Robotic 3D Vision

Lecture 1: Introduction

WS 2020/21

Dr. Niclas Zeller
Artisense GmbH

Organization

Lecturer:

- Dr. Niclas Zeller (niclas.zeller@in.tum.de)

Teaching Assistant:

- Patrick Wenzel

Course Webpage:

- https://vision.in.tum.de/teaching/ws2020/robot_vis
- Slides will be made available on the webpage

Organization

- Structure: Lecture (3h) + Exercise (1h)
  - 5 ECTS credits

- Study programme: **M.Sc. Informatics**

- Place & Time
  - Lecture: Wed 14:00 – 15:30 online
  - Lecture/Exercises: Fri 14:00 – 15:30 online

- Exam
  - Written Exam
  - Date: TBD

Ackowledgement

- This lecture was initially developed by Dr. Jörg Stückler, who held this lecture in winter term 2017/18 at TUM.
- Therefore, he deserves the acknowledgement for defining the content of this lecture as well as creating these slides, which where only partially modified.

Exercises and Demos

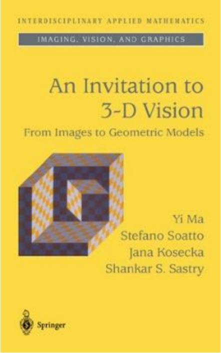
- Exercises
  - Typically 1 exercise sheet every 2 weeks (theoretical and Matlab-based assignments)
  - Hands-on experience with the algorithms from the lecture
  - Exercise sheet will be handed out about one week before the class
  - Solutions will be provided at the end of the term
  - Exercises are not mandatory to take the exam
  - First exercise class: Friday Nov. 13, 2020 14.00 – 15:30

Course Requirements

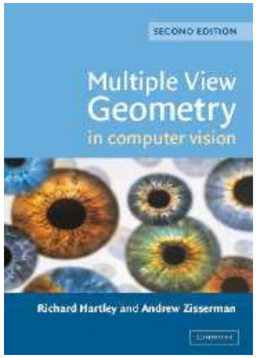
- We will build on basics from previous lectures
  - Computer Vision II: Multiple View Geometry
    https://vision.in.tum.de/teaching/ss2020/mvg2020

Textbooks

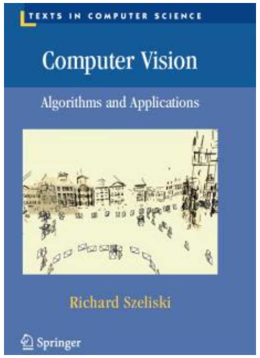
- No dedicated textbook for the class
  - Related topics can be found in



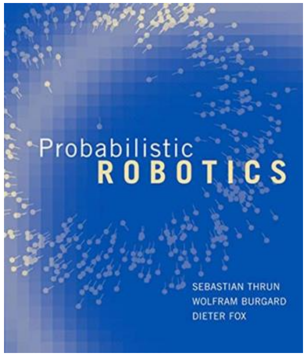
An Invitation to 3D Vision, Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, Springer, 2004



Multiple View Geometry in Computer Vision, R. Hartley and A. Zisserman, Cambridge University Press, 2004



Computer Vision – Algorithms and Applications, R. Szeliski, Springer, 2006



Probabilistic Robotics, S. Thrun, Wolfram Burgard, Dieter Fox, MIT Press, 2006

Questions about the Lecture

- Please ask questions at any time during the lecture
- If you have questions about the topic feel free to contact me via email

Robots in Complex Environments

Image credit: Amazon

Image credit: DHL

Image credit: Waymo

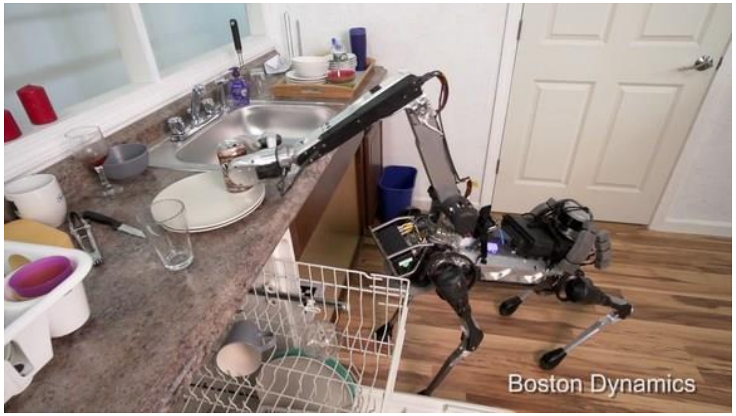


Image credit: Boston Dynamics

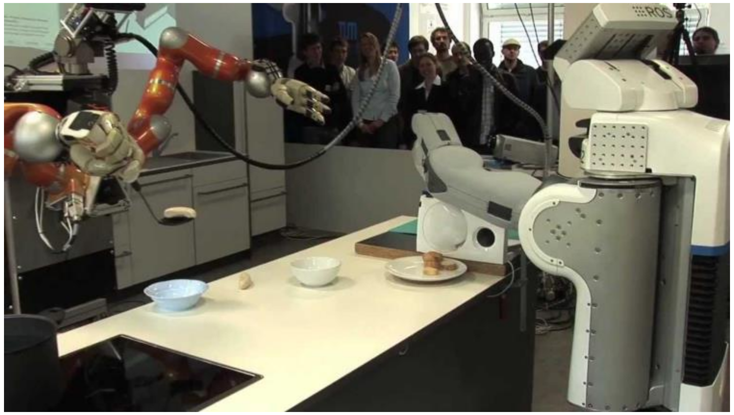


Image credit: IAS TUM / UBremen

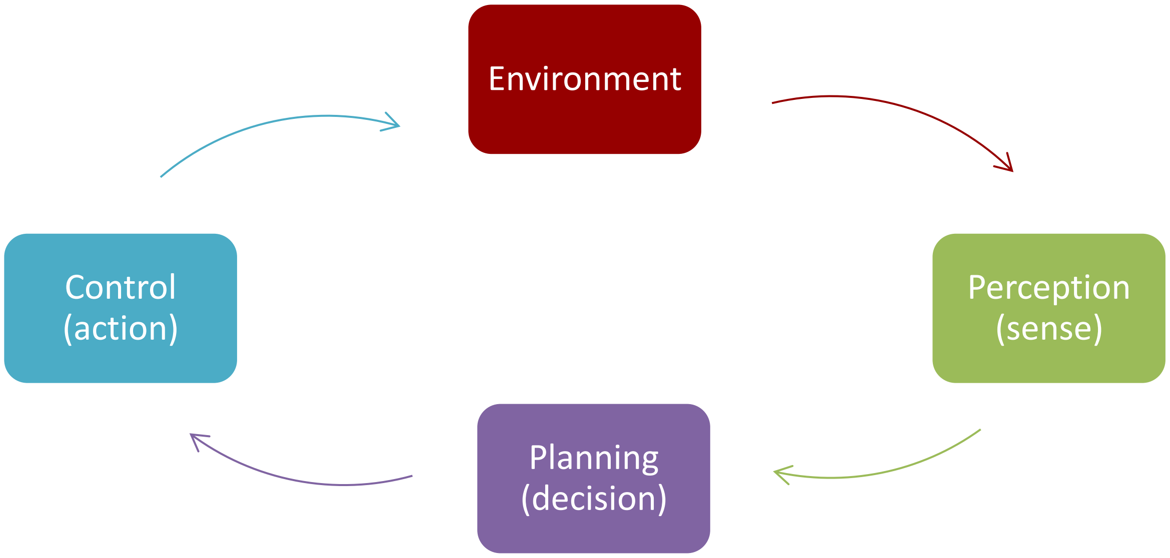
Robotic Perception

Video credit: KUKA Robots & Automation

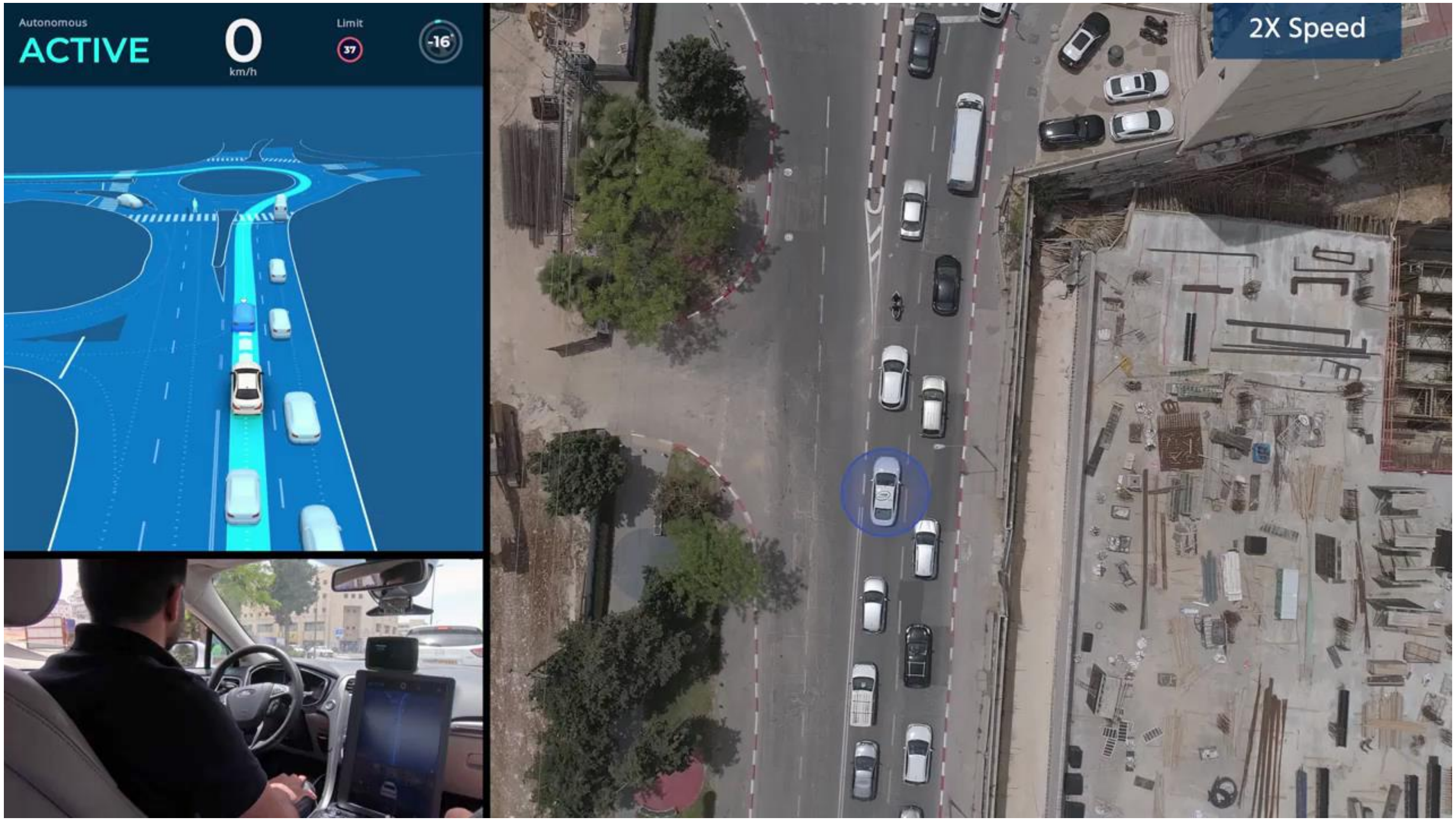
https://www.youtube.com/watch?v=tIIJME8-au8

Robotic Perception

- We want to build robots which are able to accomplish tasks simiar as or even better then humas
- Generally solved in a modular fashion
  - Holds for any kind of autonomous application
    - Robotics, autonomous driving, etc.



Robotic Perception



Video credit: Mobileye

https://www.youtube.com/watch?v=OtlJNomXb2s

Robotic Perception

(Stückler, Schwarz, Behnke, Frontiers 2016)

What We Will Cover Today

- Why Vision for Robotic Perception?
- What is Robotic 3D Vision?
- Terminology of
  - Visual Odometry
  - Visual-Inertial Odometry
  - Visual Simultaneous Localization and Mapping
  - Map Representations
  - Dense vs. Sparse Reconstruction
  - Indirect and Direct Methods
  - Visual 3D Object Detection and Tracking

Sensors for Robotic Perception

Vision

+ low power consumption
+ dense 2D projection
+ appearance
+ high frame-rate
- indirect distance

Laser

+ accurate distance
- power consumption
- sparse
- low frame-rate
- interference

Inertial (IMU)

+ linear acceleration
+ gravity
+ rotational velocity
+ high frame-rate
- noise & bias
- only derivertives (no abs. position)
- no sensing of environment

Proprioceptive

+ forward kinematics

(+ forward dynamics)

- no sensing of environment

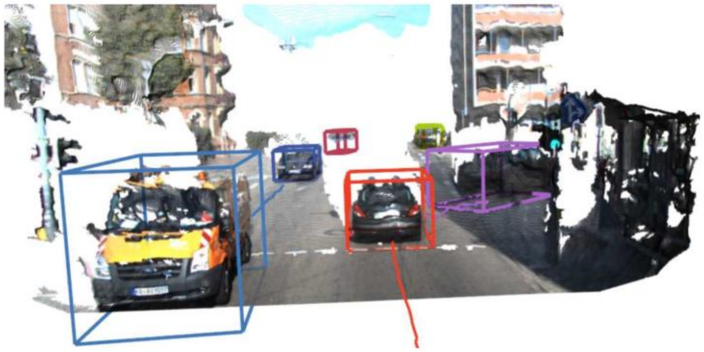
Tactile

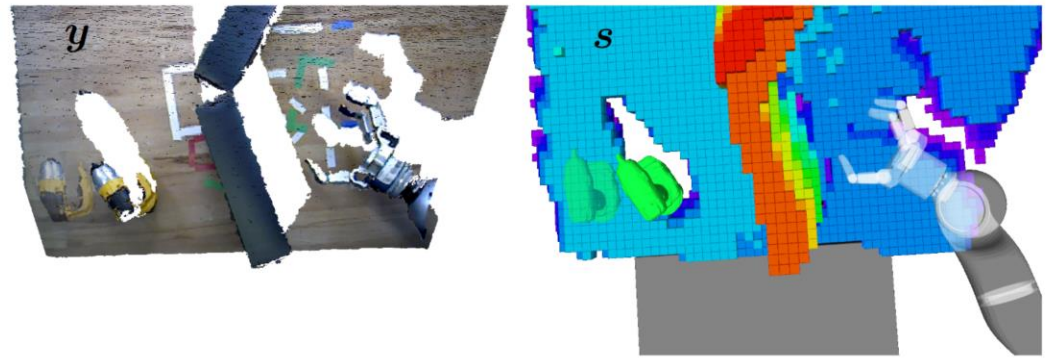
+ contact with environment

RGB-D

+ depth image
- power consumption
- interferecne

Robotic 3D Vision



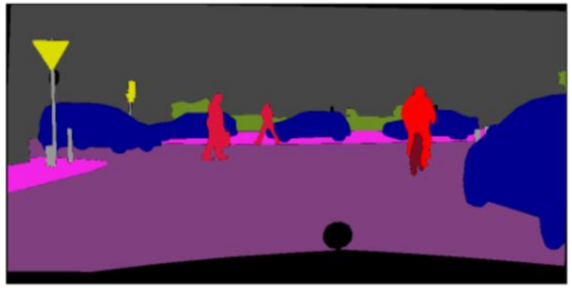


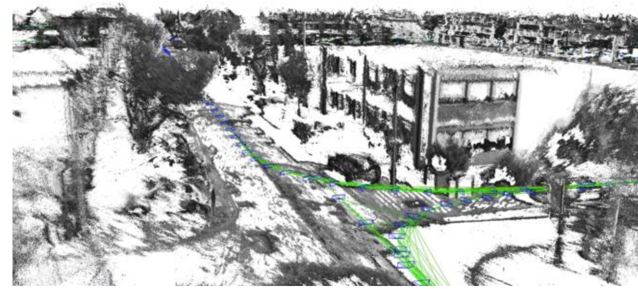
- Robots require 3D scene understanding
  - Where is the robot in the environment?
  - What is the shape (structure) of the environment?
  - Where are task-relevant objects?

- 3D Vision: 3D scene understanding from camera images

Images from: (Osep et al., ICRA 2016), (Kappler et al., arXiv 2017)

Why Vision?





Vision provides robots with rich information about the world

- Dense 2D measurements of the 3D world, in contrast to, for example, laser scanners or ultrasonic range scanners
- RGB/grayscale measurements of the appearance of objects available to detect and recognize objects
- Range/depth (third dimension) assessable by stereo
- Lightweight and low power consumption (passive cameras)

Images from: (Pohlen et al., CVPR 2017), (Engel, Stückler, Cremers, IROS 2015)

Types of Cameras

Monocular camera

- Structure from motion (chicken-and-egg problem)
- Scale ambiguity

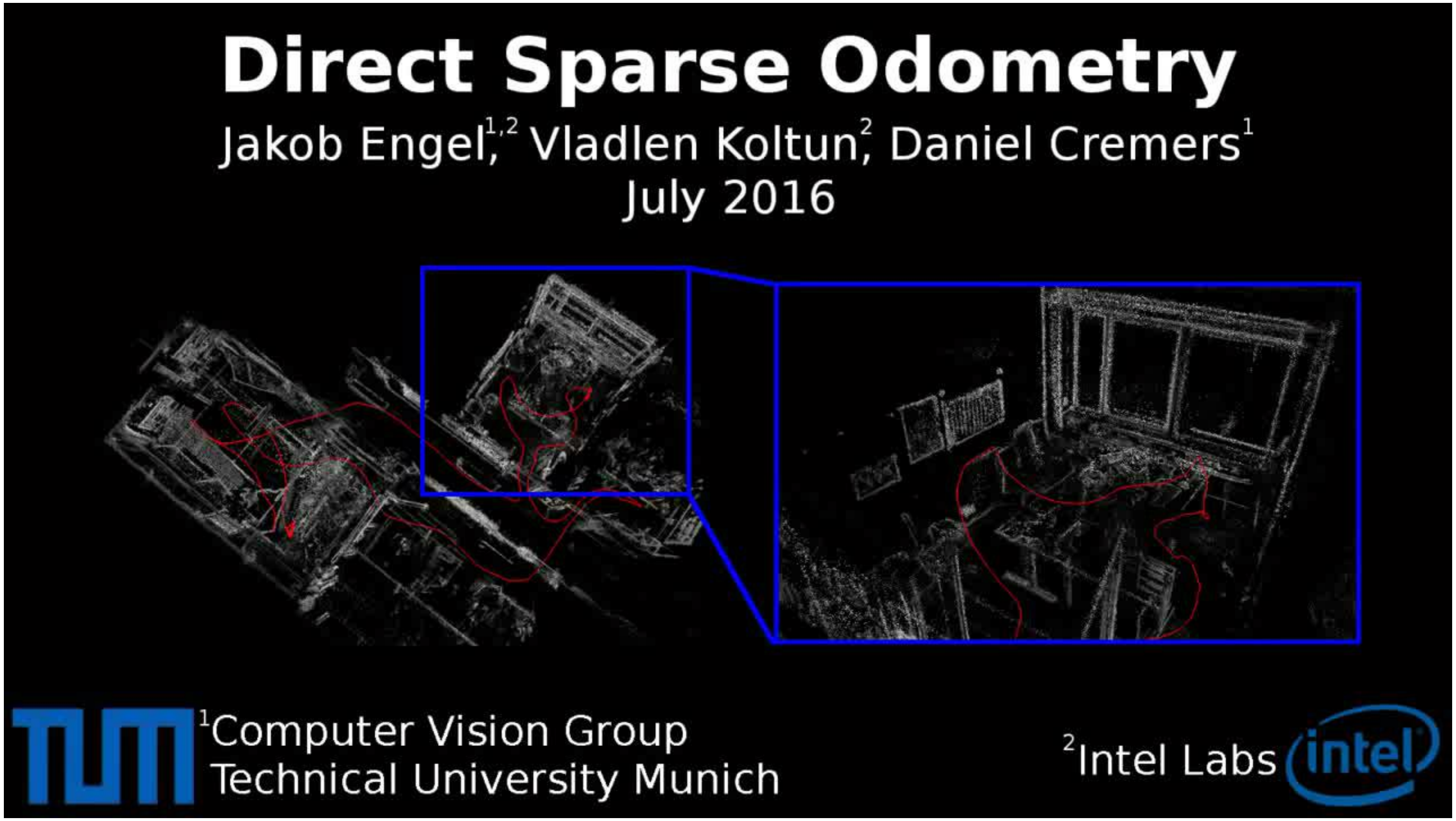
Stereo camera

- Depth from stereo in fixed configuration
- Scale observable
- Fixed baseline

RGB-D camera

- Directly measures per-pixel depth
- Active sensing
- Structured light or time-of-flight

Visual Odometry



(Engel, Koltun, Cremers, T-PAMI 2018)

How does the robot move?

https://www.youtube.com/watch?v=C6-xwSOOdqQ

What is Visual Odometry?

Visual odometry (VO)…

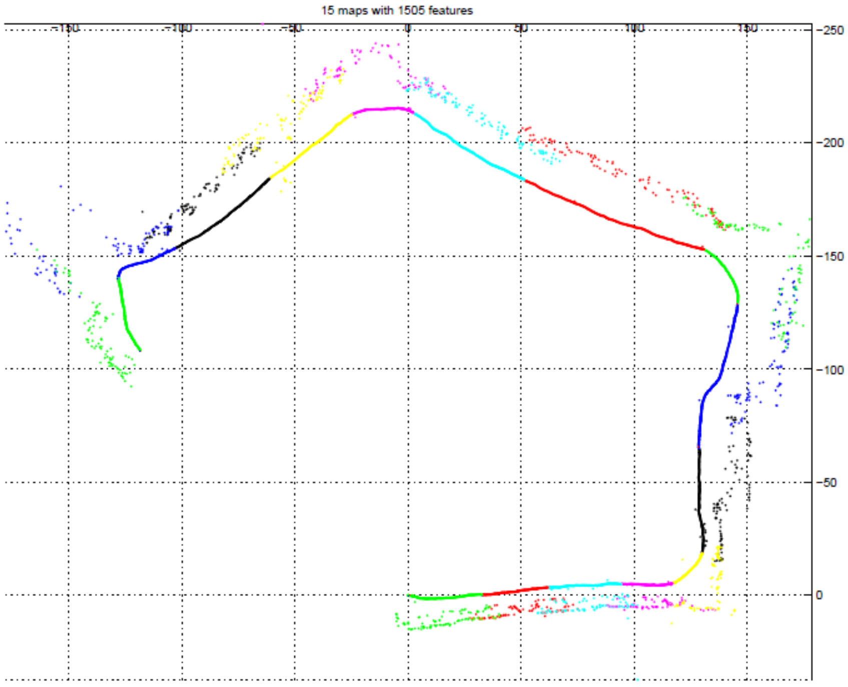
- … is a variant of tracking
  - Track the current pose, i.e. position and orientation, of the camera with respect to the environment from its images
  - Only considers a limited set of recent images for real-time constraints

- … involves a data association problem
  - Motion is estimated from corresponding interest points or pixels in images, or by correspondences towards a local 3D reconstruction

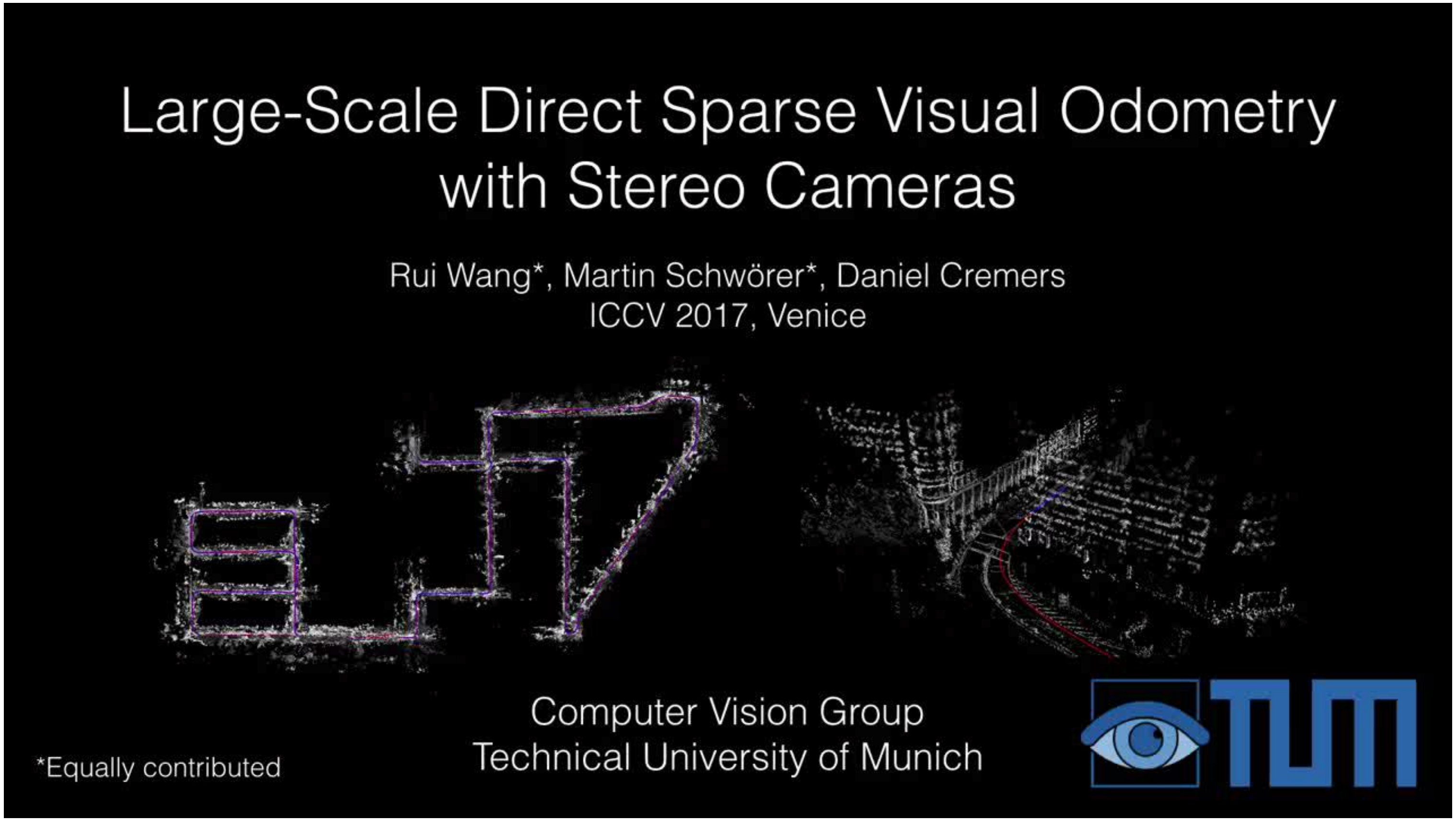
What is Visual Odometry?

Visual odometry (VO)…

- … is prone to drift due to its local view
- … is primarily concerned with estimating camera motion
  - 3D reconstruction often a "side product". If estimated, it is only locally consistent
- Monocular VO is not able to observe the absolute scale
  - Scale drift



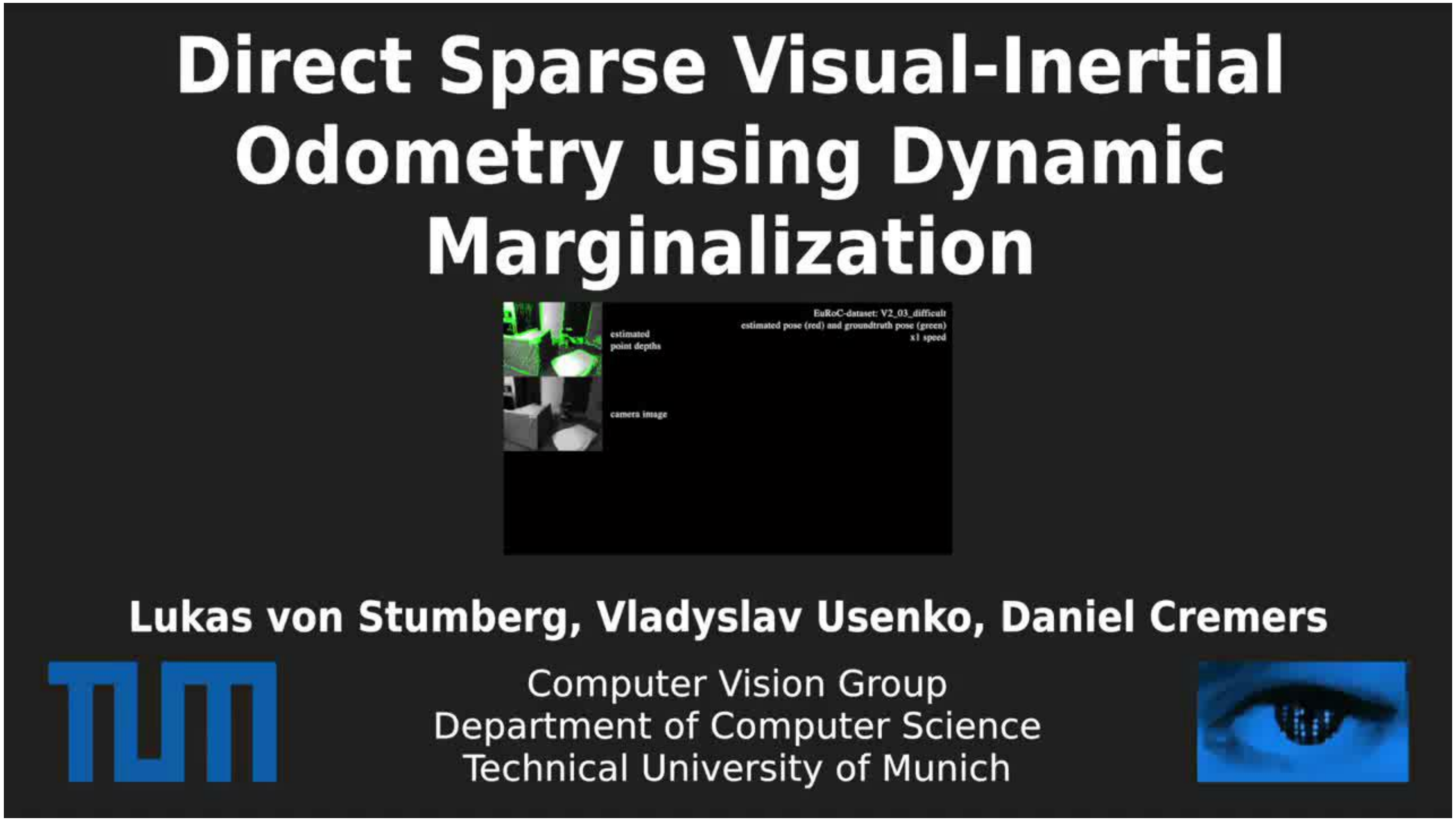
Stereo Visual Odometry



(Wang, Schwörer, Cremers, ICCV 2017)

https://www.youtube.com/watch?v=A53vJO8eygw

Visual-Inertial Odometry



(von Stumberg, Usenko, Cremers, ICRA 2018)

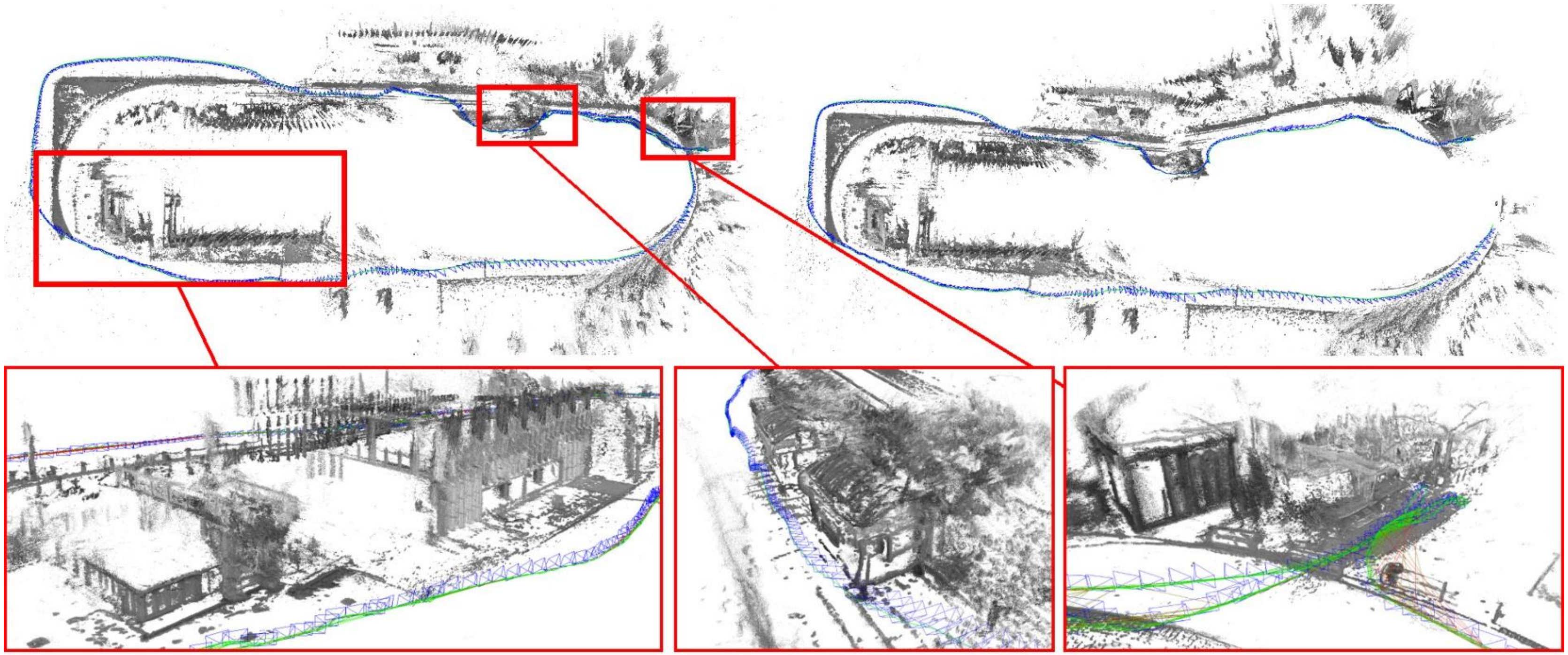
https://www.youtube.com/watch?v=GoqnXDS7jbA

What is Visual-Inertial Odometry?

Visual-inertial odometry (VIO)…

- … complements visual odometry with inertial measurements
  - Visual measurements provide up to 6-DoF relative motion using the environment as reference
  - Inertial sensors measure 3D linear accelerations and angular velocities, typically at much higher frame-rate than images
  - Gravity is also included in the acceleration measurements serving as an absolute external reference
  - Pure integration of gravity-compensated linear accelerations and angular velocities drifts
  - Vision helps to reduce integration drift, estimate sensor biases, discern gravity from motion-induced accelerations
  - Inertial measurements help to compensate degenerate cases of pure visual tracking (textureless areas, fast motion, etc.)

Simultaneous Localization and Mapping



(Engel, Schöps, Cremers, ECCV 2014)

Where is the robot and what is the 3D structure of the environment?

Simultaneous Localization and Mapping

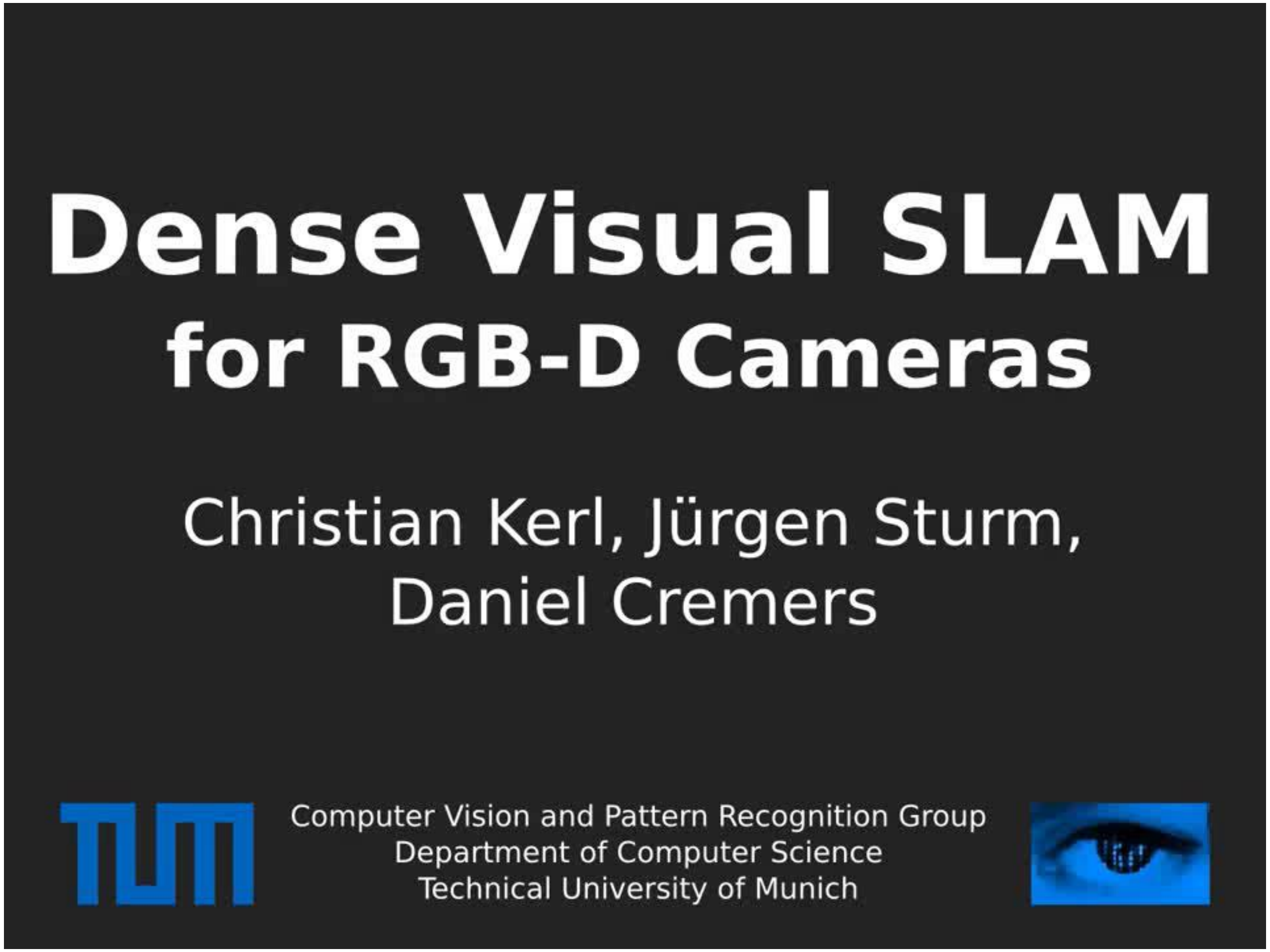
(Engel, Schöps, Cremers, ECCV 2014)

https://www.youtube.com/watch?time_continue=22&v=aBVXfqumTXc&feature=emb_logo

What is Visual SLAM?

- Visual simultaneous localization and mapping (VSLAM)…
  - Tracks the pose of the camera in a map, and simultaneously
  - Estimates the parameters of the environment map (f.e. reconstruct the 3D positions of interest points in a common coordinate frame)
- Loop-closure: Revisiting a place allows for drift compensation
  - How to detect a loop closure?
- Global and local optimization methods
  - Global: bundle adjustment, pose-graph optimization, etc.
  - Local: incremental tracking-and-mapping approaches, visual odometry with local maps. Often designed for real-time.
  - Hybrids: Real-time local SLAM + global optimization in a slower parallel process (f.e. LSD-SLAM)

Visual SLAM with RGB-D Cameras



(Kerl, Sturm, Cremers, IROS 2013)

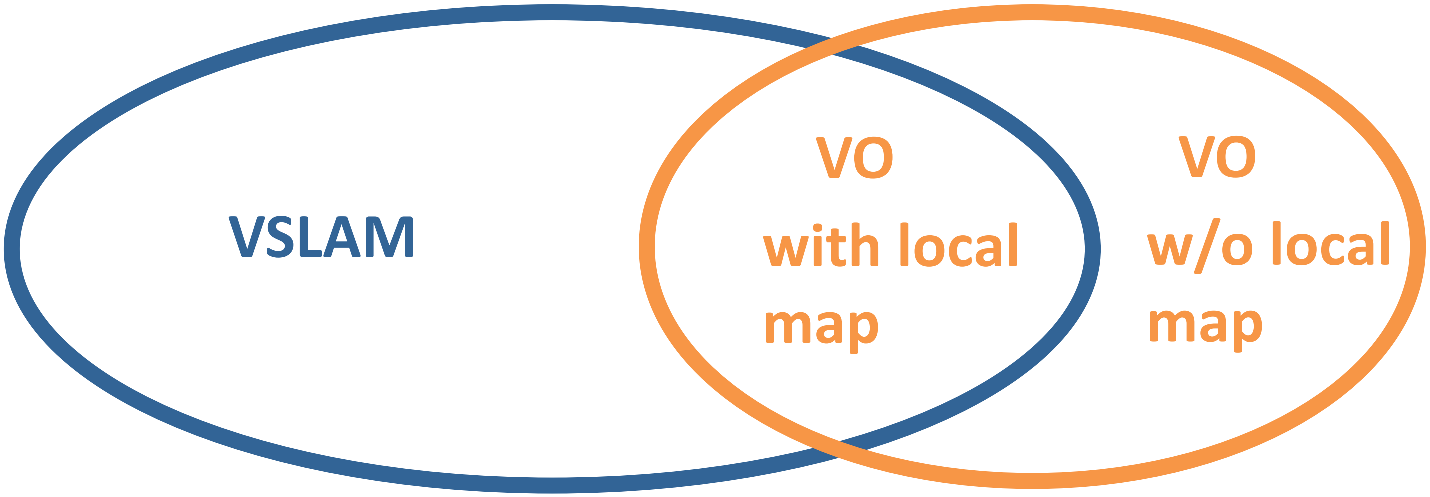
https://www.youtube.com/watch?v=jNbYcw_dmcQ

Visual SLAM using Bundle Adjustment

(Mur-Artal, Tardos, T-RO 2017)

https://www.youtube.com/watch?v=ufvPS5wJAx0

VO vs. VSLAM



Structure from Motion

- Structure from Motion (SfM) denotes the joint estimation of
  - Structure, i.e. 3D reconstruction, and
  - Motion, i.e. 6-DoF camera poses,

  from a collection (i.e. unordered set) of images

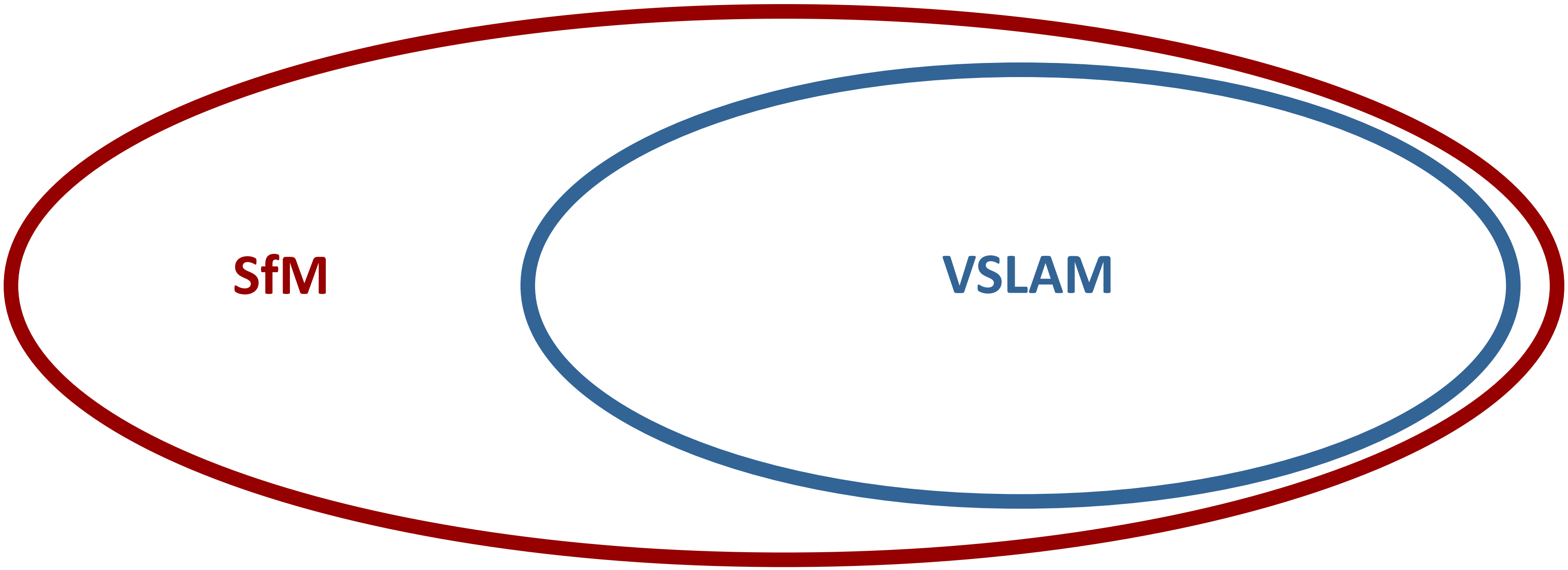
- Typical approach: keypoint matching and bundle adjustment

- In general no interest in real-time processing

Structure from Motion

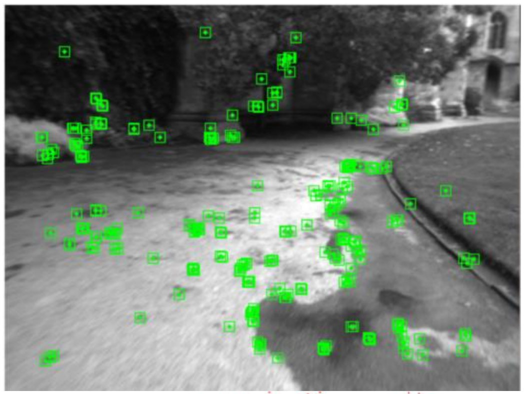
(Schönberger, Frahm, CVPR 2016)

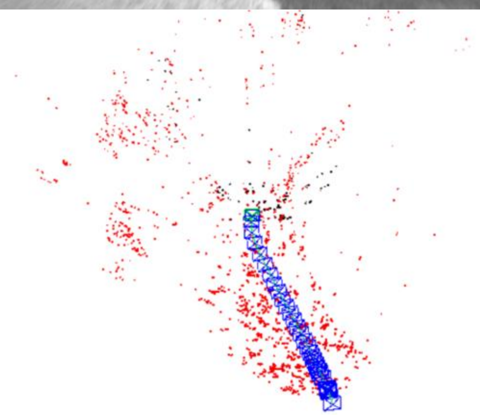
VSLAM vs. SfM



Sparse vs. Dense Reconstruction

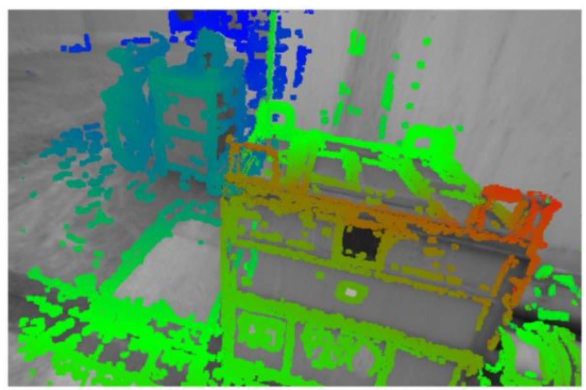
Sparse (ORB-SLAM)





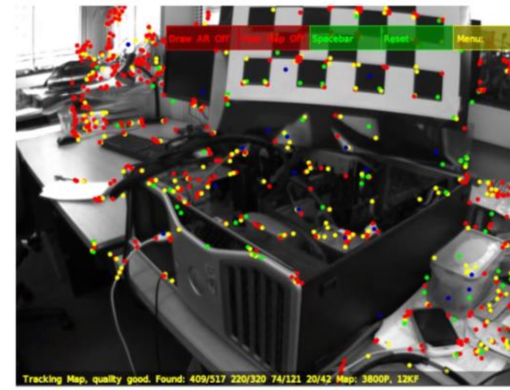
(Mur-Artal and Tardós, T-RO 2015)

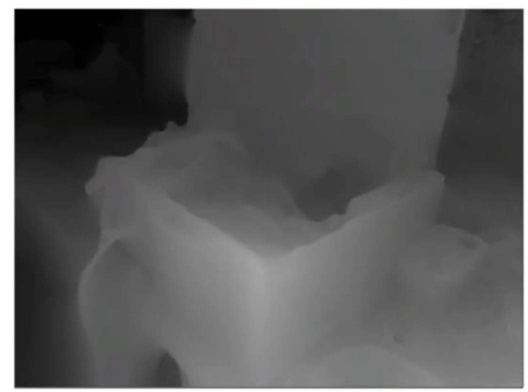
Semi-Dense (LSD-SLAM)



(Engel et al., ECCV 2014)

Dense (DTAM)

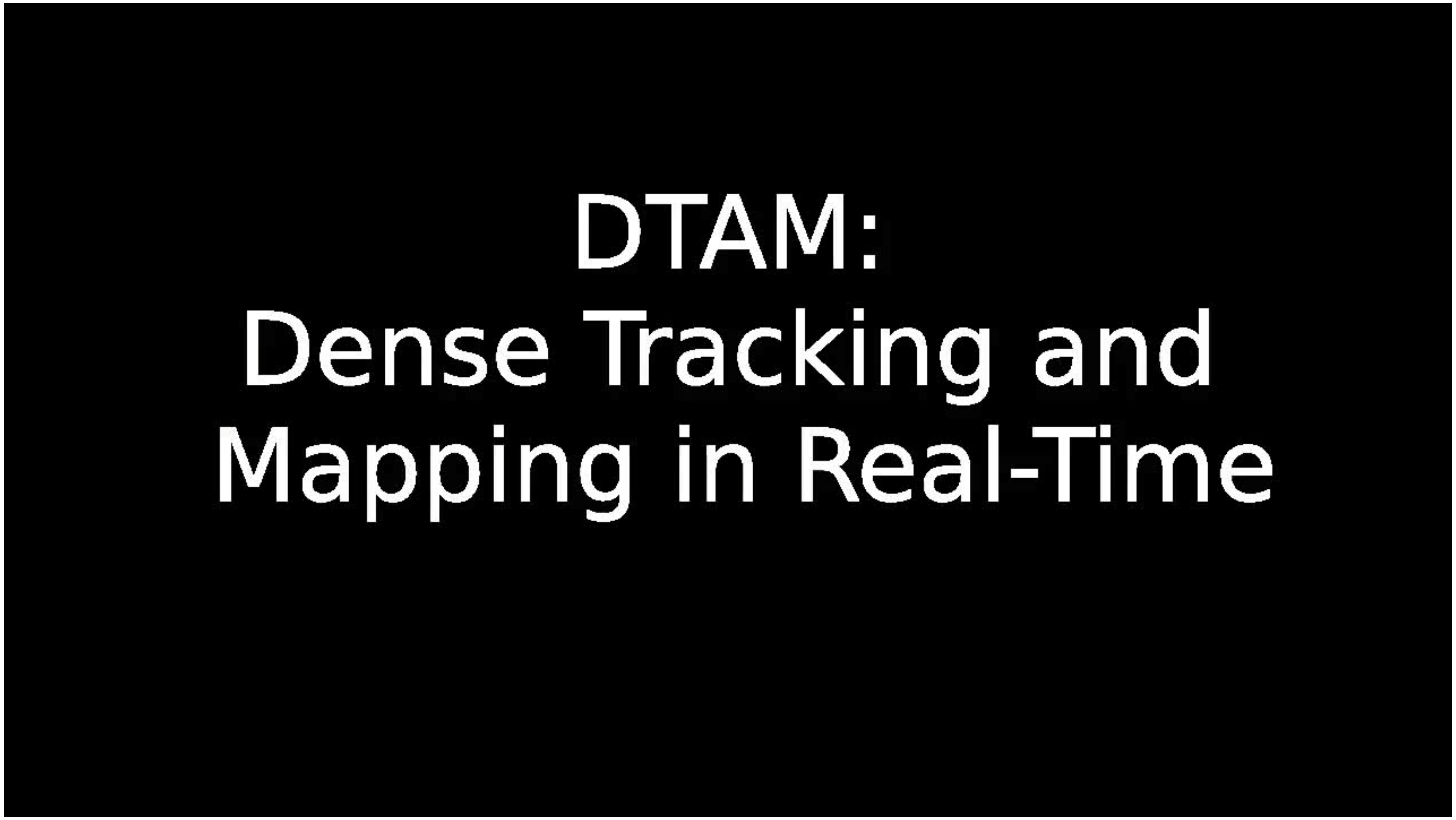




(Newcombe et al., ICCV 2011)

Good for VO/VSLAM = Good for robotic perception?

Dense VSLAM with a Single Camera



(Newcombe et al., DTAM: Dense Tracking and Mapping in Real-time, ICCV 2011)

https://www.youtube.com/watch?v=Df9WhgibCQA

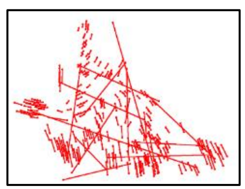
Indirect vs. Direct VO/SLAM

Indirect

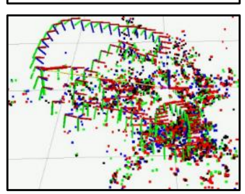
Input images

Extract and match keypoints (SIFT,BRIEF,…)

Track: min. reprojection error (point distances)



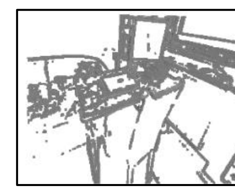
Map: estimate keypoint parameters (f.e. 3D coordinates)



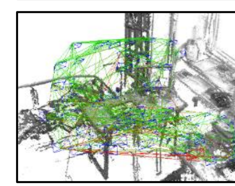
Direct

Input images

Track: min. photometric/ geometric error pixel-wise



Map: estimate per-pixel depth from photoconsistency



Indirect vs. Direct VO/SLAM

- Direct methods formulate image alignment objective in terms of photometric error (e.g. intensities)

$$E(\boldsymbol{\xi}) = \sum_i \left\| I_1(\boldsymbol{u}_i) - I_2\left(\pi(\boldsymbol{u}_i, \boldsymbol{\xi}_1^2)\right) \right\|_\gamma$$

- Indirect methods formulate image alignment objective in terms of reprojection error of geometric primitives (e.g. points, lines)

$$E(\boldsymbol{\xi}) = \sum_i \left\| \boldsymbol{x}_{1i} - \pi(\boldsymbol{x}_{2i}, \boldsymbol{\xi}_2^1) \right\|_\gamma$$

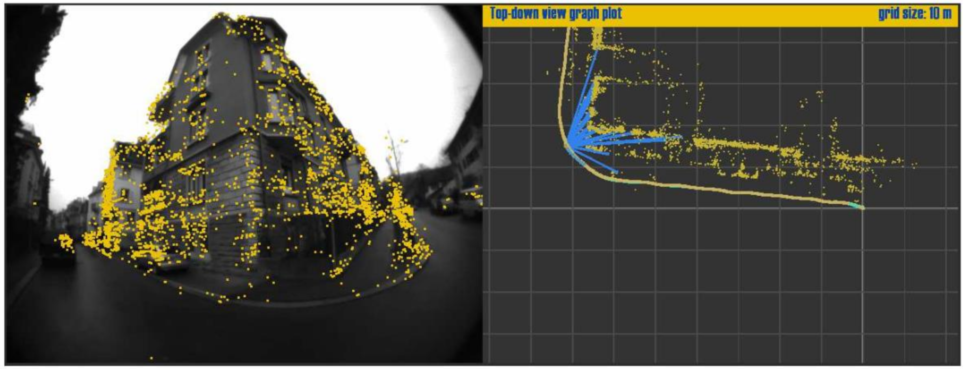
Indirect vs. Direct VO/SLAM

- Both approaches have their advantages

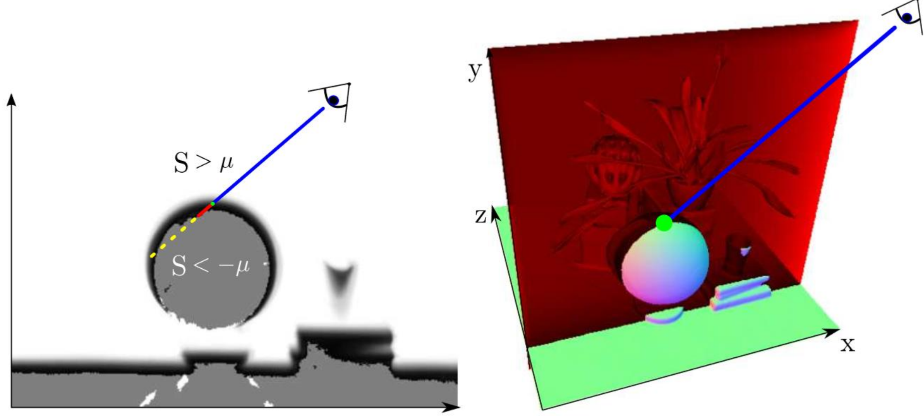
- Indirect
  - In general uses only a sparse set of primitives (e.g. points) for estimation
  - Highly dependent on feature detector
    - Often has problems in textureless areas
    - Dependented on detection accuracy
  - Better convergency
    - Since estimation is performed based on establised (matched) correspondences between images
  - Less influenced by camera/sensor properties
    - Vignetting, changing exposure, rolling shutter, etc.

- Direct
  - Can make used of entire image information
  - Intensities can be optimized up to sub-pixel accuracy
  - Needs good initialization
    - Since no correspondences are eastablished
  - Camera/sensor properties need to be treated explicitely
    - Compensation of vignetting, exposure changes, rolling shutter, etc-

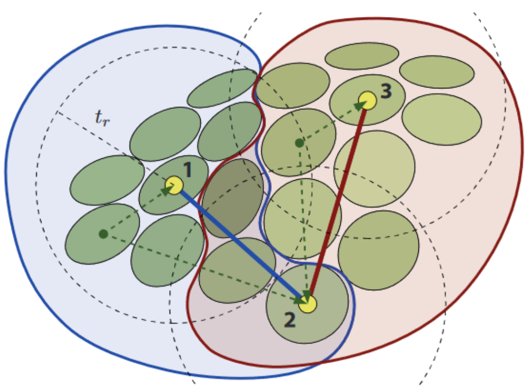
How Should We Represent The Map?



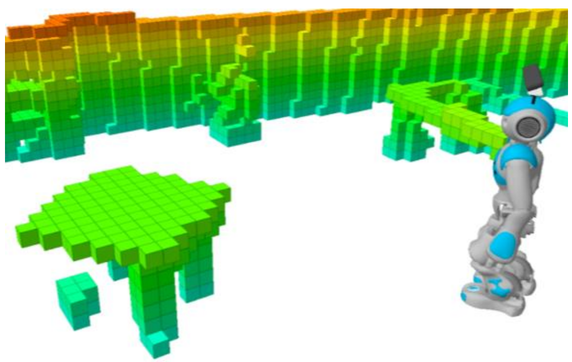
Sparse interest points



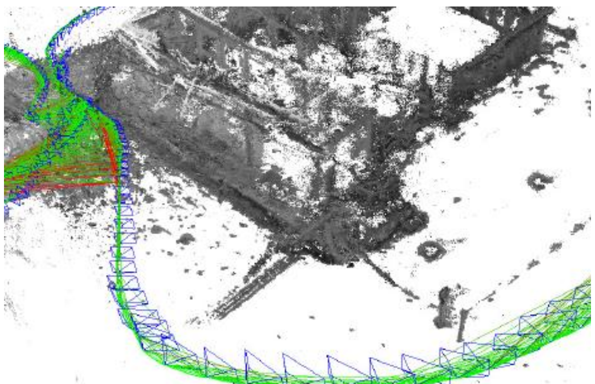
Volumetric, implicit surface



Explicit surface
(surfels, mesh,…)



Volumetric, occupancy

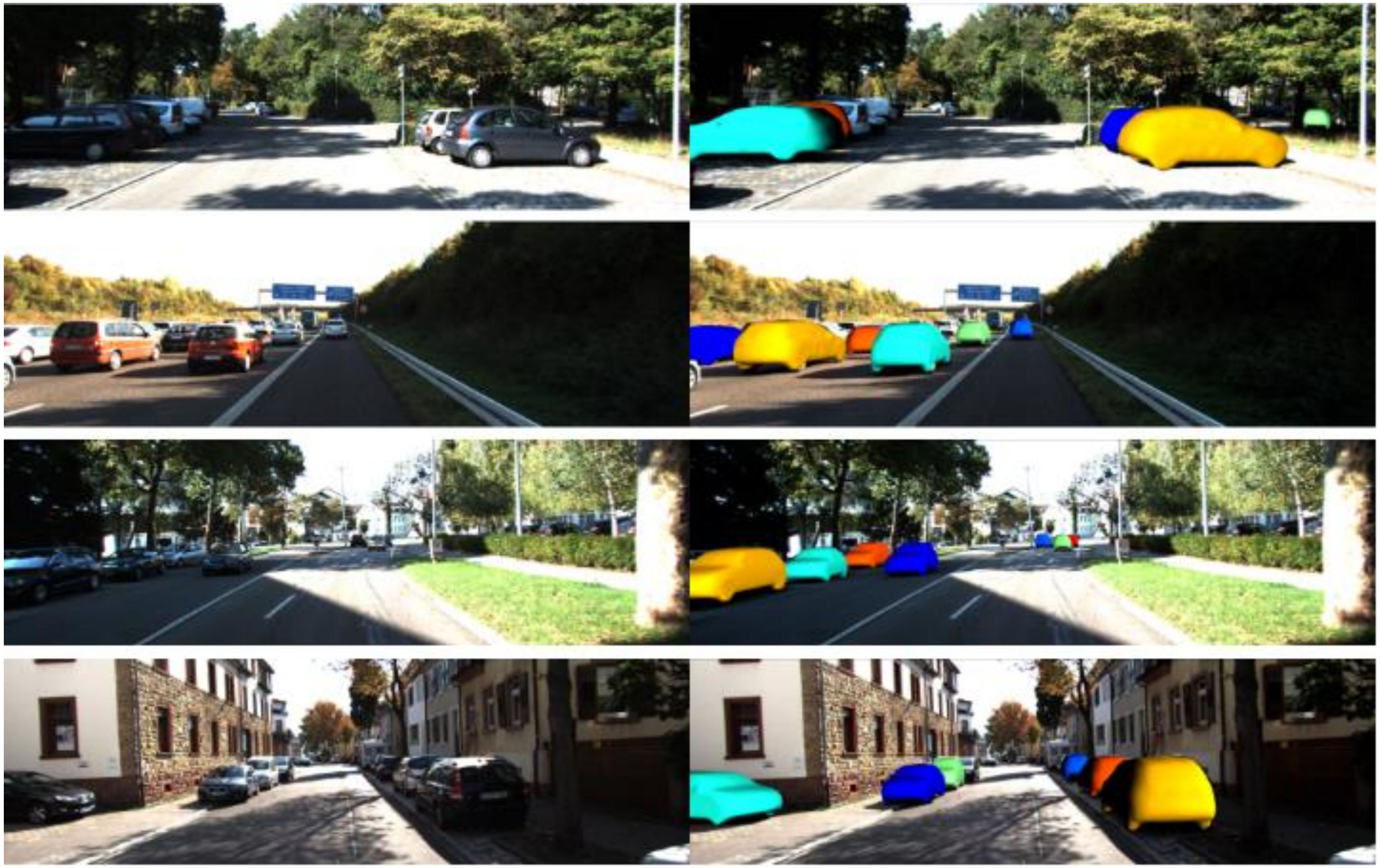


Keyframe-based maps

Good for VO/VSLAM = Good for robotic perception?

(Lynen et al., RSS 2015), (Newcombe, 2015), (Weise et al., 2009), (Maier et al., 2012), (Engel et al., ECCV 2014)

3D Object Detection and Tracking



(Wang, Stückler, Cremers ICRA 2020)

3D Object Detection and Tracking

ECCV 2018
European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Stereo Vision–based Semantic 3D Object and Ego–motion Tracking for Autonomous Driving

Peiliang Li, Tong Qin and Shaojie Shen | **HKUST UAV Group**

http://uav.ust.hk/

(Li, Qin, Shen, ECCV 2018)

https://www.youtube.com/watch?v=nE2XtCvPEDk

3D Object Detection and Tracking

- Visual 3D object detection…
  - …finds an object in an image and
  - …estimates its 6-DoF pose from the image

- Visual 3D object tracking…
  - …tracks the 6-DoF pose of an object in an image sequence
- Multi-object tracking involves data association

Course Contents

- Image formation, multi-view geometry, SE3 (recap)
- Probabilistic filtering, non-linear least squares
- Visual odometry
- Visual-inertial odometry
- Visual SLAM
- Dense reconstruction
- Map representations
- 3D object detection and tracking

Thanks for your attention!

Slides Information

- These slides have been initially created by Jörg Stückler as part of the lecture “Robotic 3D Vision” in winter term 2017/18 at Technical University of Munich.
- The slides have been revised by myself (Niclas Zeller) for the same lecture held in winter term 2020/21
- Acknowledgement of all people that contributed images or video material has been tried (please kindly inform me if such an acknowledgement is missing so it can be added).