Computer Vision I

Introduction

13.10.2014

TECHNISCHE UNIVERSITÄT DARMSTADT

Stefan Roth

visual inference

Computer Vision I

TECHNISCHE UNIVERSITÄT DARMSTADT

- Lecturer
  - Stefan Roth `<sroth AT cs.tu-darmstadt...>`
  - Office hours: Wednesdays, 13:45 – 14:45
- Teaching assistants
  - Stephan Richter + Jochen Gast
  - Office hours: TBA
- Course staff email
  - `<cv1staff AT gris.tu-da...>`
  - Please direct any questions here!

Course Material

- Course web page: http://goo.gl/5iGMoj
  - Pointers to readings
- Moodle ("HRZ Moodle")
  - Slides and homework assignments
  - To come
  - We will be using it for handing in homework assignments etc.
- Mailing list
  - *Vote*: ~~Mailing list~~ or Moodle? ✔

Forum

- There is a forum set up at the Fachschaft's website [http://www.fachschaft.informatik.tu-darmstadt.de/forum/viewforum.php?f=290]
  - Please use it to ask questions of public interest.
  - You are encouraged to discuss with each other.
  - *However:* Please do not share solutions or give strong hints about the solutions to the homework problems.

Course language

- will be *English*.
  - This applies to lectures, exercises, announcements, etc.

- Why?
  - Essentially all computer vision publications and books are written in English.
  - Knowing the original terms is crucial.

- If strongly preferred, you may contact the course staff in German.
  - English is encouraged though, because we may use your (anonymized) question to clarify points to the entire class.

Organization

- Class type: IV4

- Lecture
  - ~2 hours a week
  - Mondays, 13:30, S3|05, room 074
  - We will cover the foundational aspects of each topic.

- Exercise
  - ~2 hours a week (irregularly, remaining times: homework)
  - Mondays, after lecture, S3|05, room 074
  - We will cover some practical aspects, and discuss the homework assignments.

Style

- Lectures
  - I would like the lectures to be at least partly interactive.
  - *You are encouraged to ask questions!*

- Exercises
  - Mostly interactive.
  - You are encouraged to ask detailed questions!

Exam & Exercises

- Exam
  - Written exam at the end of the semester.
  - Can be taken in English or German.
  - Exam date will be set and announced soon.

- Exercises
  - Regular homework assignments; exercises will be graded.
  - Exercise points are part of your final grade (vorlesungsbleitende Prüfung)!
  - Details: TBA

- Grading
  - 2/3 exam, 1/3 exercise

Homework assignments

- Mostly programming assignments.
  - *MATLAB*, standard environment for scientific computing.
- Sometimes smaller pen and paper exercises

- The homework exercises are crucial for actually “digesting” the material from the lecture.
  - Hence your points count…

- Time commitment
  - Solving the homework problems will require a substantial time commitment.
  - Keep in mind that 6CP equal a workload of approximately 180h / semester.

Collaboration policy

- I do not tolerate plagiarism in any way!
- You may (and are encouraged to) discuss general class topics.
- But each handed in homework solution must be your own!
  - Any sources you used (other than provided by us) must be cited.
  - Details: TBA on the first homework assignment sheet.
- Questions? Problem cases?
  - *Talk to us.*

Readings

- Course book
  - *Computer Vision: Algorithms and Applications* by Richard Szeliski (Springer, 2011)
  - Good news: PDF available online - [http://szeliski.org/Book]

- Additional readings
  - Papers and tutorials
  - Will be available on the web or course page.

- Reading assignments
  - Please read the assigned material *in advance* of the next class

How does it fit into your course plan?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Elective
  - Part of Human Computer Systems (HCS) track.
- Related classes
  - Human Computer Systems: prerequisite, with exceptions
  - Computer Vision II (Roth, SS)
  - Maschinelles Lernen - Statistische Verfahren I (Peters, SS)
  - Maschinelles Lernen - Statistische Verfahren II (Roth, WS)
  - Bildverarbeitung (Sakas, SS)
- Theses and projects
  - Topics in computer vision and machine learning.

Preliminary Syllabus

- Introduction and Overview
- Image formation
- Appearance-based matching
- Global and local features
- Bags of visual words
- Object detection
- Motion estimation
- Single & two-view geometry
- Two-view geometry
- Stereo & segmentation

Subject to change!

Goal of today's lecture

- Get *intuitions* about
  - What is computer vision?
  - Why is it useful?
  - Why is it hard / interesting?
  - How might we go about it?

What is computer vision?

[from Steve Seitz]

What does it mean to “see”?

see [1] |siː|

verb (**sees**, **seeing**, **saw** |sɔː| ; past participle **seen** |siːn|) [with obj.]

1 perceive with the eyes; discern visually: *in the distance she could see the blue sea* | [no obj.] *: Andrew couldn't see out of his left eye* | figurative *: I can't* ***see into*** *the future*.

• [with clause] be or become aware of something from observation or from a written or other visual source: *I see from your appraisal report that you have asked for training*.

[Oxford English dictionary, slide from Michael Black]

What does it mean to "perceive"?

TECHNISCHE UNIVERSITÄT DARMSTADT

perceive |pəˈsiːv|
verb [with obj.]
1 become aware or conscious of (something); come to realize or understand: *his mouth fell open as he perceived the truth* | [with clause] *: he was quick to perceive that there was little future in such arguments*.
• become aware of (something) by the use of one of the senses, especially that of sight: *he perceived the faintest of flushes creeping up her neck*.
2 interpret or regard (someone or something) in a particular way: *if Guy does not* ***perceive*** *himself* ***as*** *disabled, nobody else should* | [with obj. and infinitive] *: some geographers perceive hydrology to be a separate field of scientific enquiry*.

[Oxford English dictionary, slide from Michael Black]

Computer Vision

- What is computer vision (or machine vision)?
  - *Developing computational models and algorithms to*
    - *… interpret digital images*
    - *… understand the visual world we live in.*
- Is that important?
- What can we (already) do with it?

[from Steve Seitz]

Face detection

e.g. Canon [powershot.com], etc.

Human pose estimation

Microsoft Kinect
[www.xbox.com/kinect]

Earth viewers

Google Street View
[www.google.com]

[www.mobile3dcity.com]

Photo browsing

Photosynth [photosynth.net]

Optical character recognition (OCR)

- Convert scanned documents to text

Digit recognition, AT&T labs
[www.research.att.com/~yann/]

License plate readers
[en.wikipedia.org/wiki/Automatic_number_plate_recognition]

[from Steve Seitz]

Traffic sign recognition

Opel

VDO

Special effects: Shape capture

The Matrix movies, ...

[from Steve Seitz]

Special effects: Motion capture

Pirates of the Carribean, Industrial Light and Magic

[from Steve Seitz]

Sports

3D soccer analysis
[www.kicker.de]

Vision in space

NASA's mars exploration rover Spirit
[en.wikipedia.org/wiki/Spirit_rover]

[from Steve Seitz]

Robotics

Robocup
[www.robocup.org]

Darmstadt
Dribblers
[FG SIM]

Medical imaging

3D imaging
MRI, CT

Image guided surgery

[from Steve Seitz]

Why is computer vision interesting (for 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 (for 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 increasingly commonplace.
  - There are a number of vision companies both here and abroad.
  - Even the big players are hiring computer vision experts.
  - Conferences are growing rapidly.

Computer Vision

- What is computer vision (or machine vision)?
  - *Developing computational models and algorithms to*
    - *… interpret digital images*
    - *… understand the visual world we live in.*
- How could we go about it?

[from Steve Seitz]

Computer Vision

- We need a formal model that describes our problem …
- … as well as an algorithm that realizes (i.e., implements) it.
  - Neither the model alone, nor the algorithm alone suffices (in the long run).
  - Both *mathematical* and *computational*.

- What *properties / cues* of the visual world can we exploit or measure?

- What *general (prior) knowledge* of the world (not necessarily visual) should we exploit?

[adapted from Michael Black]

Case Study: Art & Vision

[from Michael Black]

Case Study: Art & Vision

[from Michael Black]

Case Study: Art & Vision

[from Michael Black]

Case Study: Art & Vision

[from Michael Black]

Case Study: (Human) Visual Cues

- Stereo parallax

Case Study: (Human) Visual Cues

- Motion parallax

[from Michael Black]

Case Study: (Human) Visual Cues

- Shadows

[from Michael Black]

Case Study: (Human) Visual Cues

- Convergence

[from Kingdom et al. 2007]

Case Study: (Human) Visual Cues

- Convergence

[from Kingdom et al. 2007]

Case Study: (Human) Visual Cues

- Context

[from Antonio Torralba]

Case Study: (Human) Visual Cues

- Context

[from Antonio Torralba]

Case Study: (Human) Visual Cues

- Context

[from Antonio Torralba]

Is human vision the reference?

Sinha and Poggio, Nature, 1996

[from Steve Seitz]

Is human vision the reference?

(c) A.Kitaoka 2003

[from Steve Seitz]

Is human vision the reference?

[from Michael Black]

Measurement vs. Interpretation

Edward H. Adelson

Measurement vs. Interpretation

Measurement vs. Interpretation

Vision

- We as humans use many *different cues* to interpret what is in an image.
  - They are not always right.
- We use information on what we regard as being a *plausible and meaningful interpretation*.
  - The measurement alone does not suffice.
- This is necessary, because an image is a *2D projection* of a *complex* 3D world.
  - A lot of information about the world is lost when we take a picture.

Ambiguity of Data

- Our image data is not only too little to fully recover and understand the “state of the visible world”.

- It may even be of *poor quality*:
  - Low resolution
  - (Sensor) noise
  - Etc.

Courtesy Criterion Collection

From Michael Black

- Our image data is always *ambiguous*.

Computer Vision

- We want to devise models and algorithms to understand the visual world much like we do.

- This means
  - we have to use a lot of *different cues*...
  - we have to deal with *ambiguity*...
  - we have to exploit what is a *plausible and meaningful interpretation*...
  - ... in order to extract information about the 3D visual world from a small amount of data.

- CV is an inverse problem.

Approach

- How do we turn these cues around?
  - “inverse graphics”?

- We need
  - to understand the geometry and physics of the world (to some extent);
  - a mathematical model of the cues that we want to exploit;
  - a mathematical model of our prior knowledge of the world;
  - a computational model and an algorithm to infer the state of the world from cues and prior knowledge.

Cameras: Introduction

- Cameras are *everywhere* these days.
  - Not just your digital still and video cameras.
  - Essentially every new cell phone has a camera built in.
    - *More cameras than people?*
  - Many new computers have a camera built in.

From [Ponce & Forsyth]

- History
  - From *"camera obscura"* (dark room)
  - Basic principle known to Mozi (470-391 BC) and Aristoteles (384-322 BC)
  - Described by Ibn al-Haytham (Alhazen) in his "Book of optics" (around 1000 AD).

[www.grand-illusions.com]

Camera obscura

"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". [Da Vinci]

[from Michael Black]

Cameras: Introduction

- Idealized standard camera model
  - Pinhole camera

From [Ponce & Forsyth]

  - Very simple, but for many purposes sufficient.
  - Usually, we work with the upright *virtual image*.

Cameras: Introduction

- Idealized standard camera model
  - Pinhole camera

CCD or CMOS

From [Ponce & Forsyth]

  - Very simple, but for many purposes sufficient.
  - Usually, we work with the upright virtual image.

Aside & Reminder: What is vision about?

- How do we make sense of this array of numbers?

"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees."

[Max Wertheimer, 1923]

[from Michael Black]

Perspective cameras

- Perspective projection with a pinhole camera:

- World point: $P = (x, y, z)^{\mathrm{T}}$
- Image point: $P' = (x', y', z')^{\mathrm{T}}$
- Focal length: f'

Perspective cameras

- Perspective projection with a pinhole camera:

- Projected depth:

$$z' = f'$$

Note: We're assuming a virtual image!

- Projective projection:

$$\frac{x}{x'} = \frac{z}{f'} \quad \Rightarrow \quad \boxed{x' = f' \cdot \frac{x}{z}}$$

$$\frac{y}{y'} = \frac{z}{f'} \quad \Rightarrow \quad \boxed{y' = f' \cdot \frac{y}{z}}$$

Orthographic cameras

- Even simpler: Orthographic camera

$$\boxed{\begin{aligned} x' &= x \\ y' &= y \end{aligned}}$$

Coordinate Systems

- Objects live in “world coordinates”.
- Camera has “camera coordinate” system.
- Image is defined in “image coordinates”.

- Transformations
  - Extrinsic camera transformation takes world into camera coordinates.
  - Intrinsic camera transformation describes the image formation process.

Coordinate Systems

- Extrinsic camera transformation
  - Described as a linear transformation
  - Using homogeneous coordinates
    - Augment vector with a constant 1: $\mathbf{x} = (x_1, x_2, x_3, 1)^{\mathrm{T}}$
  - Combination of rotation and translation:

$$(\mathbf{R}, \mathbf{t}) \in \mathbb{R}^{4\times 4}$$

$$\mathbf{x}^C = (\mathbf{R}, \mathbf{t}) \cdot \mathbf{x}^W$$

Coordinate Systems

- Intrinsic camera transformation
  - Usually described as a linear transformation + perspective division
  - Pinhole camera model is a special case
  - Can deal with more general camera models as well

$$\mathbf{K} \in \mathbb{R}^{4\times 4}$$

$$\mathbf{x}^I = \frac{1}{z^C}\mathbf{K}\cdot\mathbf{x}^C$$

Images

- Images arriving at our CCD or CMOS sensor are *spatially discrete* with individual pixels.

- But is the visual world spatially discrete? No.
  - Light hits the sensor everywhere.
  - Spatially continuous intensity function:

$$I(x, y), \quad I : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$$

- The image sensor performs a "sampling" of this function.
  - Turns it into a discrete array of intensity values.

Images

- Idealized spatial sampling (1D analogy)

  - Details: Signal processing, Fourier theory

- For our purposes here, we will regard the image as spatially discrete.
  - Array of pixels

Readings for next week

- Introduction (Ch. 1)
  - What is computer vision? (Sec. 1.1)
  - For the interested: History of computer vision (Sec. 1.2)

- Image formation (Ch. 2)