Ex4: homework

VL5 (18.11): Feature Extraction and Matching
Ex5: homework

VL6 (25.11): Two View Geometry
Ex6: Robust Panoramic Stitching; explain homework

VL7 (2.12): 3D reconstruction from multiple views
Ex7: homework
Course Overview (total 14 lectures)

VL8 (9.12): Dense Motion Estimation
Ex1: homework

VL9 (16.12): Image Segmentation
Ex2: homework

VL10 (6.1): Object Recognition (in progress)
Ex3: Object Recognition; explain homework

Ex4: homework

VL12 (20.1): Model-based Vision (in progress)
Ex5: homework

VL13 (27.1): Having Fun with Images (in progress)
Ex6: homework

VL14 (3.2): Wrap-Up: 100 things we have learned (in progress)
Ex5: homework
Related Lectures

• Machine Learning (D. Schlesinger) WS 13/14 (2/2/0)
  • Basics on machine learning from data (not necessarily image)

• Computer Vision 2: Models, Inference and Learning SS 14 (4/2/0)
  • Requirement: ML WS 13/14 and CV WS 13/14
  • Goes mathematically deeper

• For doing a Thesis/PhD in the CVLD all three classes are compulsory

• Computer graphics (Prof. Gumhold) (Introduction, I, II)
  3D Scanning with structured light; Illumination models; Geometry
Exams and Exercises

- Exam: in person (maybe written exams in future)

- Exercises/homework:
  - There are 3 blocks
  - Each block has several exercises with different points
  - You have to collect in total 10 points to sit the exam
  - The exercises have to be handed in until end of semester (ideally after each block)
  - Last possible date to hand in is end of semester (end of January)

- Collaboration:
  - You are encouraged to discuss the topics
  - You are not allowed to copy any code for the homework from other people
Before we start ... some Advertisement

**CVLD Overview**

Interactive Image and Data manipulation

Applied Optimization, Models, and Learning

3D Scene Understanding

Inverse rendering from moving images

Benchmarking and Label collection

Image Analysis for System Biology
We have a lot of different topics to offer: ranging from theoretical to practical ones

Main emphasis:

- 3D Image Editing
- 3D Scene understanding
- Learning and Inference in undirected Graphical models

A list of topics will be available soon on: http://www.inf.tu-dresden.de/index.php?node_id=1864&ln=en
A project work in the CVLD is a good stepping stone if you:

- want to do a PhD in computer vision, graphics, machine learning
- want to become a researcher or software developer in one of the big research labs (Microsoft Research, Google, Adobe, TechniColor, etc)
- If you are interested in doing a start-up
- Other “computer vision related” industry

**Example:** Master Thesis on “Video Matting” jointly with Adobe Seattle (e.g. for Adobe Adobe After Effects)
• Komplex Praktikum:
  • Segmentation in 3D images (3D Image Editing)
  • real-time Pose estimation (3D scene understanding)

• Einfuehrungspraktikum:
  6 Exercises to get to know Computer Vision (ranging from physics based vision to semantic vision)
Before we start....

• Please give Feedback – during the lecture, after the lecture or via email

• This my first full course I give at University

• Feedback helps me to adjust the level.
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 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”?

**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

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

Very many sources of variability

Image
Image-formation model

Scene type

Street scene

Scene geometry
Image-formation model

Scene type

Scene geometry

Object classes

<table>
<thead>
<tr>
<th>Street scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sky</td>
</tr>
<tr>
<td>Building×3</td>
</tr>
<tr>
<td>Road</td>
</tr>
<tr>
<td>Sky</td>
</tr>
<tr>
<td>Sidewalk</td>
</tr>
<tr>
<td>Tree×3</td>
</tr>
<tr>
<td>Person×4</td>
</tr>
<tr>
<td>Bicycle</td>
</tr>
<tr>
<td>Car×5</td>
</tr>
<tr>
<td>Bollard</td>
</tr>
<tr>
<td>Bench</td>
</tr>
</tbody>
</table>

[Slide Credits: John Winn, ICML 2008]
### Image-formation model

<table>
<thead>
<tr>
<th>Scene type</th>
<th>Scene geometry</th>
<th>Object classes</th>
<th>Object position</th>
<th>Object orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Street scene</td>
<td>Sky</td>
<td>Sky</td>
<td>Building×3</td>
<td>Road</td>
</tr>
<tr>
<td></td>
<td>Sidewalk</td>
<td>Tree×3</td>
<td>Road</td>
<td>Person×4</td>
</tr>
<tr>
<td></td>
<td>Bicycle</td>
<td>Car×5</td>
<td>Pedestrian×4</td>
<td>Bench</td>
</tr>
<tr>
<td></td>
<td>Bollard</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Slide Credits: John Winn, ICML 2008]
Image-formation model

Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape

Street scene

[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
The “Scene Parsing” challenge --- a “grand challenge” of computer vision

(Probabilistic) Script = \{\text{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 Vision can be seen as “inverse graphics”
Example of a recent work

[Image of a building and sky]

Scene graph

Input

[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.

[Input Image] [3D Cuboid Detector] [Output Detection Result] [Synthesized New Views]

[Xiao et al. NIPS 2012]

[Sussman, Lamport, Guzman 1966]

[Slide credits Andrew Blake]
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?

- 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 is efficiently is open challenged (see later)
How can we interpret visual data?

- What general (prior) knowledge of the world (not necessarily visual) can be exploited?
- 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 efficiently is open challenged (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 – harder to describe

• Describe Image Texture

Real Image  zoom  Not a real Image  zoom

• Microscopic Images. What is the true shape of these objects
Example: State-of-the Art Denoising

Ground truth  Input ($\sigma = 40$)  Out. $\text{PSNR}_\text{RTF}_{\text{ALL}}$  Output EPLL

Ground truth  Dust on Lens  “Perceptual Error”  “normal Error”

[Janscary, Nowozin, Rother, ECCV 12]
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 Image - local
What the computer sees

 Ambient Light
An unlikely 3D representation
(hard to see for a human)

True colors in 3D world

Direct Light
The most likely 3D representation
This is what humans see implicitly. Ideally the computer sees the same.

2D

3D

A

B

A

B

A

B
The importance of Prior knowledge

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

Which monster is bigger?
The importance of Prior knowledge

Which monster is bigger?

In the 2D Image

1 meter

2 meter

In the 3D world (true)
Two Explanations:

a) People are different height and room right shape
b) People are same height but room weirdly shaped
Human Vision can be fooled
Male or Female
How can we interpret visual data?

- 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 is efficiently is open challenged (see later)
Cue: Appearance (Colour, Texture) for object recognition

![Image of a cow in a field]
Cue: Outlines (shape) for object recognition
Guess the Object

- Colour
- Texture
- Shape

[from John Winn ICML 2008]
Guess the object

- Colour
- Texture
- Shape

[from John Winn ICML 2008]
Cue: Context for object recognition
Cue: Context for object recognition
Cue: stereo vision (2 frames) for geometry estimation

Ground truth

Algorithmic output
Cue: Multiple Frames for geometry estimation
Cue: Convergence for geometry estimation

Lines with same vanishing point may also be parallel in 3D
Cue: Shading & shadows for geometry and Light estimation
Texture gradient for geometry estimation
The “Scene Parsing” challenge --- a “grand challenge” of computer vision

(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 in reach

To simplify the problem:

1) Richer Input:
   - Modern sensing technology
   - Moving images
   - User involvement

2) Rich Data to learn from:
   - use the web
   - crowdsourcing to get labels
     (online games, mechanical turk)
   - Powerful graphics engines

3) For many practical applications:
   We do not have to infer the full probabilistic script
Kinect has simplified (revolutionized) computer vision

[1zadi et al. ´11]
Animate the world

[Chen et al. UIST ‘12]
New hardware design ...

**Digits: Freehand 3D Interactions Anywhere Using a Wrist-Worn Gloveless Sensor**

David Kim\(^1,2\), Otmar Hilliges\(^1\), Shahram Izadi\(^1\), Alex Butler\(^1\), Jiawen Chen\(^1\), Iason Oikonomidis\(^1,3\), Patrick Olivier\(^2\)

Voice-over by
Emily Whiting
ETH Zürich

\(^1\)Microsoft Research, UK
\(^2\)Culture Lab, Newcastle University, UK
\(^3\)FORTH, University of Crete, Greece

{b-davidk, otmarh, shahrami, dab, jiawen}@microsoft.com, oikonom@ics.forth.gr, p.l.olivier@ncl.ac.uk
Kinect Body Pose estimation and tracking
Kinect Body Pose estimation and tracking
behind the scene ...

Graphics simulation

Synthetic (graphics)  Real (hand-labelled)
Body tracking and Gesture Recognition has many applications

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

StartUp 2012: Try Fashion online
Real-time pedestrian detection
Real-time Face recognition

e.g. Canon powershot
General Object recognition & segmentation

Good results ...

[TextonBoost; Shotton et al, '06]
General Object recognition & segmentation

Failure cases...

[TextonBoost; Shotton et al, ‘06]
Image Search
Start-Up Company: Like.com

Vestal

Vestal Alpha Bravo Canvas Watch Blue/Silver/Black

$76 online

Write a review  Add to Shortlist  Browse Watches »

Go commando with the no-nonsense, military-inspired Vestal Alpha Bravo Canvas Watch. The mineral crystal glass face and stainless-steel housing can take a beating when you're on a mission in hostile territory, and the durable canvas strap provides a comfortable fit for all-day... more »

Visually Similar Items

Vestal
$75.99

GioiaPura
$139.71

Diesel
$98.57

Marc Ecko
$81.25

Vestal
$76.46

Govnil
$668.13

Sponsored

$76.46
Free shipping
Backcountry.com

Shop »

$76.46
Free shipping
DogFunk.com

Shop »

$81.00
Extreme Supply
4.5 stars (186)

Shop »

See all stores from $70
Interactive Image manipulation

[Agrawal et al ’04]
Interactive Image manipulation

(a) Original photograph  (b) User scribbles  (c) Reflectance  (d) Illumination  (e) Re-texturing
Image manipulation - stitching
Image manipulation - stitching
Video manipulation
Image de-convolution

Input

Output

Output – kernel

[Schmidt, Rother, Nowozin, Jancsary, Roth 2013] Best Student Paper award
Image de-convolution (other domains)

input

output

Raw

Deconvolved
Video Editing

[Rav-Acha et al. ‘08]
Automatic Video Summary (StartUp: Magisto)
Automatic Photo Summary - Commercial

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

Pirates of the Caribbean, Industrial Light and Magic
Robotics

Robocup

Nasa Mars exploration
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.
Interactive Segmentation
Model versus Algorithm

Example: Interactive Segmentation

Given \( z \); derive binary \( x \):

**Model:** Energy function \( E(x) \) (implicitly models a statistical model \( P(x|z) \))

**Algorithm** to minimization: \( x^* = argmin_x E(x) \)
Goal: formulate $E(x)$ such that

Optimal solution $x^* = \arg\min_x E(x)$
How does the energy looks like?

Energy function (sum of terms $\theta$):

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

Unary terms

Pairwise terms
How does the energy looks like?

Visualization:
Undirected graphical models

\[ \theta_{ij}(x_i, x_j) \quad \text{“pairwise terms”} \]

\[ \theta_i(x_i) \quad \text{“unary terms”} \]
Unary term

User labelled pixels

Gaussian Mixture Model Fit
Unary term

\( \theta_i(x_i = 0) \)
Dark means likely background

\( \theta_i(x_i = 1) \)
Dark means likely foreground

Optimum with unary terms only
Pairwise term

$$\theta_{ij}(x_i, x_j) = |x_i - x_j| \text{ “Ising Prior”}$$

When is $$\theta_{ij}(x_i, x_j)$$ small, i.e. likely configuration?

This models the assumption that the object is spatially coherent.
Next step could be: model shapes of starfishes
Energy minimization (optimization)

\[ E(x) = \sum_i \theta_i(x_i) + \omega \sum_{i,j} |x_i - x_j| \]
The key Questions

• What type of **modelling language** should be chosen: undirected or directed discrete Graphical models, Continuous-Domain models

![Diagram of Graphical Models](image)

• How does the exact **model** look like:
  • What is the structure
  • How do the terms look like

• Can we **learn** the Model from Data:
  • Learn structure
  • Learn potential functions

• How do we **optimize** the model (perform inference):
  • fast, approximate
  • Exactly solvable?
  • NP-hard?
Another Example: Model versus Algorithm

Input: Image sequence

Output: New view

Model: Minimize a binary 4-connected pair-wise graph
(choose a colour-mode at each pixel)

[Data courtesy from Oliver Woodford]

[Fitzgibbon et al. ‘03]
Another Example: Model versus Algorithm

Why is the result not perfect? Model or Optimization
Why is computer vision interesting (to you)?

- It is a challenging problem that is far from being solved

- It combines insights and tools from many fields and disciplines:
  - Mathematics and statistics
  - Cognition and perception
  - Engineering (signal processing)
  - And of course, computer science
Why is computer vision interesting (to you)?

• Allows you to apply theoretical skills
  ... that you may otherwise only use rarely.

• Quite rewarding:
  • Often visually intuitive and encouraging results.

• It is a growing field:
  • Cameras are becoming more and more popular
  • There are a lot of companies (big, small, startups) working in vision
  • Conferences are growing rapidly.
Relationship to other fields – my personal view

- Biology
- Robotics
- Human-Computer Interaction
- Medicine
- (many more)

Applications

- Machine Learning
- Optimisation
- Computer Graphics
- Sensing hardware

Computer Vision
Reading for next class

This lecture: Chapter 1 (in particular: 1.1)

Next lecture:
• Chapter 3 (in particular: 3.2, 3.3) - Basics of Digital Image Processing
• Chapter 4.2 and 4.3 - Edge and Line detection