Lehrstuhl Informatik 8
Computergrafik und
Multimedia

RWTH AACHEN UNIVERSITY

Challenges in Dynamic Visual Scene Understanding: Beyond Tracking

Bastian Leibe

Computer Vision Group
Computer Science 8
RWTH Aachen University

DAGM UP Workshop, Saarbrücken, 03.09.2013

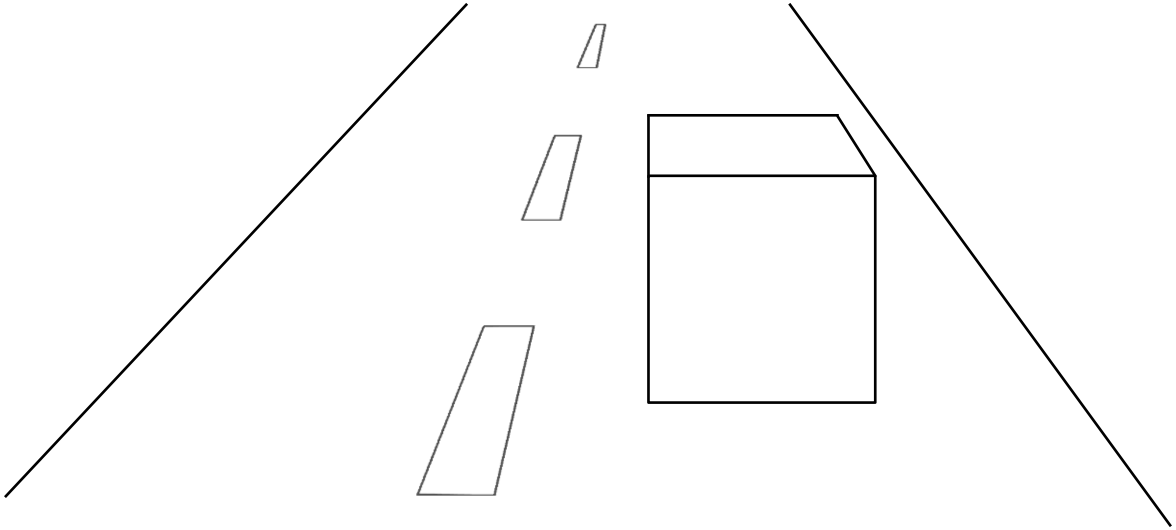
erc
European Research Council
Established by the European Commission

funded by ERC StG 307432
CV-SUPER

UMIC

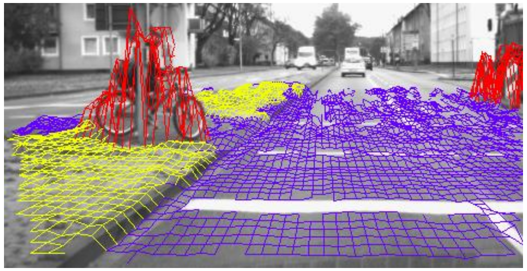
Computer Vision Group

Pedestrian Detection in Cars – Why?



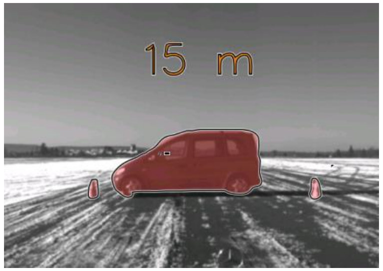
- **It is NOT necessary to detect obstacles!**

Dense Stereo



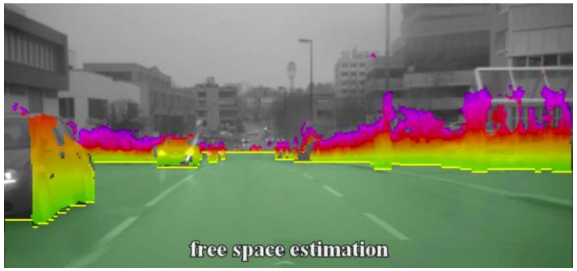
[Oniga & Nedevschi, TVT’09]

Optical flow



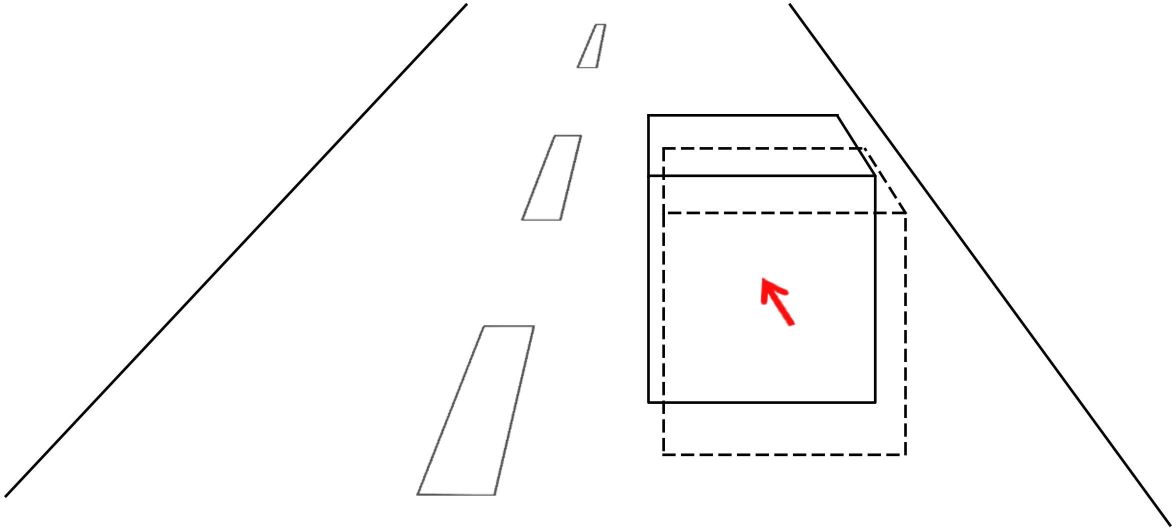
[Wedel et al., DAGM’07]

Road modeling



[Wedel et al., TITS’09]

Pedestrian Detection in Cars – Why?



- **It is NOT even necessary to track them!**

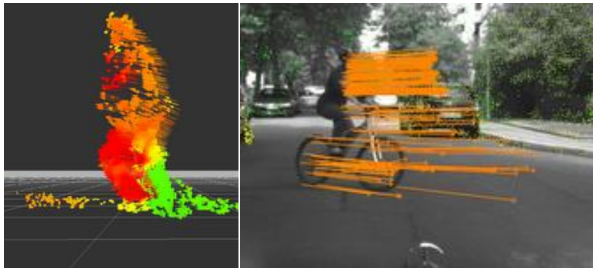
Particle based Occupancy Grids

[Danescu et al., TITS'11]

LIDAR based tracking

[Teichman & Thrun, RSS']

Scene Flow, Dense6D



[Wedel et al., ECCV'08]
[Franke et al., '12]

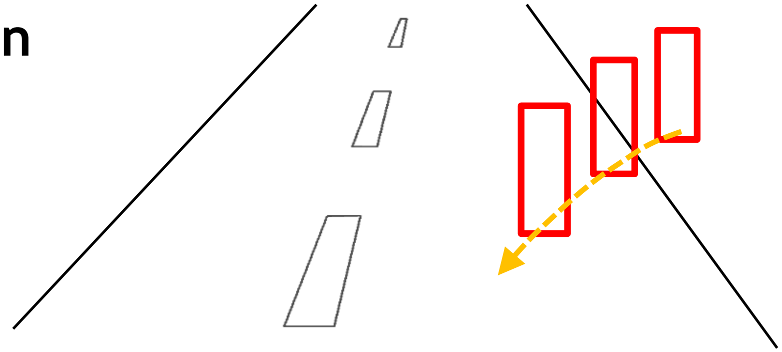
Two Main Reasons for Object Detection

- Robustness
  - Tracking requires f-g segmentation

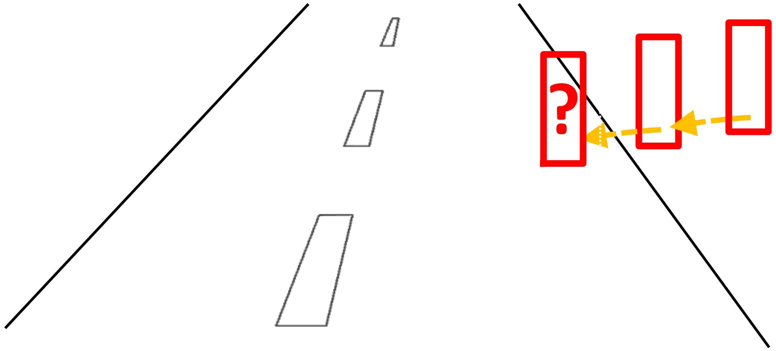
  ⇒ Very challenging task

  - Pedestrians are important

  ⇒ Detection failure is not an option

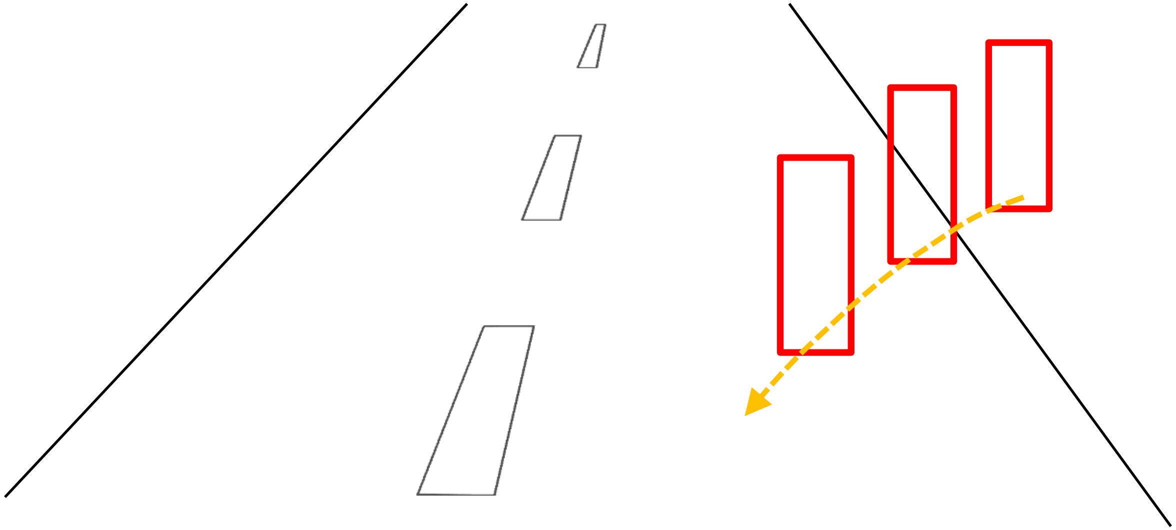


- Semantics
  - Use class-specific motion models to make better predictions



To what extent do we live up to those promises?

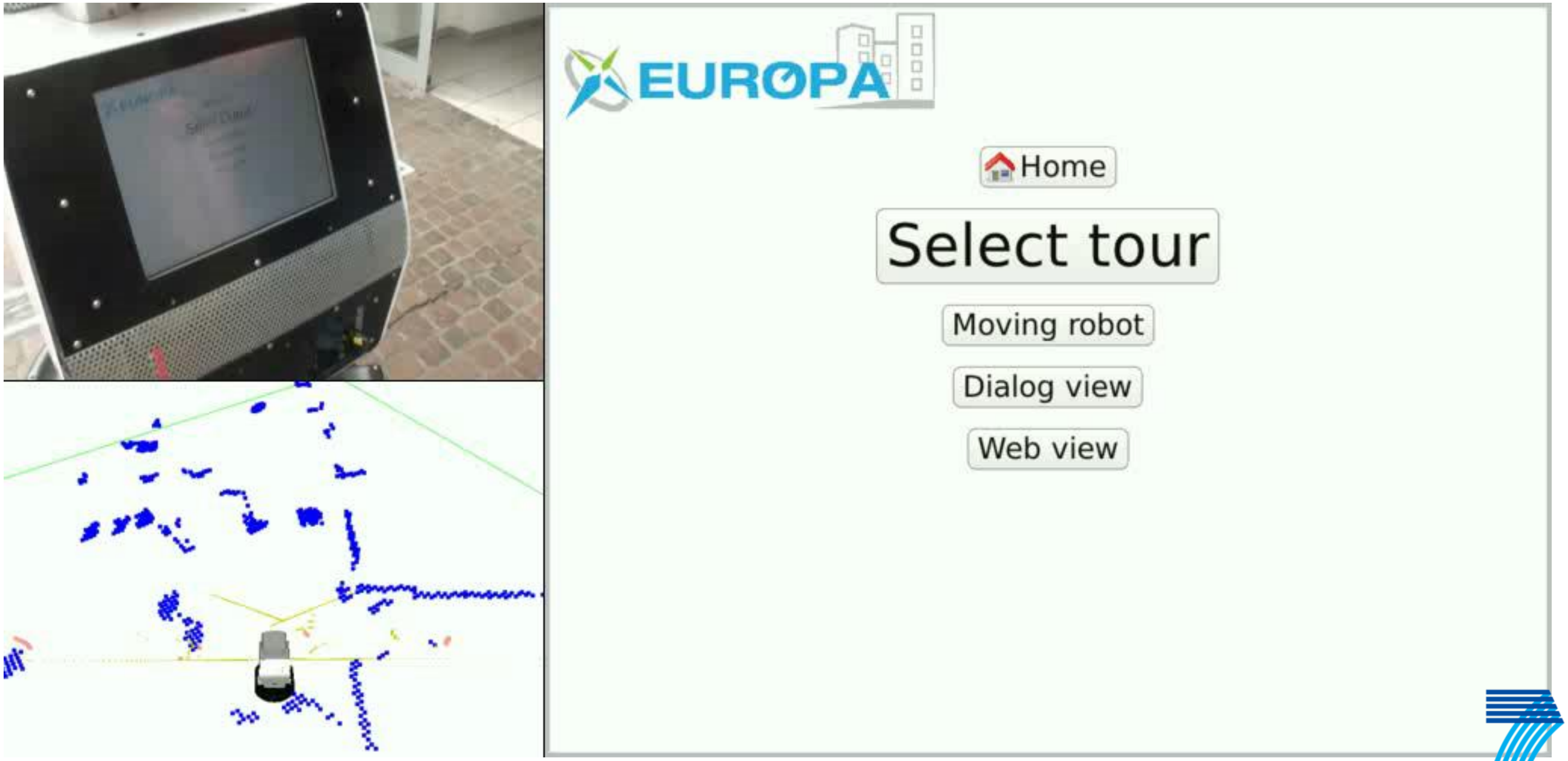
Mobile Object Detection & Tracking



- Standard approach: Tracking-by-Detection
  - Detect all objects in each frame
  - Link detections into trajectories
  - Multi-hypothesis handling for additional robustness

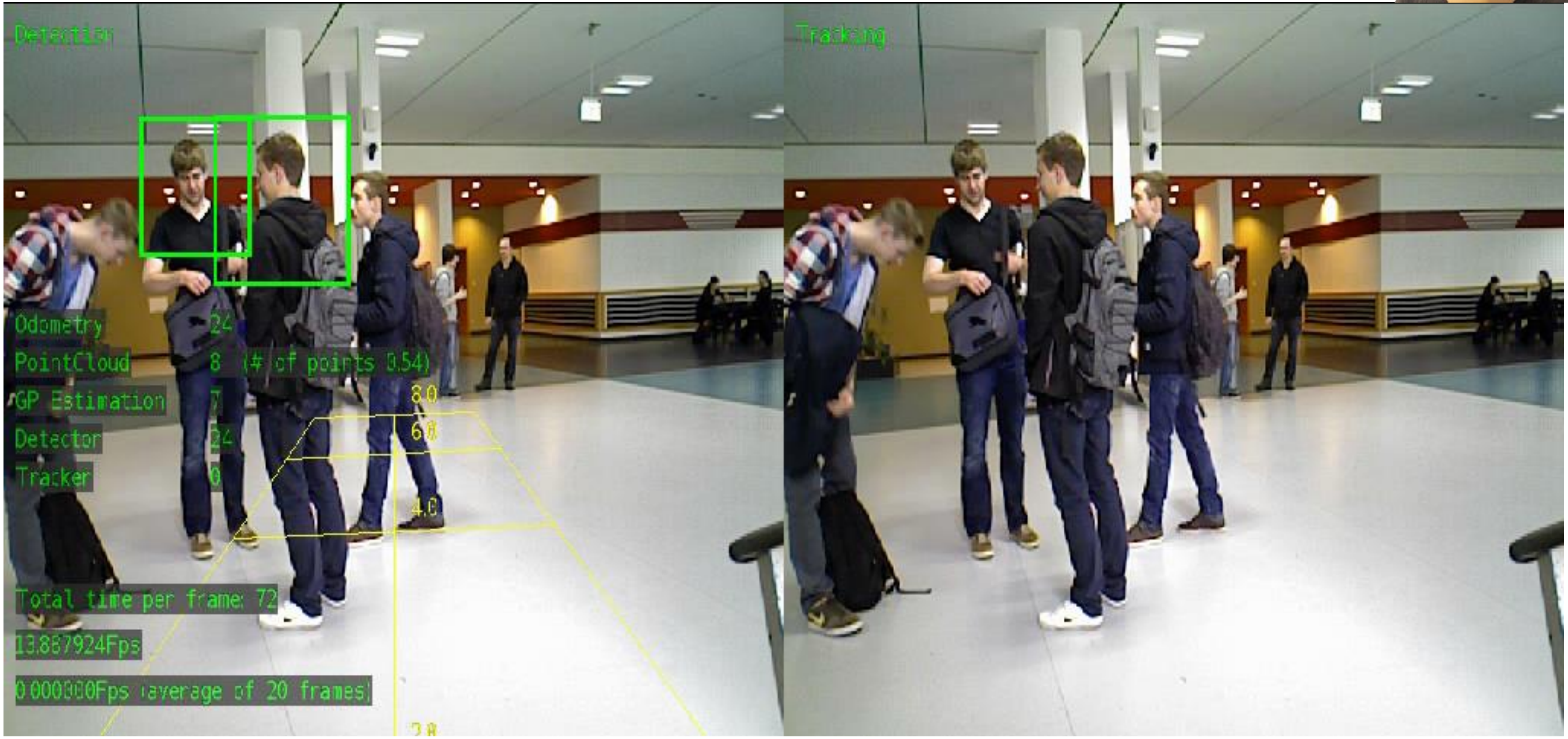
⇒ *Successfully used for tracking pedestrians and cars*

Real-Time Application on a Mobile Robot



link to the video

Most Recent Version (Demo at CVPR'13)



```
Odometry          24
PointCloud        8  (# of points 0.54)
GP Estimation     7
Detector          24
Tracker           0

Total time per frame: 72
13.887924Fps
0.000000Fps (average of 20 frames)
```

- **Kinect-based head-worn setup**
  - Person detection + Tracking + Visual odometry + GP estimation
  - Result: 20-35 fps on single CPU core (no GPU involved!)
    15 fps with additional far-range detector (on the GPU)

So, Are We Done?

- Limitations
  - Tracking a single object class (typically pedestrians or cars)
  - How can we scale this to 100s of categories?

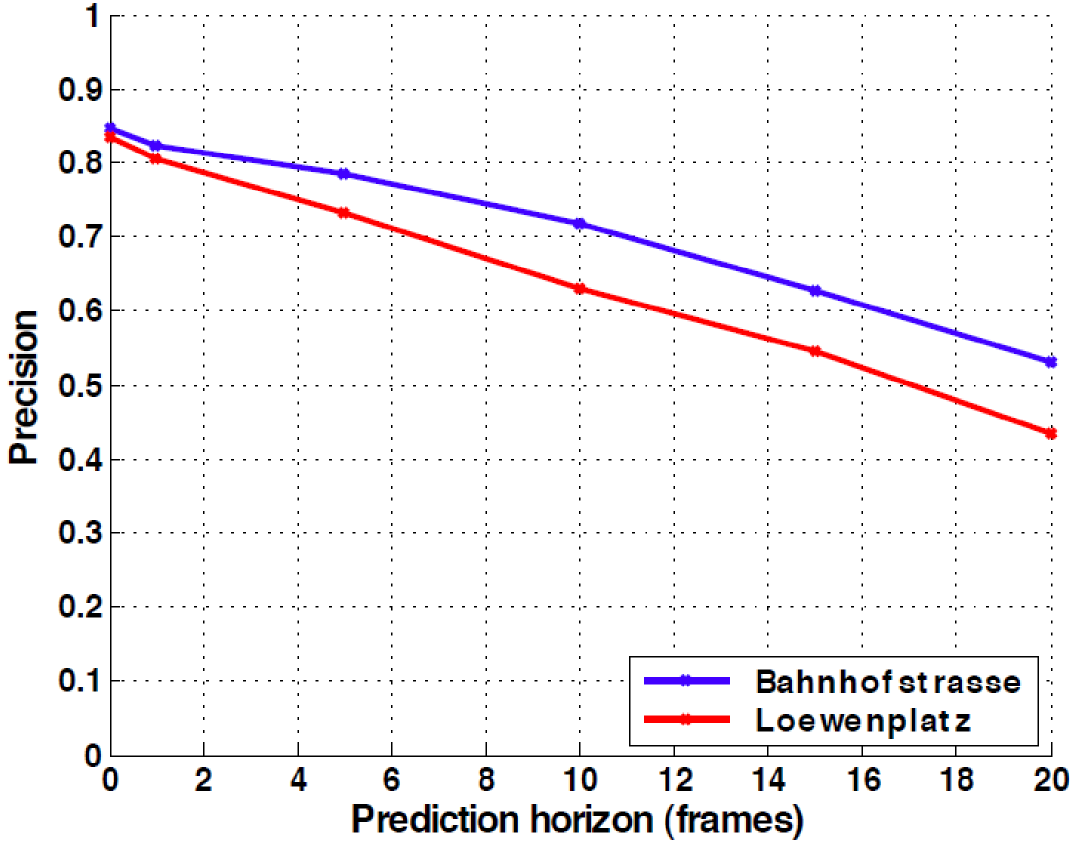
⇒ ***We can't. Tracking-by-detection is inherently not scalable.***

So, Are We Done?

- Limitations
  - Tracking a single object class (typically pedestrians or cars)
  - How can we scale this to 100s of categories?
  - At least we can make predictions for the tracked classes, right?

⇒ *Not really. Only short-term predictions are reasonably good.*

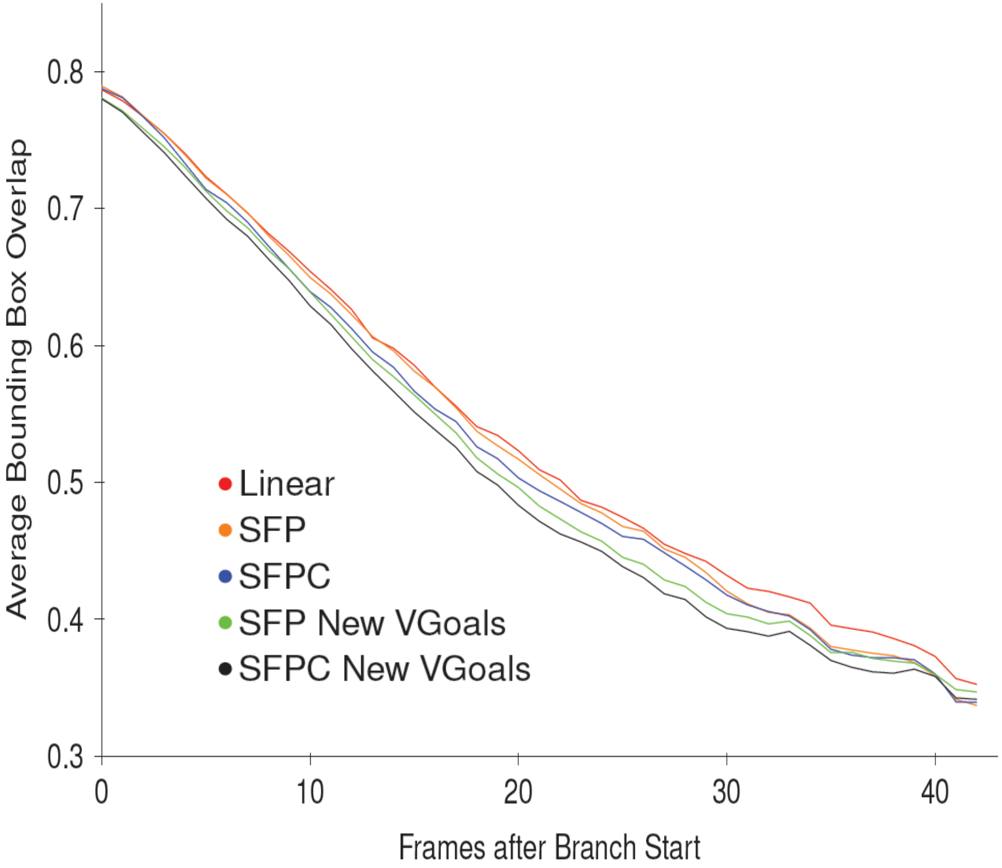
KF Tracking Prediction is of Limited Use

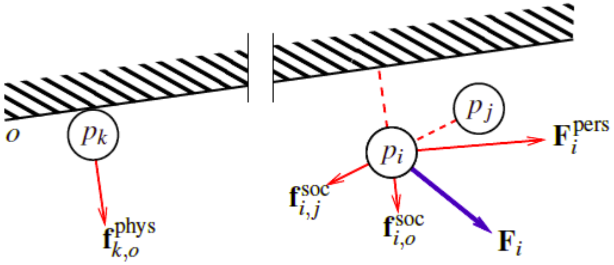


- KF prediction accuracy drops significantly beyond 1.5s
  - Within this time frame, people are mostly ballistic

[Ess, Schindler, Leibe, Van Gool, IJRR'10]

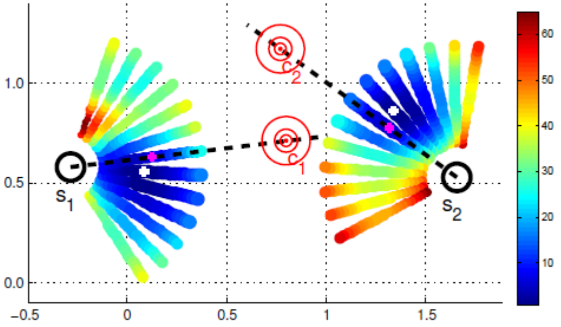
Even Social Walking Models Don’t Help Much




o
p_k
p_j
p_i
$\mathbf{F}_i^{\text{pers}}$
$\mathbf{f}_{k,o}^{\text{phys}}$
$\mathbf{f}_{i,j}^{\text{soc}}$
$\mathbf{f}_{i,o}^{\text{soc}}$
$\mathbf{F}_i$

Force-based model

[Luber et al., ICRA’10]



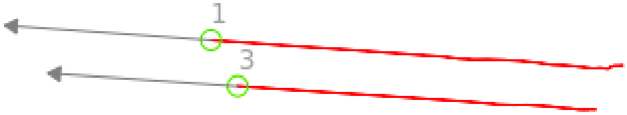
LTA model

[Pellegrini et al., ICCV’09]

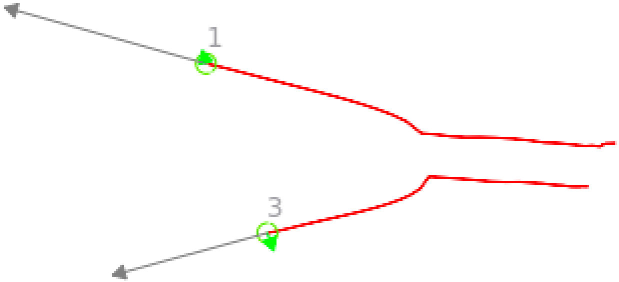
- **Hard to outperform linear prediction on average**
  - **There are too many factors that need to be modeled...**

[P. Fischer, D. Mitzel, B. Leibe, unpublished]

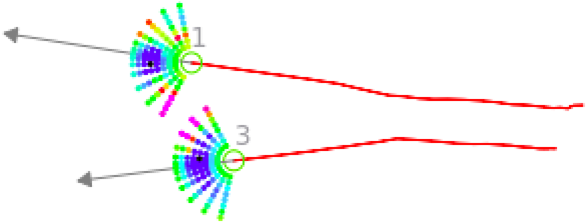
Limits of Social Walking Models: Groups



Linear prediction

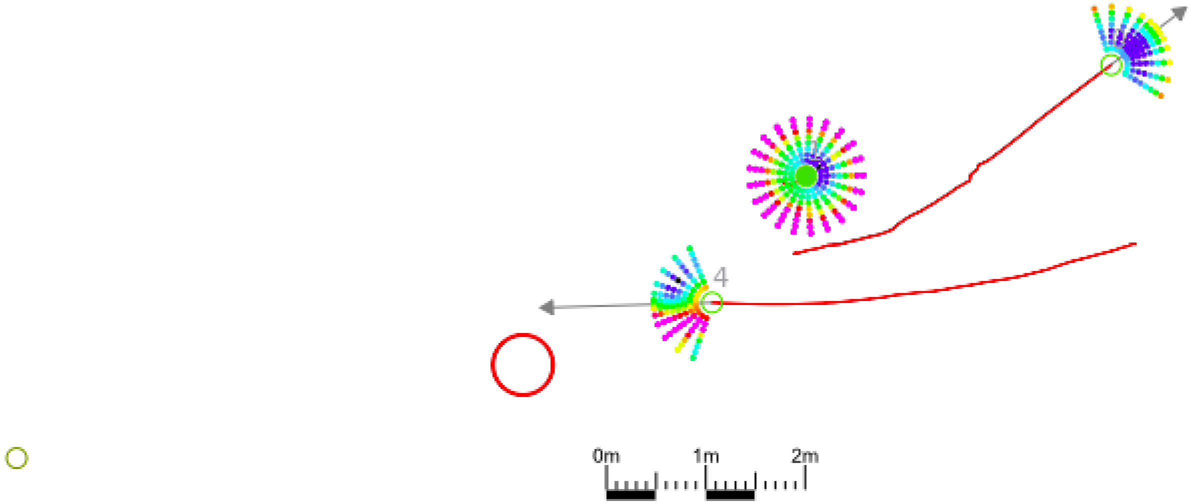


Force-based models



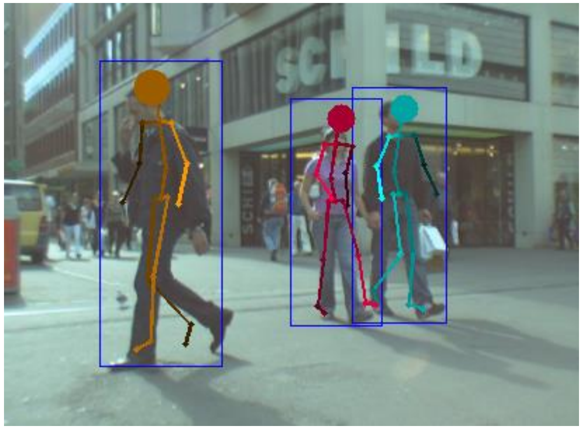
LTA model

Limits of Social Walking Models: Goal Locations

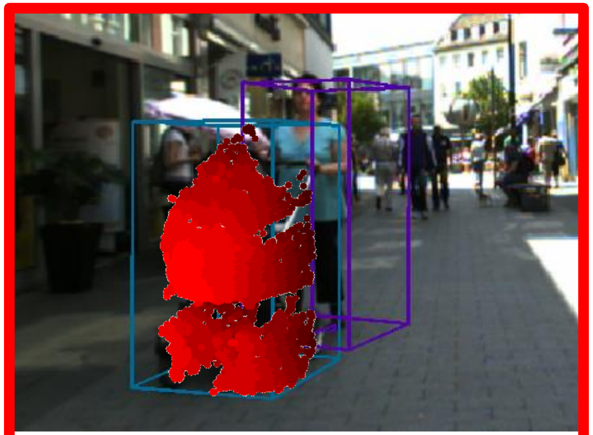


- **Difficulties**
  - To calculate evasive behavior, goal location needs to be known
  - Resulting behavior varies wildly with changing goal location
  - Goal locations are often not visible in the image
  - Is a person walking towards its goal or is it evading something?

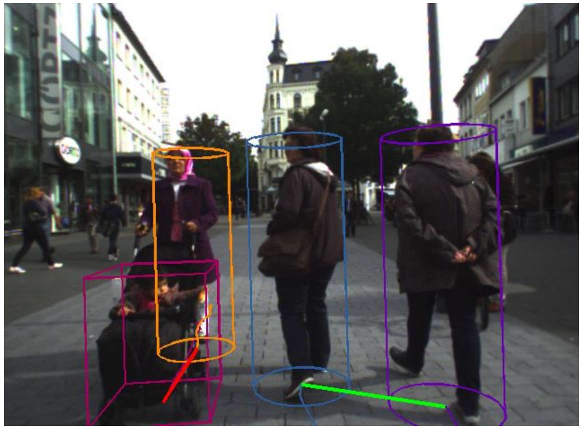
Postulate: We Need More Detailed Analysis...



...of people

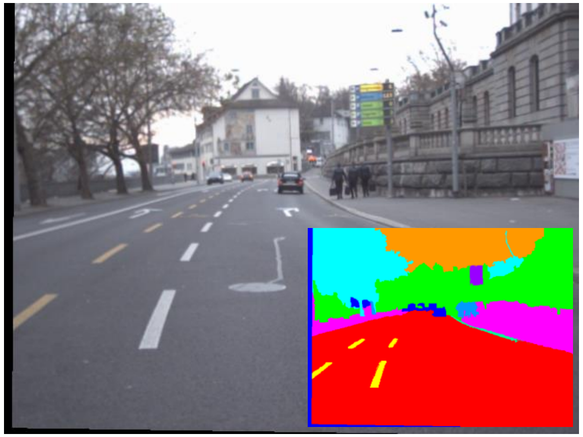


...of objects

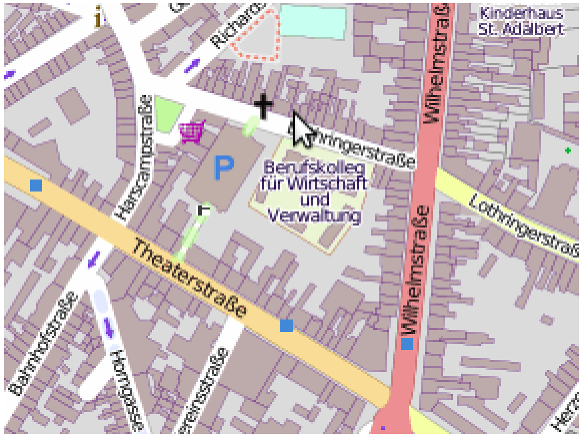


...of interactions

...of social behaviors

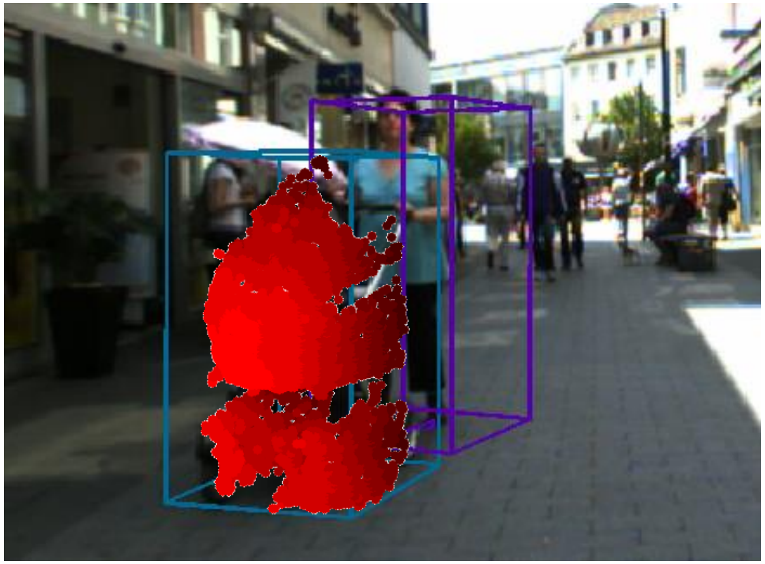


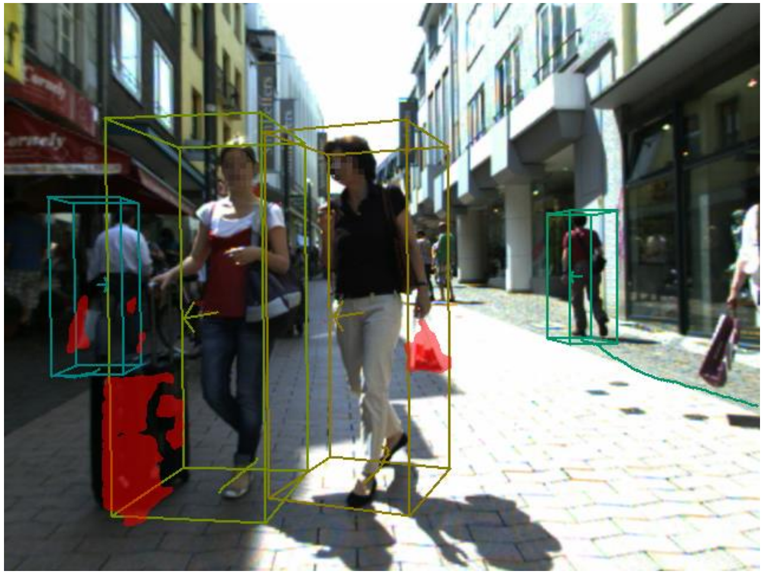
...of the environment



...of the surroundings

Tracking Unknown Objects





- Goal
  - Recognize and track large variety of unknown objects
- Challenges
  - Large variety of objects, pre-trained detectors not feasible
  - Segmentation problem: What is an object?

Approach: Tracking-*before*-Detection

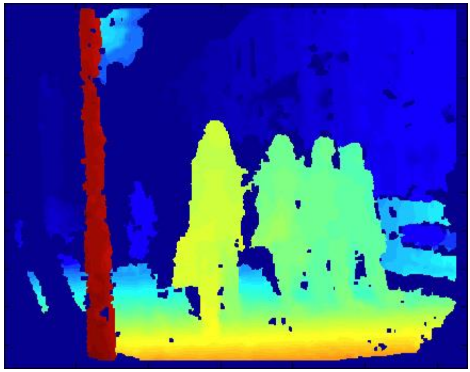
Reversing the traditional pipeline...

- Basic idea
  - Extract a (potentially overcomplete) set of object candidates
  - Try to track each of them for several frames.
  - If we manage to do this for a candidate, it's probably an object.
  - We can then still apply a *classifier* to determine its category...
  - ...or *postpone* this to a later point (when it's better visible).

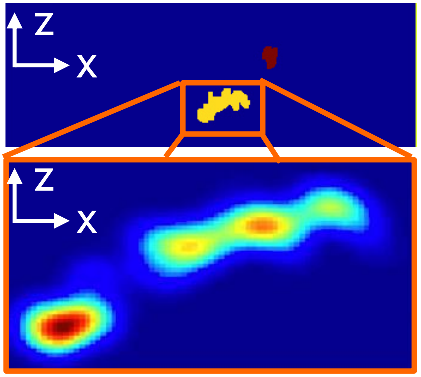
- In order to do this from a mobile setup, we need
  - A generic object candidate generation method
  - A robust low-level tracking approach

Stereo Tracking-*before*-Detection

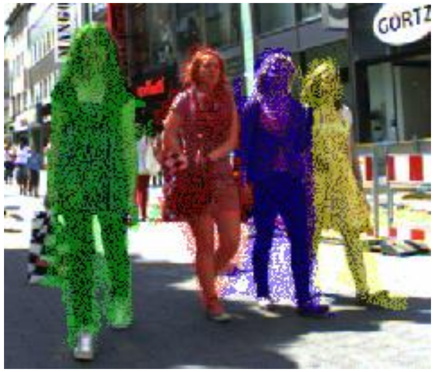
Image



Depth Map



Ground projection

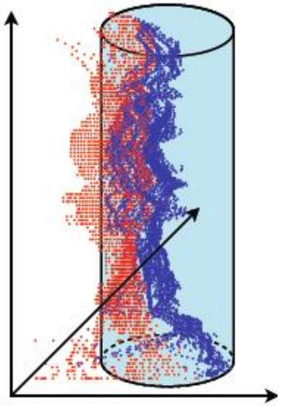


Segmented ROIs

2. Region-of-Interest (ROI) extraction
   - Estimate ground plane from stereo depth
   - Project 3D points onto ground plane
   - Segment individual objects in projection image
3. Track all objects using 3D information
   - Use ICP for 3D point cloud registration

   ⇒ Tracking entirely in 3D

   ⇒ Problem: Limited depth resolution!

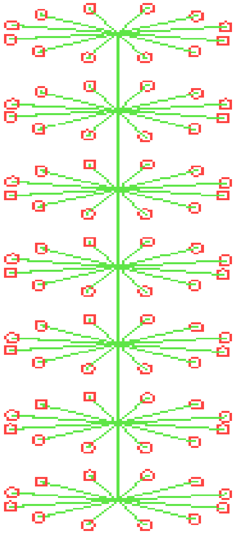


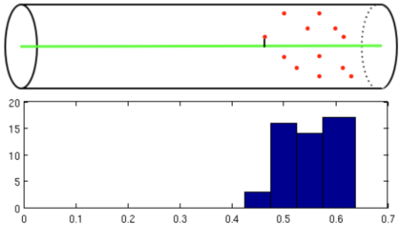
ICP Tracking

[D. Mitzel, B. Leibe, CORP’11, ECCV’12]

Model: Generalized Christmas Trees (GCT)

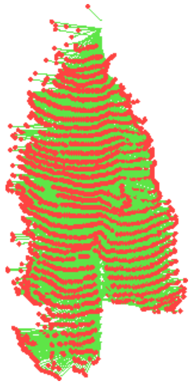
- Idea
  - Integrate depth measurements over time to smooth out noise
  - Build up object model online
- GCT Model structure
  - Central axis
  - Uniformly sampled rays at different height levels
  - Distance distribution per ray
- Model captures
  - Surface details (median depth)
  - Variation caused by noise and articulations



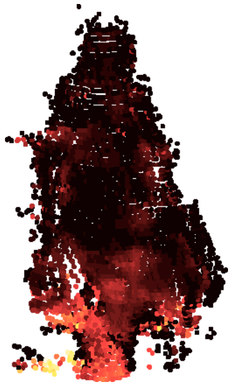


Dist. distribution per ray

Object



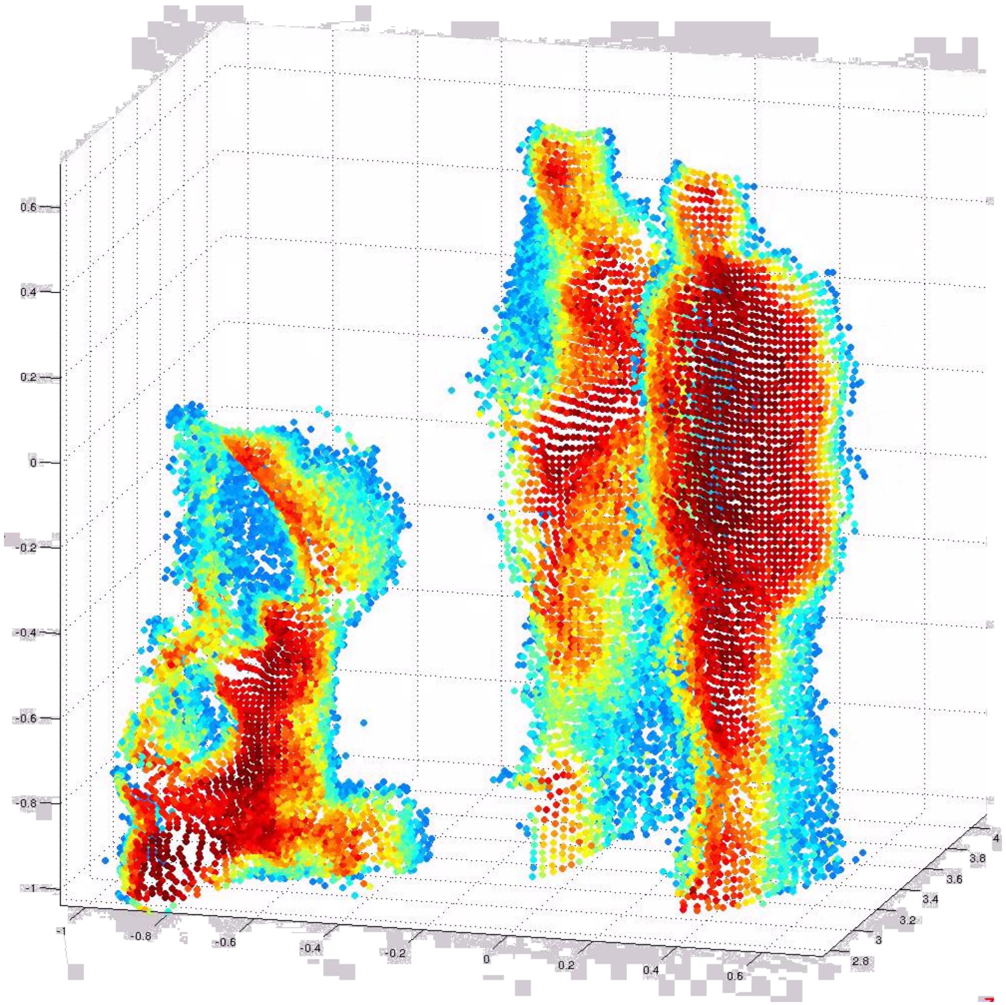
Median depth

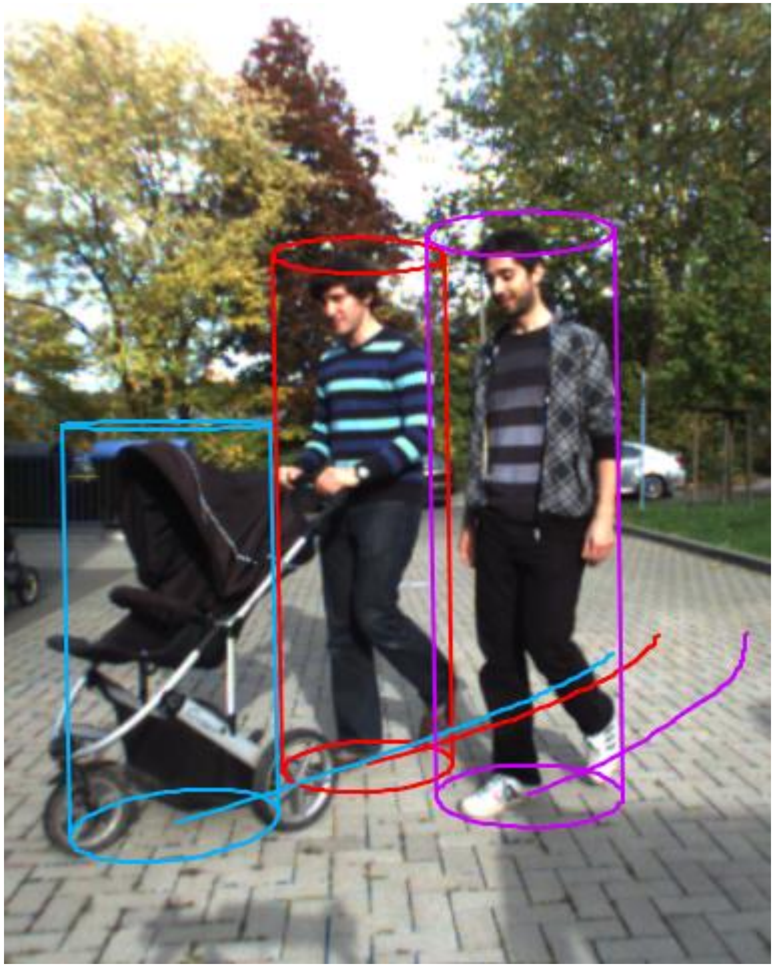


Variances

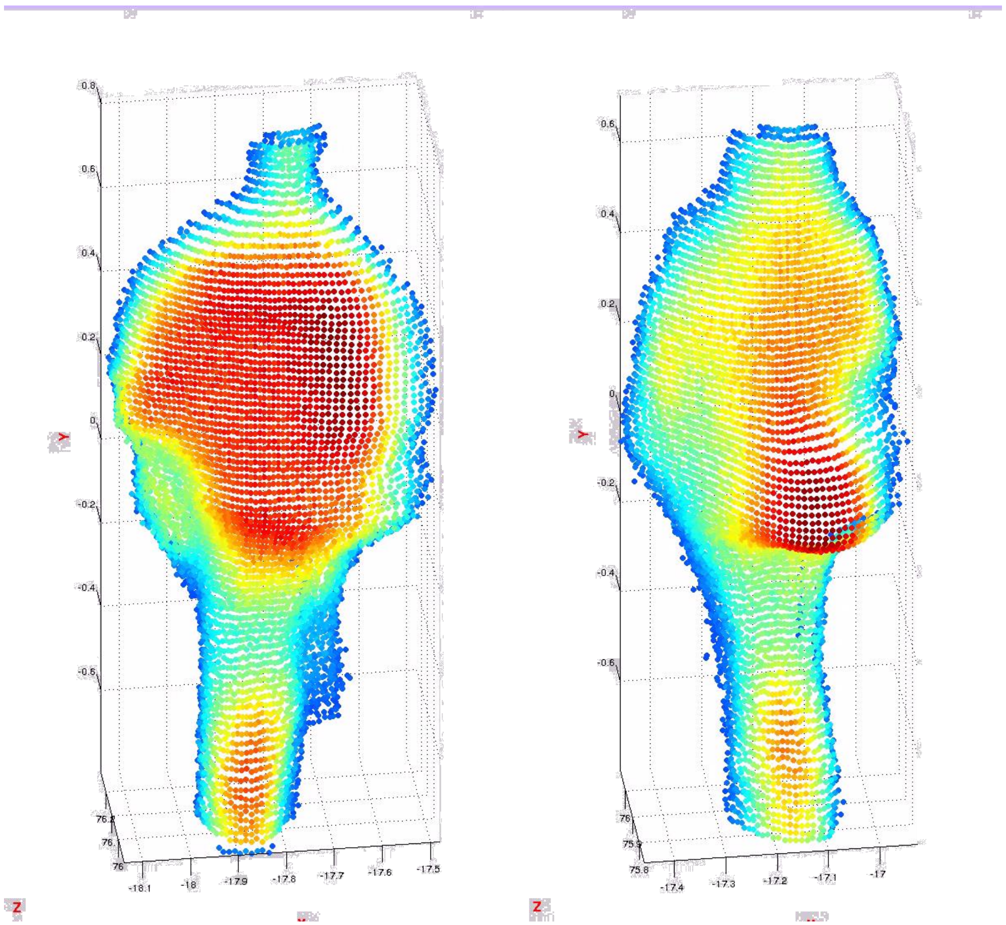
[D. Mitzel, B. Leibe, ECCV'12]

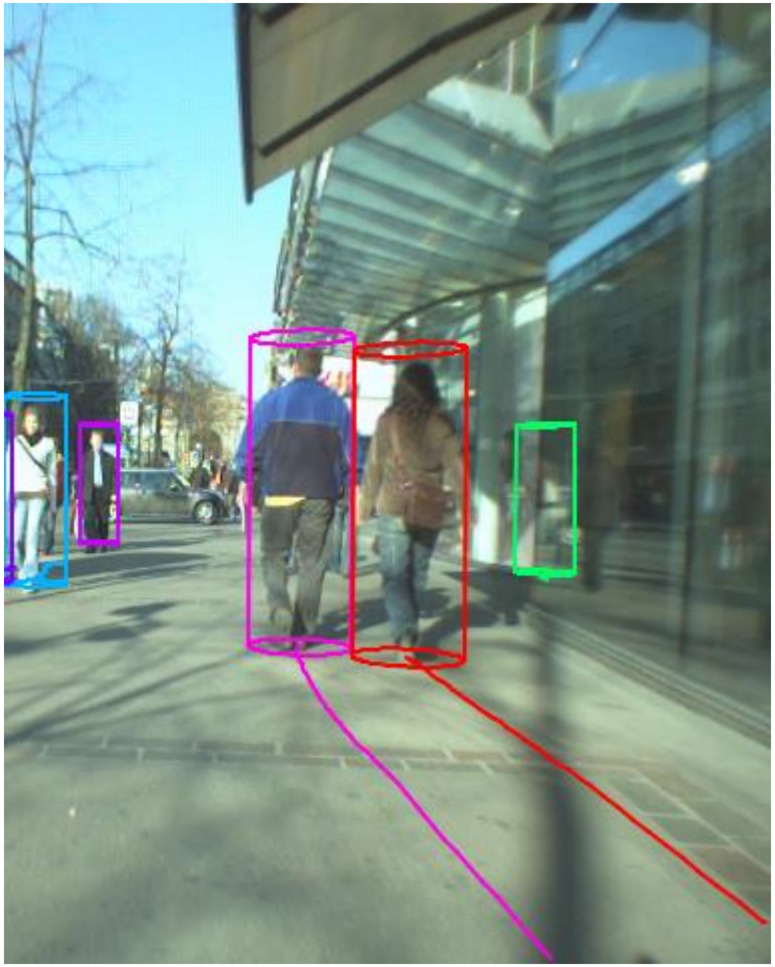
Example GCTs



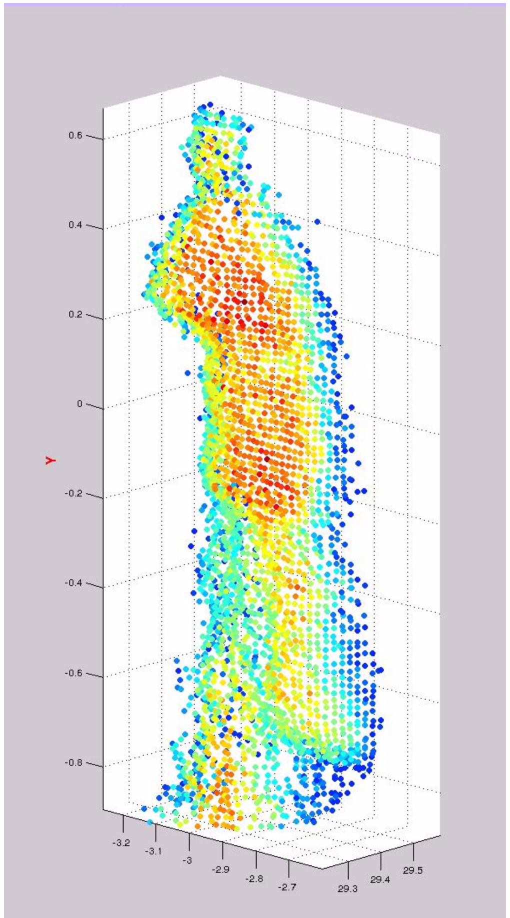


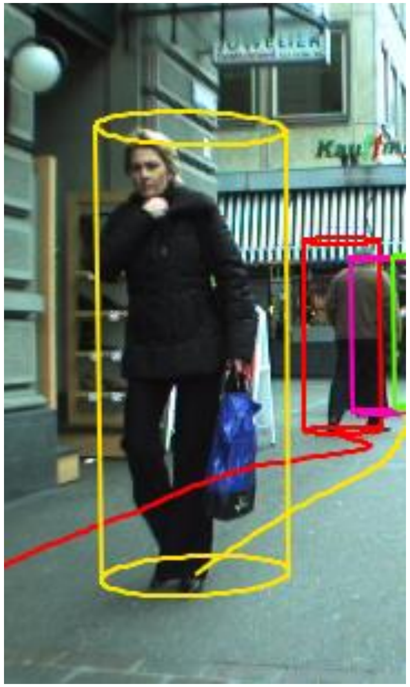
Example GCTs (2)

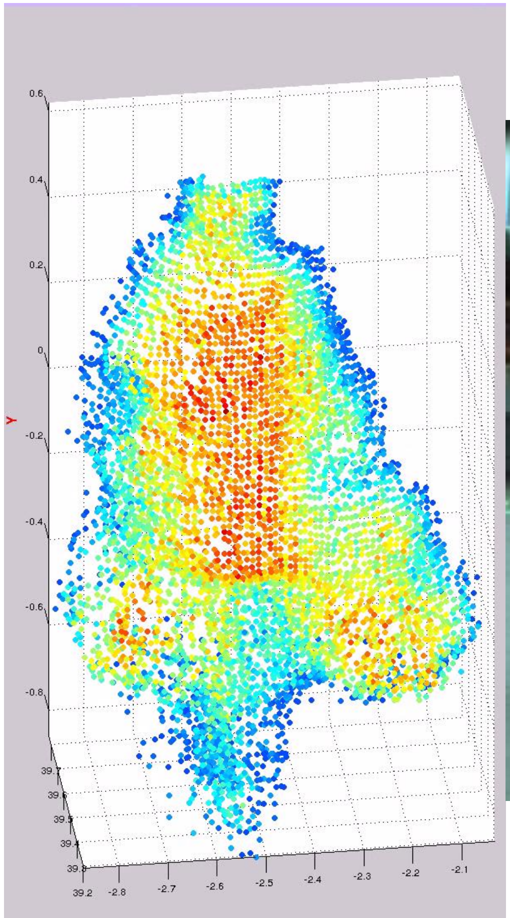


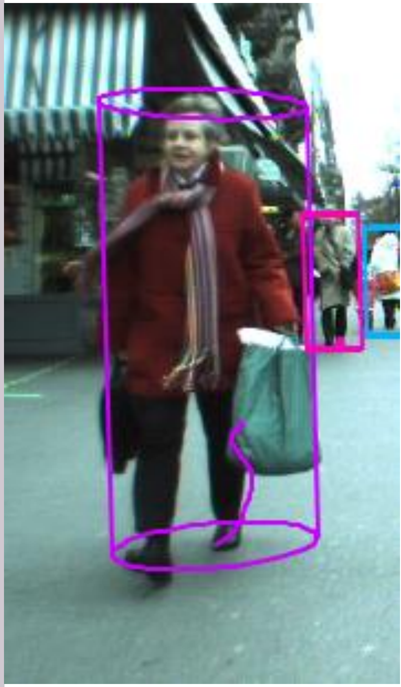


Example GCTs (3)

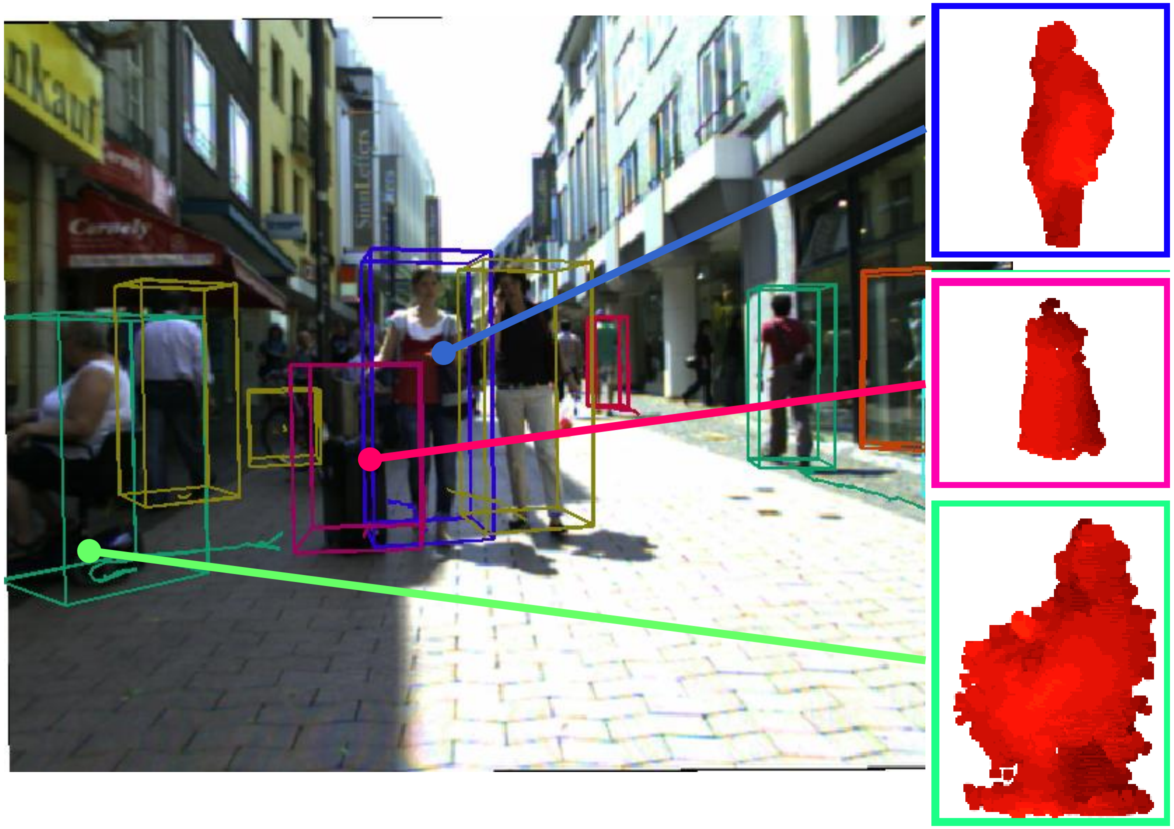








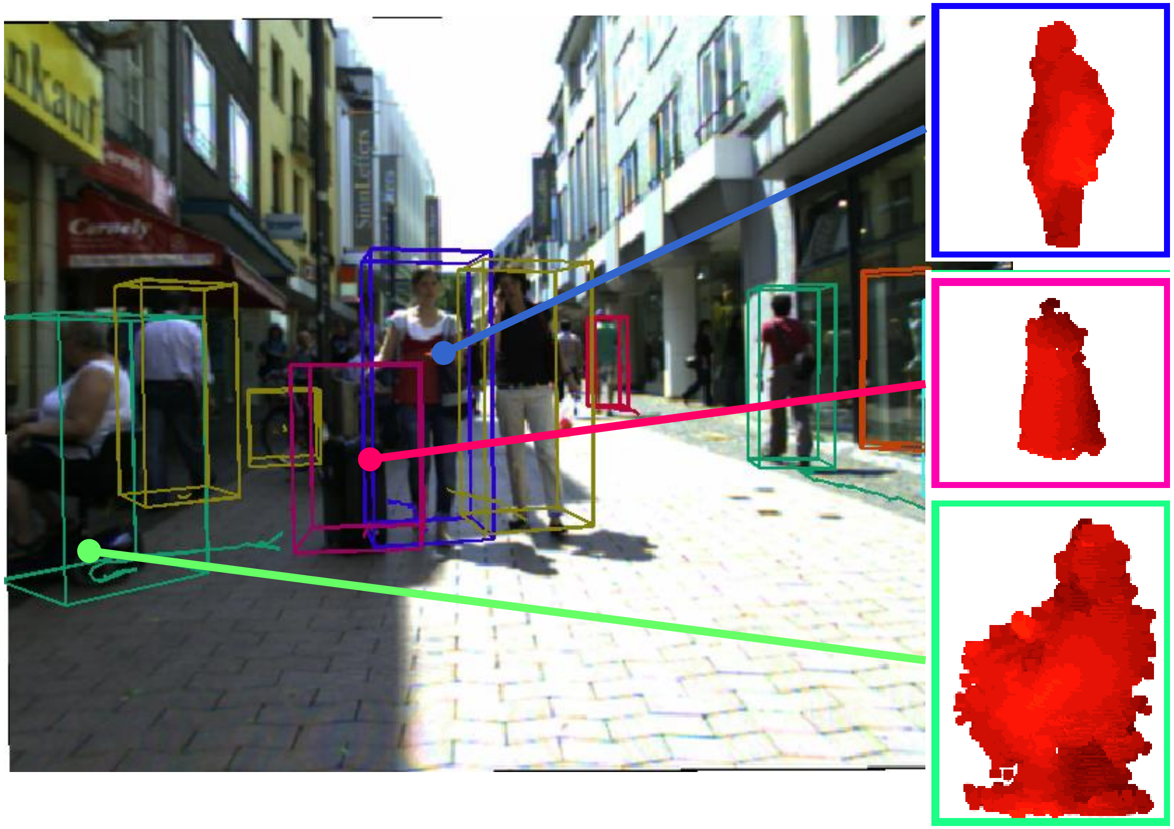
Tracking Known and Unknown Objects...

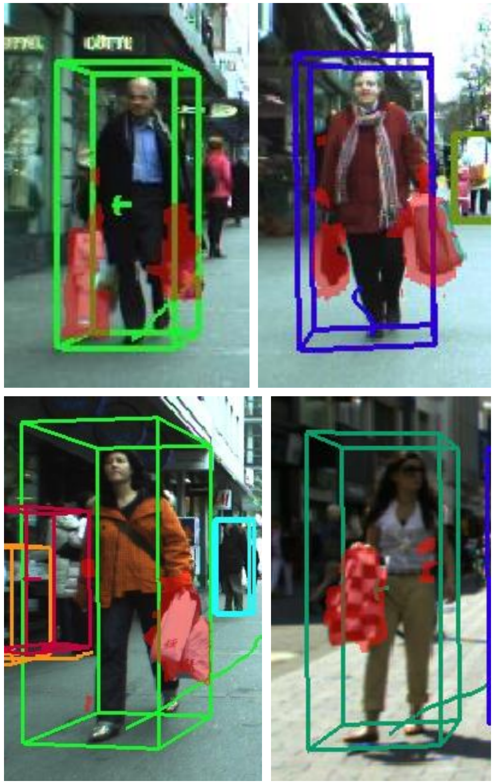


- Tracking-*before*-detection pipeline
  - Tracking fully based on ICP, detector only for verification
  - Build up 3D object models online

[D. Mitzel, B. Leibe, ECCV'12]

Tracking Known and Unknown Objects...

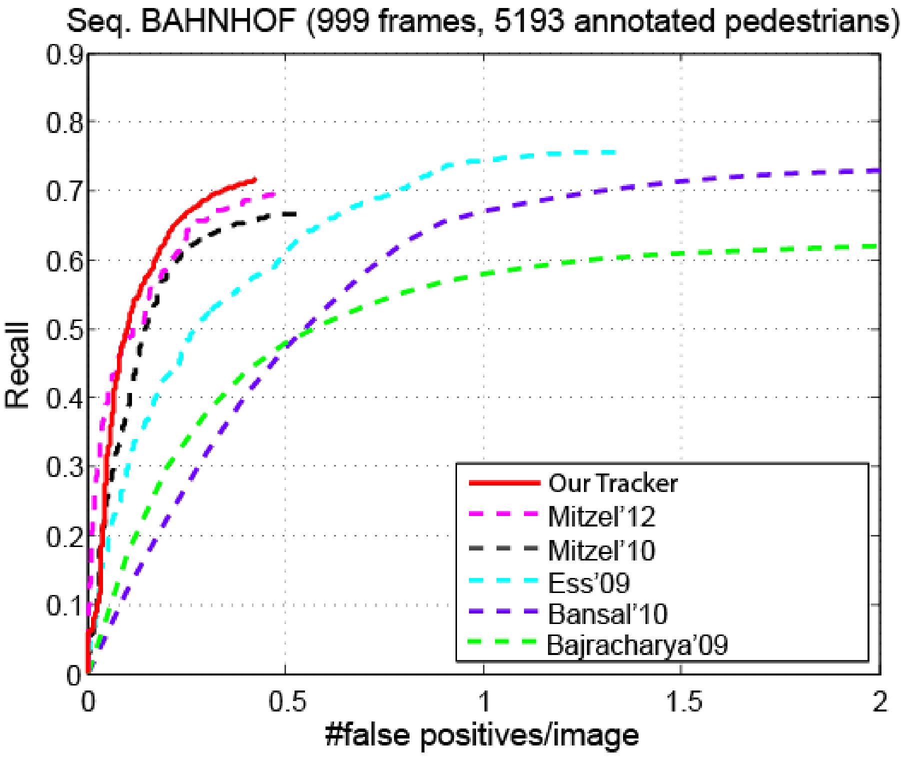


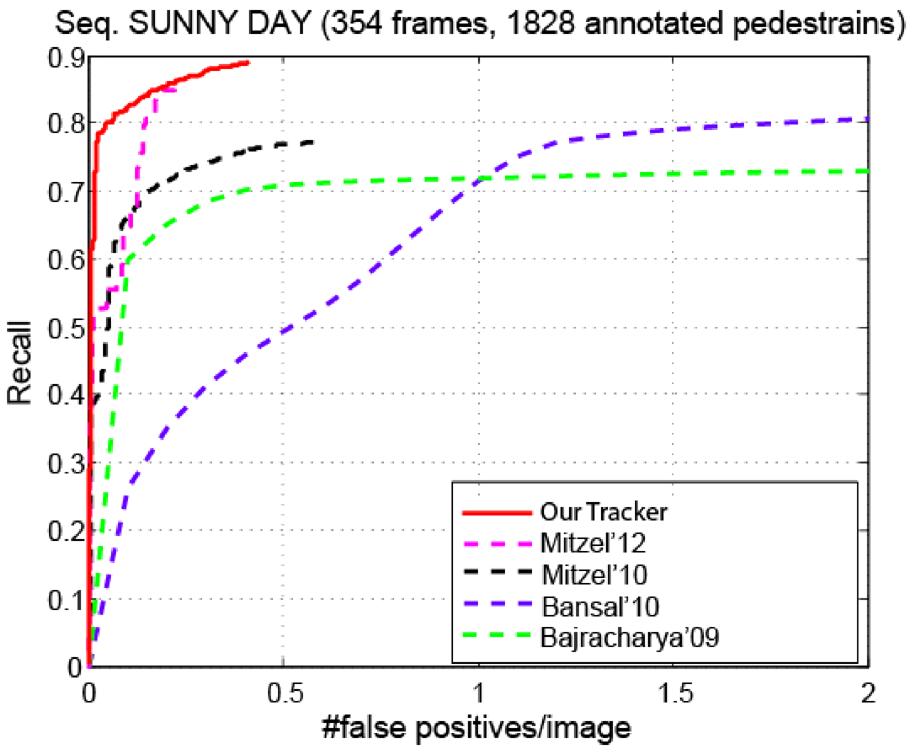


- Tracking-*before*-detection pipeline
  - Tracking fully based on ICP, detector only for verification
  - Build up 3D object models online
  - Detect carried items by comparing with 3D person model

[D. Mitzel, B. Leibe, ECCV'12]

Quantitative Tracking Performance



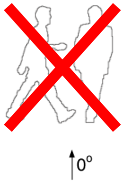


- **Results on ETH Pedestrians**
  - **Considerably improved robustness over tracking-by-detection**
  - **GCTs improve over plain ICP, enable more detailed analysis**

⇒ ***New standard component to build upon***

[D. Mitzel, B. Leibe, ECCV'12; T. Baumgartner, D. Mitzel, B. Leibe, CVPR'13]

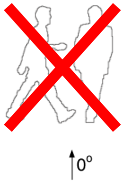
Mobile Tracking in Densely Populated Settings



(Tracking based on stereo depth only, no detector verification)

[D. Mitzel, B. Leibe, ECCV'12]

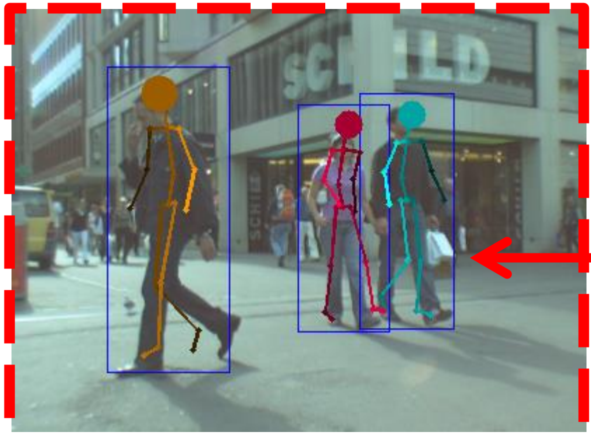
Mobile Tracking in Densely Populated Settings



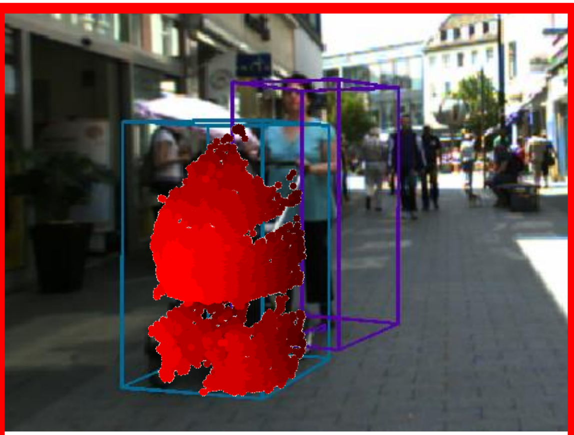
(Tracking based on stereo depth only, no detector verification)

[D. Mitzel, B. Leibe, ECCV'12]

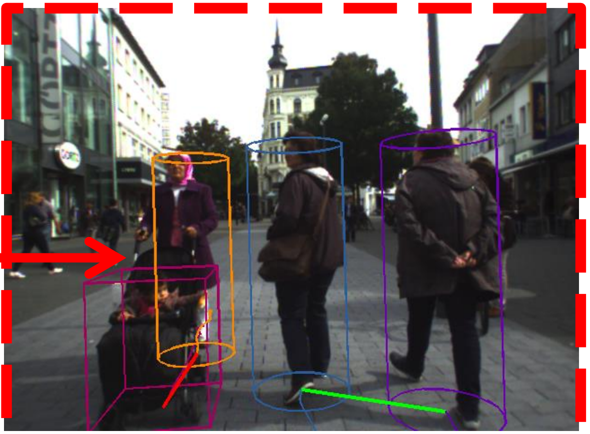
Towards More Detailed Analysis...



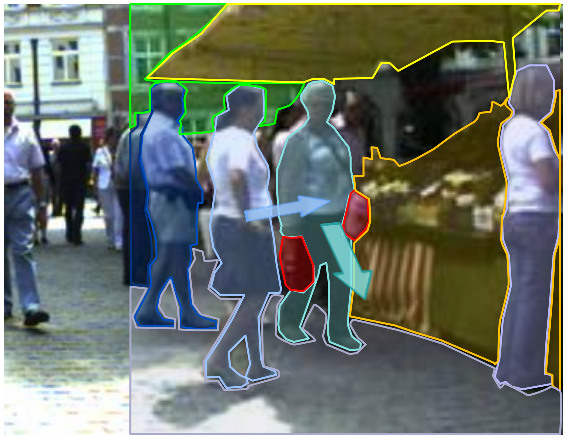
...of people



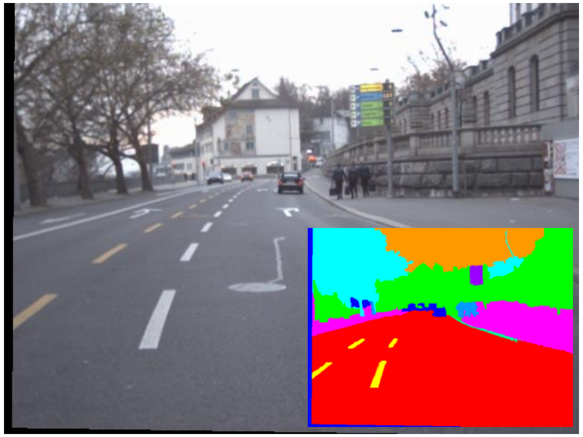
...of objects



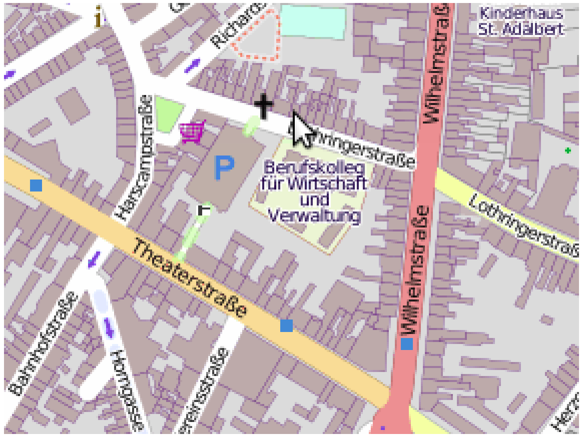
...of interactions



...of social behaviors



...of the environment



...of the surroundings

Conclusions

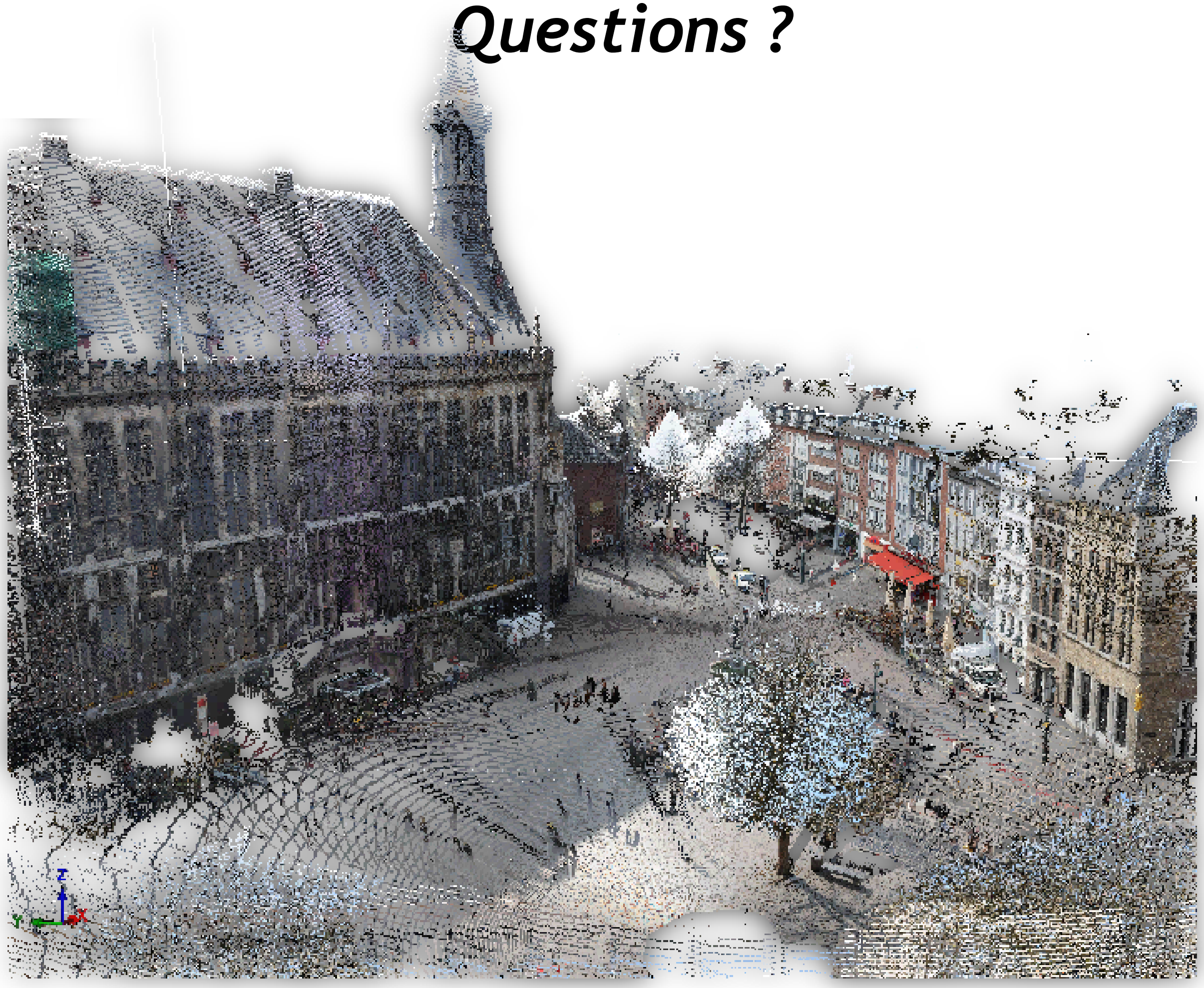
- Tracking for Dynamic Scene Understanding
  - Revisited the goals of using recognition for this
  - Tried to generalize what we have achieved so far

  ⇒ Limits: Tracking-by-detection not scalable to many categories

  ⇒ Limits: Making good predictions is still an elusive goal
- To make progress, we need a more detailed analysis
  - Of people
  - Of objects
  - Of interactions and social behaviors
  - Of the semantics of the environment
- Proposed starting point for such an analysis
  - Approach for tracking arbitrary objects
  - Object-centric representation for partial 3D shape analysis (GCT)

Thank you very much!

Questions ?



Dennis Mitzel

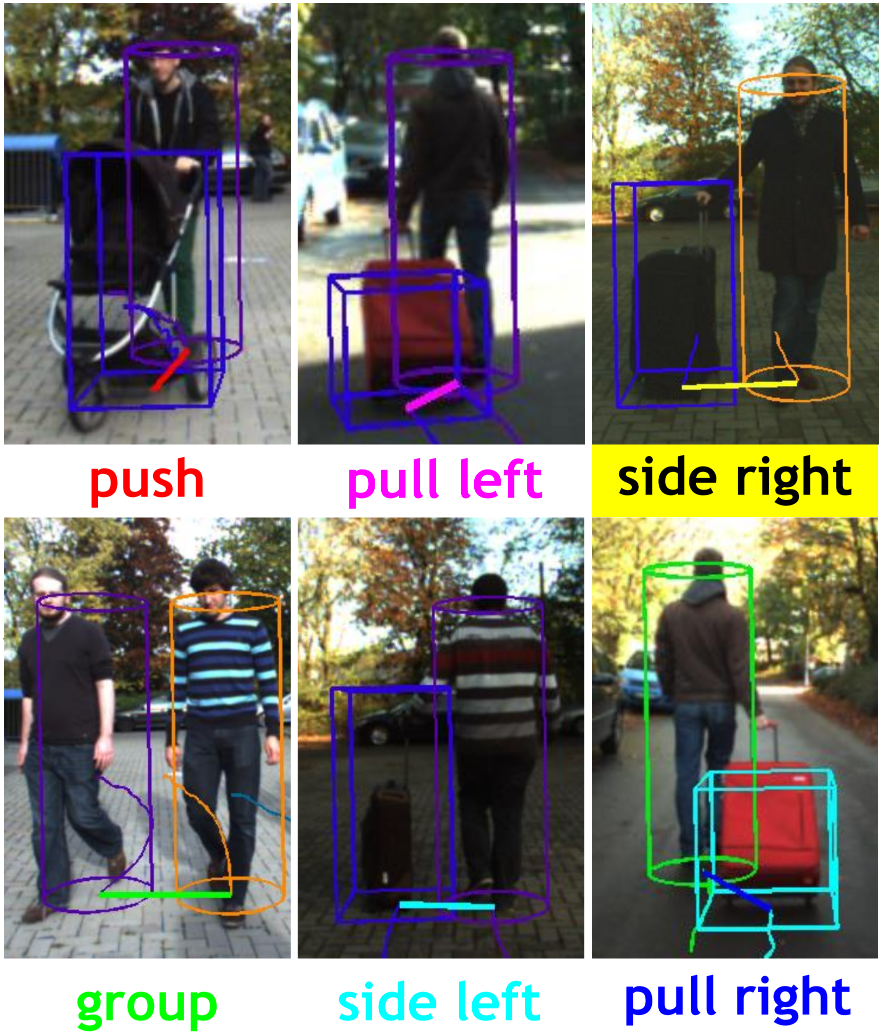
http://www.vision.rwth-aachen.de/

European Research Council
Established by the European Commission

ERC StG 307432
CV-SUPER

New RWTH Interaction Dataset

- 325 video sequences
  - Stereo camera setup
  - More than 15k frame pairs
  - 153 training seq. / 172 test
- Annotations:
  - Segmented 3D point clouds
  - 6 + 1 object classes (*person*, *stroller*, *2-wheel bag* *4-wheel bag*, *walking aid*, *autonomously moving*, *noise*)
  - 6 + 1 interaction classes (*push*, *pull left*, *side left*, *pull right*, *side right*, *group*, *none*)



[T. Baumgartner, D. Mitzel, B. Leibe, CVPR'13]