Vision Algorithms for Mobile Robotics

Lecture 01
Introduction

Davide Scaramuzza
Today’s Class

• About me
• What is Computer Vision?
• Example of Vision Applications
• Specifics of this course
• Introduction to Visual Odometry
Who am I?

Current positions

- Professor of Robotics, Dep. of Informatics and Neuroinformatics (UZH & ETH)

Education

- PhD from ETH Zurich with Roland Siegwart
- Post-doc at the University of Pennsylvania with Vijay Kumar & Kostas Daniilidis

Highlights

- Coordinator of the European project sFly on visual navigation of micro drones
  - Which introduced the PX4 autopilot and visual navigation of drones

Spinoffs & Tech Transfer

- Zurich-Eye, enabling machines to see, now Facebook-Oculus Zurich
- Former strategic advisor of Dacuda, now Magic Leap Zurich
- Fotokite, aerial filming made simple, incubated in my lab
My Research Background

Computer Vision
- Visual Odometry and SLAM
- Sensor fusion
- Camera calibration

Autonomous Robot Navigation
- Self driving cars
- Micro Flying Robots
My lab

http://rpg.ifi.uzh.ch

Closed to bahnhof Oerlikon,
Andreasstrasse 15, 2nd floor
Our Research Areas

Visual-Inertial State Estimation
[IJCV’11, PAMI’13, RSS’15, TRO’16]

End-to-End Learning
[RAL’16-17]

Vision-based Navigation of Flying Robots
[AURO’12, RAM’14, JFR’15]

Event-based Vision
[-IROS’3, ICRA’14, RSS’15, PAMI’17]
Parrot: Autonomous Inspection of Bridges and Power Masts

Albris drone

Automated take off, self-check & calibration
Dacuda 3D (now Magic Leap Zurich)

- Fully immersive VR (running on iPhone)
- Powered by SVO

[Dacuda's 3D division] [Magic Leap logo]
Zurich-Eye (now Oculus Zurich)

Vision-based Localization and Mapping Solutions for Mobile Robots

Today’s Class

• What is Computer Vision?
• Example of Vision Applications
• Specifics of this course
• Overview of Visual Odometry
What is computer vision?

Automatic extraction of “meaningful” information from images and videos

Semantic information

Geometric information (this course)
Vision Demo?

Terminator 2

We are almost there!
Why study computer vision?

➢ Relieve humans of boring, easy tasks
➢ Enhance human abilities: human-computer interaction, visualization, augmented reality (AR)
➢ Perception for autonomous robots
➢ Organize and give access to visual content
Vision in humans

- **Vision** is our most powerful sense
- Retina is $\sim 1000 \text{mm}^2$. Contains 130 million **photoreceptors** (120 mil. rods and 10 mil. cones for color sampling)
- Provides **enormous** amount of information: data-rate of $\sim 3 \text{GBytes/s}$
  - Half of primate cerebral cortex is devoted to visual processing!
- To match the eye resolution we would need a **500 Megapixel** camera. But in practice the acuity of an eye is **8 Megapixels** within a **15-degree field of view** (around the fovea)!
What A Baby Can See Every Month For The First Year Of Its Life

“Your baby sees things best from 15 to 30 cm away. This is the perfect distance for gazing up into the eyes of mom or dad. Any farther than that, and the newborn sees mostly blurry shapes because they're nearsighted. At birth, a newborn's eyesight is between 20/200 and 20/400.”

Why is vision hard?

How do we go from an array of number to recognizing a fruit?

What we see

What a computer sees
Origins of computer vision


He is the inventor of ARPANET, the current Internet
Related disciplines

- Cognitive science
- Robotics
- Image processing
- Artificial intelligence
- Machine learning
- Cognitive science
- Graphics
- Computer vision
- Robotics
Computer Vision vs Computer Graphics

Images $\xrightarrow{\text{Computer Vision}}$ Model

Computer Graphics

Inverse problems: analysis and synthesis.
Today’s Class

• About me
• What is Computer Vision?
• Examples of Vision Applications
• Specifics of this course
• Image Formation
Optical character recognition (OCR)

Technology to convert scanned docs to text

Digit recognition, AT&T labs, using CNN, by Yann LeCun (1993), now head of Deep Learning at Facebook
http://yann.lecun.com/

License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition
Face detection

Now in all new digital cameras and smartphones

(NB. Paul Viola is now Vice President of Amazon Prime Air)
Object recognition (in mobile phones)

This is becoming real:
- Lincoln Microsoft Research
- Point & Find, Nokia
- SnapTell.com (Amazon)
- Google Goggles
Special effects: shape and motion capture
Sports

• Augmented Reality

The AC Liveline system is bringing augmented reality to sailing.

2013 America’s Cup
Medical imaging

3D imaging
MRI, CT

Image guided surgery
Grimson et al., MIT
3D Reconstruction by Multi-View Stereo

YouTube Video
3D Reconstruction: Multi-View Stereo

YouTube Video
Microsoft Photosynth

The Photosynth Technology Preview is a taste of the newest – and, we hope, most exciting – way to view photos on a computer. Our software takes a large collection of photos of a place or an object, analyzes them for similarities, and then displays the photos in a reconstructed three-dimensional space, showing you how each one relates to the next.

http://labs.live.com/photosynth/

Based on Photo Tourism technology developed by Noah Snavely, Steve Seitz, and Rick Szeliski
Pix4D

- EPFL startup – Now a company
Automotive safety

- **Mobileye**: Vision systems in high-end Tesla, BMW, GM, Volvo models. Bought by **Intel in 2017 for 15 billion USD**!
  - Pedestrian collision warning
  - Forward collision warning
  - Lane departure warning
  - Headway monitoring and warning
Vision-based interaction: Xbox Kinect
Lot of Computer Vision in Modern Smartphones

iPhone X

- Ambient light sensor
- Proximity sensor
- Flood illuminator
- Infrared camera
- Speaker
- Microphone
- Front camera
- Dot projector
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 (made by 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.
Vision-based Autonomous Drone Navigation

Works in GPS-denied Environments (EU project SFLY)

[Scaramuzza et al., Vision-Controlled Micro Flying Robots: from System Design to Autonomous Navigation and Mapping in GPS-denied Environments, IEEE RAM, September, 2014]
Dacuda’s mouse scanner

- World’s first mouse scanner, Distributed by LG, Logitech, etc.
- Dacuda was bought by Magic Leap in 2017 and is now Magic Leap Zurich (focusing on Augmented Reality)
Microsoft HoloLens
Google Tango
Augmented Reality with Google Tango and Apple ARKit
Current state of the art

• These were just few examples of current systems
  – Many of these are less than 5 years old

• Computer Vision is a very active field of research, and rapidly changing
  – Many new applications and phone apps in the next few years

• To learn more about vision applications and companies
  – David Lowe maintained an excellent overview of vision companies until 2015
    • http://www.cs.ubc.ca/spider/lowe/vision.html
Let’s have a 10 minute break with Google Tango and Microsoft Hololens Demos
Today’s Class

• About me
• What is Computer Vision?
• Example of Vision Applications
• Specifics of this course
• Overview of Visual Odometry
Organization of this Course

- **Lectures:**
  - 10:15 to 12:00 every week
  - Room: ETH LFW C5, Universitätstrasse 2

- **Exercises:**
  - 13:15 to 15:00: Starting from the 3rd week. Then almost every week.
  - Room: ETH HG E 1.1

- **Official course website:** [http://rpg.ifi.uzh.ch/teaching.html](http://rpg.ifi.uzh.ch/teaching.html)
  - Check it out for the PDFs of the lecture slides and updates
Learning Objectives

• **High-level goal:** learn to implement current visual odometry pipelines used in mobile robots (drones, cars, Mars rovers), and Virtual-reality (VR) and Augmented reality (AR) products: e.g., Google Tango, Microsoft HoloLens

• You will also learn **to implement the fundamental computer vision algorithms** used in mobile robotics, in particular: feature extraction, multiple view geometry, dense reconstruction, object tracking, image retrieval, visual-inertial fusion, event-based vision.

• In order to learn these competences, **participation in the exercise sessions is critical although not mandatory!**
## Course Schedule

For updates, slides, and additional material: [http://rpg.ifi.uzh.ch/teaching.html](http://rpg.ifi.uzh.ch/teaching.html)

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Description of the lecture/exercise</th>
<th>Lecturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.09.2017</td>
<td>10:15 - 12:00</td>
<td>01 – Introduction</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td>28.09.2017</td>
<td>10:15 - 12:00</td>
<td>02 - Image Formation 1: perspective projection and camera models</td>
<td>Guillermo Gallego</td>
</tr>
<tr>
<td>05.10.2017</td>
<td>10:15 - 12:00</td>
<td>03 - Image Formation 2: camera calibration algorithms</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 1: Augmented reality wireframe cube</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>12.10.2017</td>
<td>10:15 - 12:00</td>
<td>04 - Filtering &amp; Edge detection</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td>19.10.2017</td>
<td>10:15 - 12:00</td>
<td>05 - Point Feature Detectors 1: Harris detector</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 3: Harris detector + descriptor + matching</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>26.10.2017</td>
<td>10:15 - 12:00</td>
<td>06 - Point Feature Detectors 2: SIFT, BRIEF, BRISK</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td>02.11.2017</td>
<td>10:15 - 12:00</td>
<td>07 - Multiple-view geometry 1</td>
<td>Guillermo Gallego</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 4: Stereo vision: rectification, epipolar matching, disparity, triangulation</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>09.11.2017</td>
<td>10:15 - 12:00</td>
<td>08 - Multiple-view geometry 2</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 5: Eight-point algorithm and RANSAC</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>16.11.2017</td>
<td>10:15 - 12:00</td>
<td>09 - Multiple-view geometry 3</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td>23.11.2017</td>
<td>10:15 - 12:00</td>
<td>10 - Dense 3D Reconstruction (Multi-view Stereo)</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 7: Intermediate VO Integration</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>30.11.2017</td>
<td>10:15 - 12:00</td>
<td>11 - Optical Flow and Tracking (Lucas-Kanade)</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 8: Lucas-Kanade tracker</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>07.12.2017</td>
<td>10:15 - 12:00</td>
<td>12 – Place recognition</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 9: Recognition with Bag of Words</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>14.12.2017</td>
<td>10:15 - 12:00</td>
<td>13 – Visual inertial fusion</td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 10: Pose graph optimization and Bundle adjustment</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
<tr>
<td>21.12.2017</td>
<td>10:15 - 12:00</td>
<td>14 - Event based vision + <strong>lab visit and live demonstrations</strong></td>
<td>Davide Scaramuzza</td>
</tr>
<tr>
<td></td>
<td>13:15 – 15:00</td>
<td>Exercise 11: final VO integration</td>
<td>T. Cieslewski/H. Rebecq/A. Loquercio</td>
</tr>
</tbody>
</table>
Exercises

- Almost every week starting from the 3rd week (check out course schedule)
- Bring your own laptop
- Each exercise will consist of coding a building block of a visual odometry pipeline. At the end of the course there will be one additional exercise dedicated to assembling all the blocks together into a full pipeline.
- Have Matlab pre-installed!
  - ETH
    - Download: https://idesnx.ethz.ch/
  - UZH
    - Download: http://www.id.uzh.ch/dl/sw/angebote_4.html
    - Info on how to setup the license can be found there.
  - An introductory tutorial on Matlab can be found here:
  - Please install all the toolboxes included in the license.
Exercises

• **Learning Goal** of the exercises: Implement a full visual odometry pipeline (similar to that running on Mars rovers and on current AR/VR devices (but actually much better 😊)).

• **Each week** you will learn how to implement a **building block** of visual odometry. The building blocks are:

  - Image sequence
  - Feature detection
  - Feature matching (tracking)
  - Motion estimation
    - 2D-2D
    - 3D-3D
    - 3D-2D
  - Local optimization
Outcome of last year exercises
Recommended Textbook

- **Robotics, Vision and Control: Fundamental Algorithms**, by Peter Corke 2011. The PDF of the book can be freely downloaded (only with ETH VPN) from Springer or alternatively from Library Genesys


- Other books:
  - *An Invitation to 3D Vision*: Y. Ma, S. Soatto, J. Kosecka, S.S. Sastry
  - *Multiple view Geometry*: R. Hartley and A. Zisserman
Instructors

• Lecturer
  – Davide Scaramuzza: sdavide (at) ifi (dot) uzh (dot) ch
  – Receiving hours: Thursday afternoon (announce yourself by email)

• Exercises

Henri Rebecq
rebecq (at) ifi (dot) uzh (dot) ch

Titus Cieslewski
titus (at) ifi (dot) uzh (dot) ch

Antonio Loquercio
loquercio (at) ifi (dot) uzh (dot) ch
Prerequisites

- Linear algebra
- Matrix calculus
- No prior knowledge of computer vision and image processing required
Grading and Exam

- The final grade is based on the oral exam (30 minutes)
- In addition, strong class participation can offset negative performance at the oral exam.
- Optional mini project: you have the option (not mandatory) to do a mini project, which consists of implementing a working visual odometry algorithm in Matlab. If the algorithm runs properly producing a reasonable result, you will be rewarded with an up to 0.5 grade increase on the final grade. However, notice that the mini project can be quite time consuming! The deadline to hand the mini project is 07.01.2018. Group work (up to 4) possible.
Class Participation

• Class participation includes
  – showing up
  – being able to articulate key points from last lecture
Today’s Class

• About me
• What is Computer Vision?
• Example of Vision Applications
• Specifics of this course
• Overview of Visual Odometry
What is Visual Odometry (VO)?

VO is the process of incrementally estimating the pose of the vehicle by examining the changes that motion induces on the images of its onboard cameras.

**Input**
- Image sequence (or video stream) from one or more cameras attached to a moving vehicle

**Output**
- \( R_0, R_1, ..., R_i \)
- \( t_0, t_1, ..., t_i \)
- Camera trajectory (3D structure is a plus)
Why VO?

- Contrary to wheel odometry, VO is **not affected by wheel slippage** on uneven terrain or other adverse conditions.

- More accurate trajectory estimates compared to wheel odometry (**relative position error 0.1% – 2%**)

- VO can be used as a complement to
  - wheel encoders (wheel odometry)
  - GPS
  - inertial measurement units (IMUs)
  - laser odometry

- Crucial for flying, walking, and underwater robots
Assumptions

- **Sufficient illumination** in the environment
- **Dominance of static scene** over moving objects
- **Enough texture** to allow apparent motion to be extracted
- **Sufficient scene overlap** between consecutive frames

Is any of these scenes good for VO? Why?
A Brief history of VO

- **1980**: First known VO real-time implementation on a robot by **Hans Moraveck** PhD thesis *(NASA/JPL)* for Mars rovers using one sliding camera (*sliding stereo*).
A Brief history of VO

- **1980**: First known VO real-time implementation on a robot by **Hans Moraveck** PhD thesis (**NASA/JPL**) for Mars rovers using one sliding camera (**sliding stereo**).

- **1980 to 2000**: The VO research was dominated by **NASA/JPL** in preparation of the 2004 mission to Mars.

- **2004**: VO was used on a robot on another planet: Mars rovers Spirit and Opportunity (see seminal paper from **NASA/JPL, 2007**).

- **2004**: VO was revived in the academic environment by **David Nister**’s «Visual Odometry» paper. The term VO became popular.
More about history and tutorials


VO vs VSLAM vs SFM
Structure from Motion (SFM)

SFM is more general than VO and tackles the problem of 3D reconstruction and 6DOF pose estimation from unordered image sets.

Reconstruction from 3 million images from Flickr.com
Cluster of 250 computers, 24 hours of computation!
Paper: “Building Rome in a Day”, ICCV’09
VO vs SFM

- VO is a *particular case* of SFM

- VO focuses on estimating the 3D motion of the camera *sequentially* (as a new frame arrives) and in *real time*.

- Terminology: sometimes SFM is used as a synonym of VO
VO vs. Visual SLAM

- **Visual Odometry**
  - Focus on incremental estimation/local consistency

- **Visual SLAM**: Simultaneous Localization And Mapping
  - Focus on globally consistent estimation
  - Visual SLAM = visual odometry + loop detection + graph optimization

- The choice between VO and V-SLAM depends on the tradeoff between performance and consistency, and simplicity of implementation.

- VO trades off consistency for real-time performance, without the need to keep track of all the previous history of the camera.

Image courtesy from [Clemente et al., RSS’07]
VO Working Principle

1. Compute the relative motion $T_k$ from images $I_{k-1}$ to image $I_k$

$$T_k = \begin{bmatrix} R_{k,k-1} & t_{k,k-1} \\ 0 & 1 \end{bmatrix}$$

2. Concatenate them to recover the full trajectory

$$C_n = C_{n-1}T_n$$

3. An optimization over the last $m$ poses can be done to refine locally the trajectory (Pose-Graph or Bundle Adjustment)
How do we estimate the relative motion $T_k$?

\[ T_k = \arg \min_T \int \int_{\tilde{\mathcal{R}}} \rho \left[ I_k \left( \pi \left( T \cdot \pi^{-1}(u, d_u) \right) \right) - I_{k-1}(u) \right] d\mathbf{u} \]
Direct Image Alignment

It minimizes the per-pixel intensity difference [1]

\[ T_{k,k-1} = \arg \min_{T} \sum_{i} \| I_k(u'_i) - I_{k-1}(u_i) \|_{\sigma}^2 \]

Dense

DTAM [Newcombe et al. ‘11]
300’000+ pixels

Semi-Dense

LSD [Engel et al. 2014]
~10’000 pixels

Sparse

SVO [Forster et al. 2014]
100-200 features x 4x4 patch
~ 2,000 pixels

Direct Image Alignment

It minimizes the per-pixel intensity difference [1]

\[ T_{k,k-1} = \arg \min_T \sum_i \| I_k(u'_i) - I_{k-1}(u_i) \|^2_\sigma \]

Dense

Semi-Dense

Sparse

Live incremental reconstruction of a large scene

Texture mapped model

Inverse depth solution

Camera image

LSD-SLAM builds a pose-graph of keyframes and associated semi-dense depth maps

SVO with a single camera on Euroc dataset

DTAM [Newcombe ‘11] REMODE [Pizzoli’14]
300’000+ pixels

LSD-SLAM [Engel’14]
~10,000 pixels

SVO [Forster’14]

100-200 x 4x4 patches \( \cong 2,000 \) pixels

VO Flow Chart

VO computes the camera path incrementally (pose after pose)

Image sequence

Front-end

Feature detection

Feature matching (tracking)

Motion estimation

2D-2D | 3D-3D | 3D-2D

Local optimization

Back-end
VO Flow Chart

VO computes the camera path incrementally (pose after pose)

1. Image sequence
2. Feature detection
3. Feature matching (tracking)
4. Motion estimation
   - 2D-2D
   - 3D-3D
   - 3D-2D
5. Local optimization

Example features tracks
VO Flow Chart

VO computes the camera path incrementally (pose after pose)

Image sequence

Feature detection

Feature matching (tracking)

Motion estimation

2D-2D   3D-3D   3D-2D

Local optimization
VO Flow Chart

VO computes the camera path incrementally (pose after pose)

Front-end

Image sequence

Feature detection

Feature matching (tracking)

Motion estimation

2D-2D  3D-3D  3D-2D

Local optimization (back-end)

Back-end

\[ C_0 \rightarrow C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow \ldots \rightarrow C_{n-1} \rightarrow C_n \]

\[ k \]

\( m \) – poses windowed bundle adjustment
Course Topics

• Principles of image formation
• Image Filtering
• Feature detection and matching
• Multi-view geometry
• Visual place recognition
• Event-based Vision
• Dense reconstruction
• Visual inertial fusion
Course Topics

• Principles of image formation
  – Perspective projection
  – Camera calibration
Course Topics

- Feature detection and matching
Course Topics

• Multi-view geometry and 3D reconstruction
Course Topics

• Multi-view geometry and 3D reconstruction

San Marco square, Venice
14,079 images, 4,515,157 points
Course Topics

• Dense reconstruction
Course Topics

• Dense reconstruction

M. Pizzoli, C. Forster, D. Scaramuzza, REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time, ICRA’14
Course Topics

- Visual place recognition
Course Topics

• Visual place recognition

<table>
<thead>
<tr>
<th>Query image</th>
<th>Most similar places from a database of millions of images</th>
</tr>
</thead>
</table>

[Images of similar places]
Course Topics

- Visual-inertial fusion
Course Topics

• Event-based vision
Course Topics

• Visual odometry