Taking Multi-Object Tracking to the Next Level: People, Unknown Objects, and Carried Items

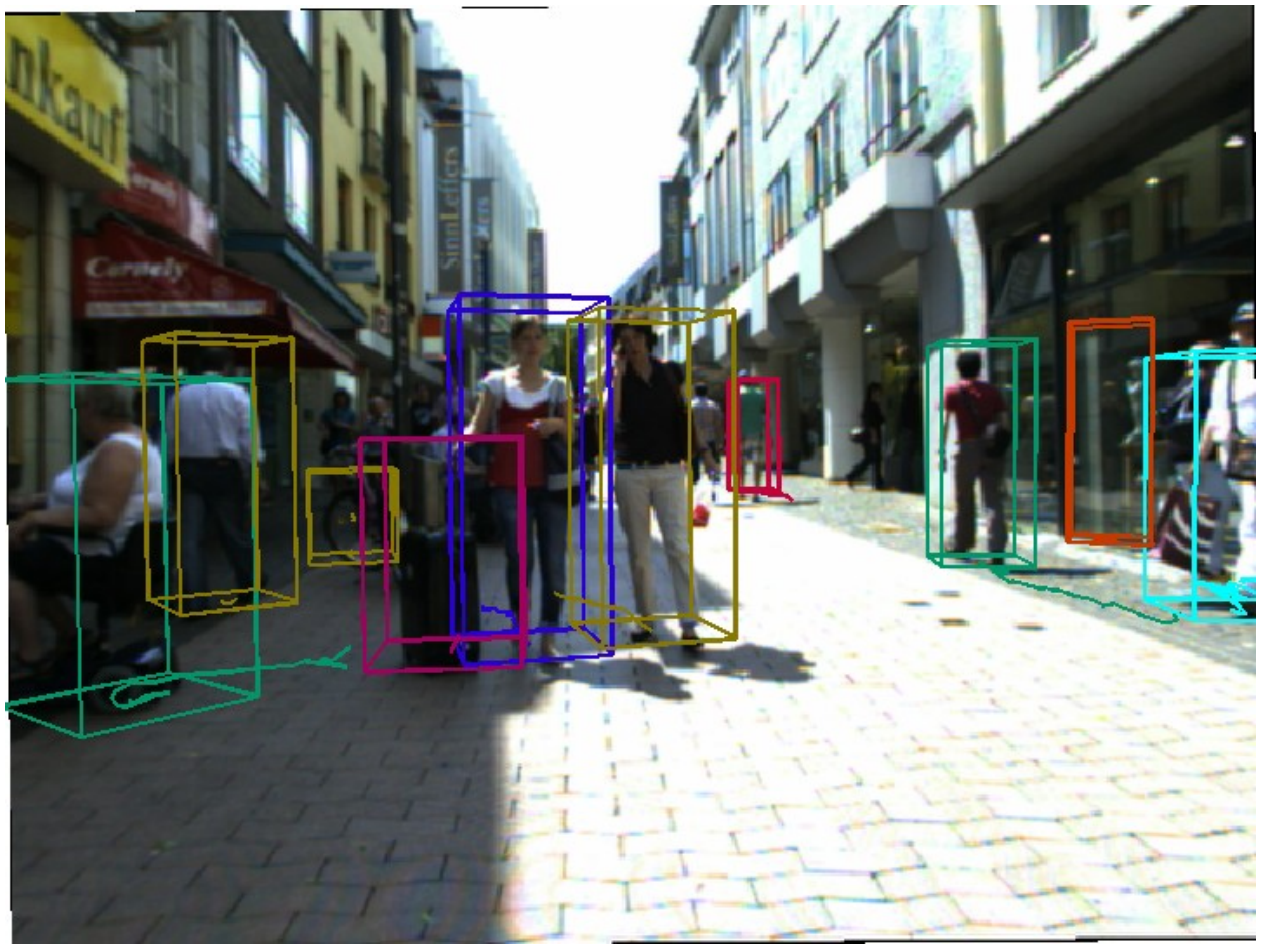
Review by Igor Bogoslavskyi

Department Of Computer Science
Albert Ludwigs University Freiburg
e-mail: bogoslai@informatik.uni-freiburg.de

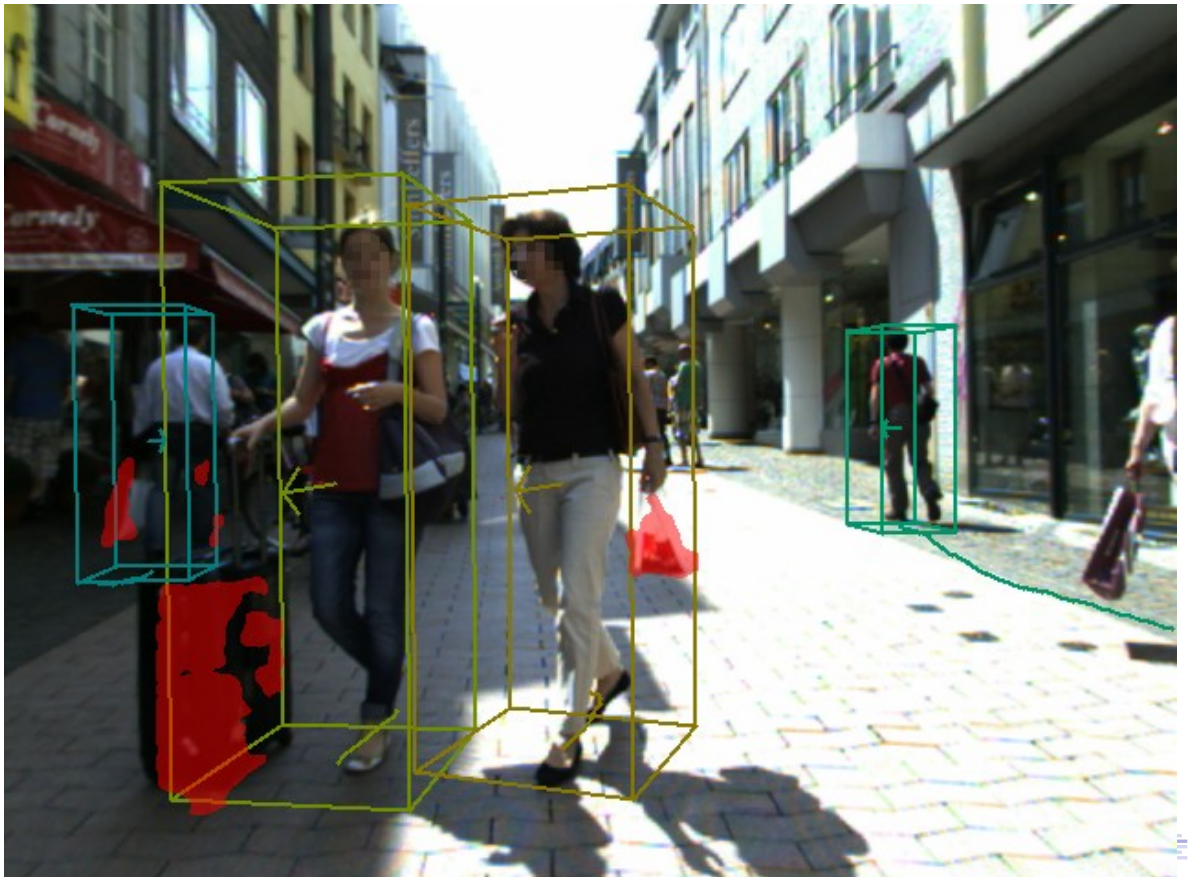
2.02.2013

Motivation

- Tracking objects in the moving scene is an important task in mobile robotics



- Previously only *tracking-by-detection* approaches. These need pre-trained detector models.
- It is important to recognize and track other objects in peoples' surroundings.
- Methods that can detect and track also novel object types and learn models for them on-line are needed.

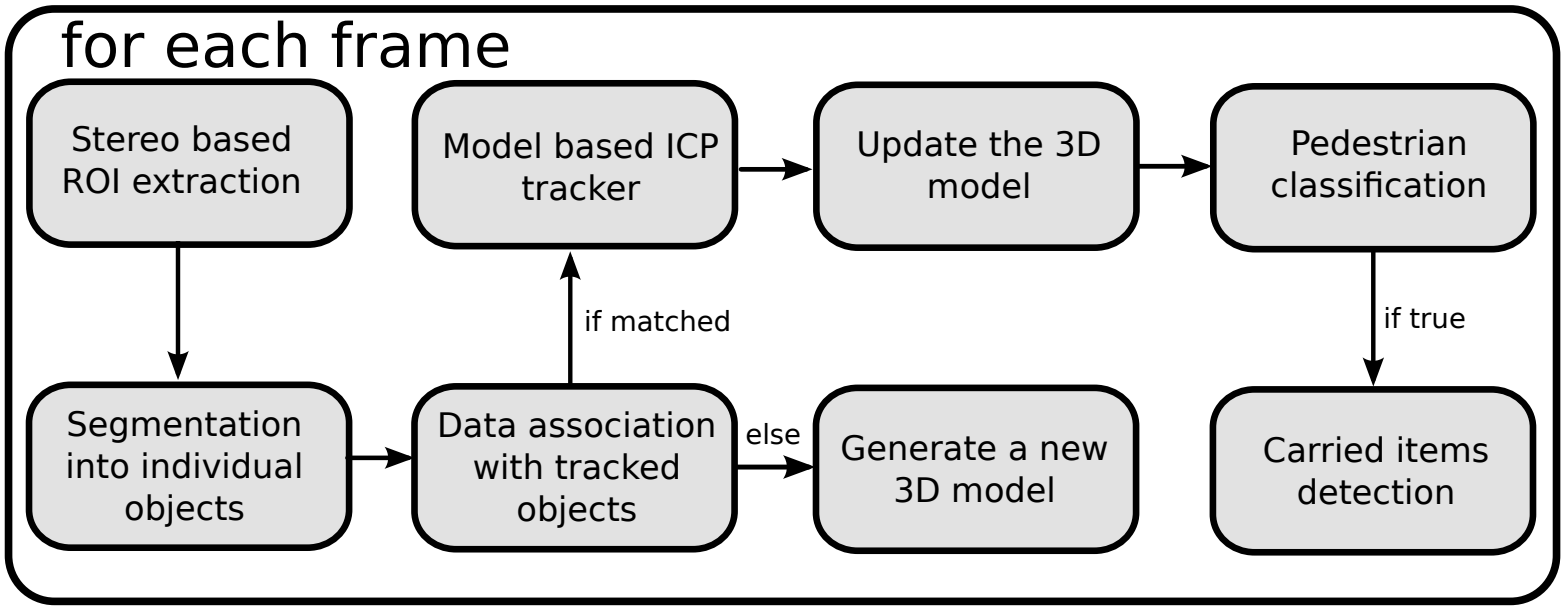


Problem Description

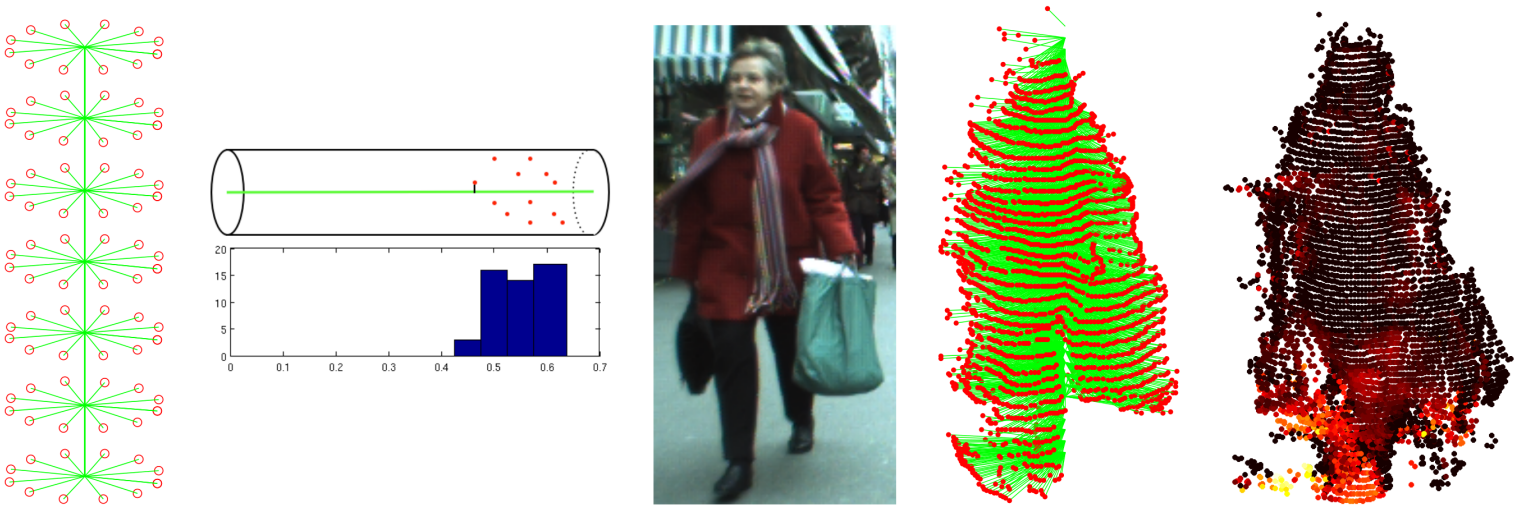
- The problem of detecting novel objects is not trivial.
- To do that one has to answer an even harder question - *what is an object*.
- This itself involves segmenting object from the video stream input.

Tracking-Before-Detection Approach

- Noisy stereo estimates are used to extract regions-of-interest (ROIs) in the input images and to segment them into candidate objects.
- Each ROI is then tracked independently in 3D.
- Each segmented object is then sent to an object classifier for classification into pedestrians and other objects.



3D Model Representation

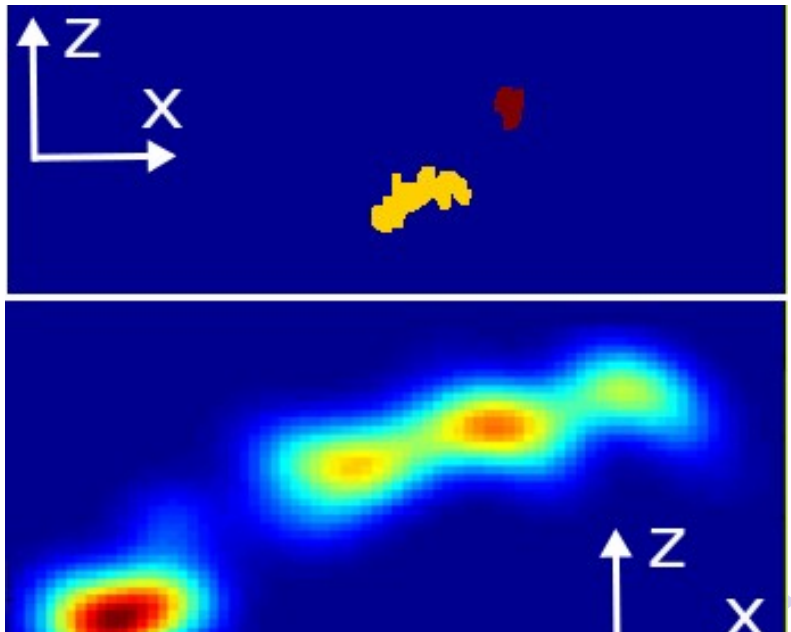


(a) (b) (c) (d) (e)

- The objects are represented by novel *Generalized Christmas Tree* (GCT) model.
- The model is composed of a vertical axis and several layers of equally spaced horizontal rays.
- Along each ray, 3D points are stored in distance histogram.
- Representation adapts to the shape of tracked objects allowing to use median points as well as variances for better visualization.

ROI Extraction and Segmentation

- Given depth image generate a set of ROIs for potential objects.
  - The 3D points within 2m height corridor are projected onto the ground plane.
  - A 2D histogram of these points is taken
  - The histogram bins are thresholded for removing noise.
  - Resulting bins are grouped into connected components (image, top)
- The connected components are found by segmenting a smoothed version of the original histogram via Quick Shift algorithm.



Quick Shift

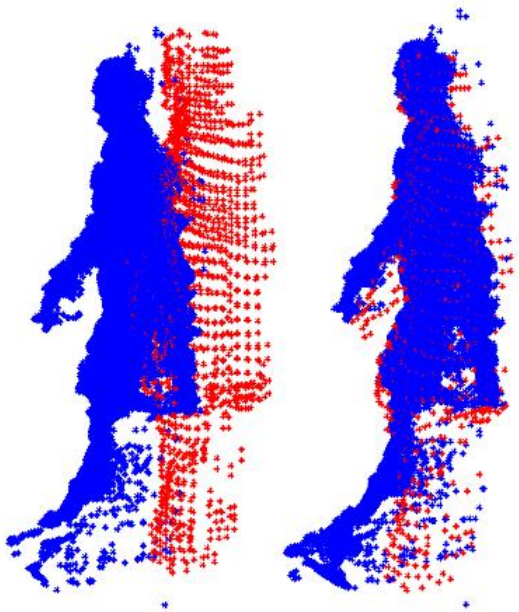
- Quick Shift finds the modes of density by shifting each point to the nearest neighbor with higher density value.
- Quick Shift is performed for every point in smoothed histogram.
- As a result we get a segmentation of the ROIs into individual objects.

Data Association

- We need to associate ROIs with existing tracks.
- This is done by matching them to each track's ROI from previous frame.
- ROIs are assumed to match if the over-union of their ground projection footprints is over 50%.
- We start a new track for all ROIs that cannot be associated.

ICP Tracking and Model Update

- Adapted version of ICP is used for aligning the model from previous step with the 3D points from overlaping ROI.
  - The red points in (a) are the 3D points from the GCT model.
  - The blue points are the 3D points from the overlapping ROI.
  - We obtain rotation and translation for moving the GCT model via ICP.
  - The last step is to update the registered GCT model with new 3D points.



(a) (b)

Object Classification and Tracking

- When the newly generated track comes into detection range, the ROI is passed to pedestrian detector for person/non-person classification.
- This is done by evaluating the detector only in a small region around the back-projected segmented 3D points.
- If the object is then continuously tracked no further classification is performed.
- To deal with occlusions the tracklet is only terminated if no ROI can be associated to a tracked object for t_{term} frames.

Pedestrian Model

- Through the tracking process the 3D model of each tracked person is constantly refined and can afterwards be used to analyze its shape and volume in detail.
- The idea is to compare the on-line model to a learned statistical shape template of pedestrians in order to detect deviations, that cannot be explained by variation in GCT volume.
- Pedestrian Model:
  - The model was learned by collecting a training set of 12 GCT models of pedestrians moving in different directions over a duration of 15-20 frames from a separate training sequence.

Carried Item Segmentation

- Carried item segmentation is based on Conditional Random Fields.

$$E(y) = \sum_{i \in V} \psi_i(y_i) + \sum_{(i,j) \in E} \psi_{ij}(y_i, y_j)$$

- Unary potentials:

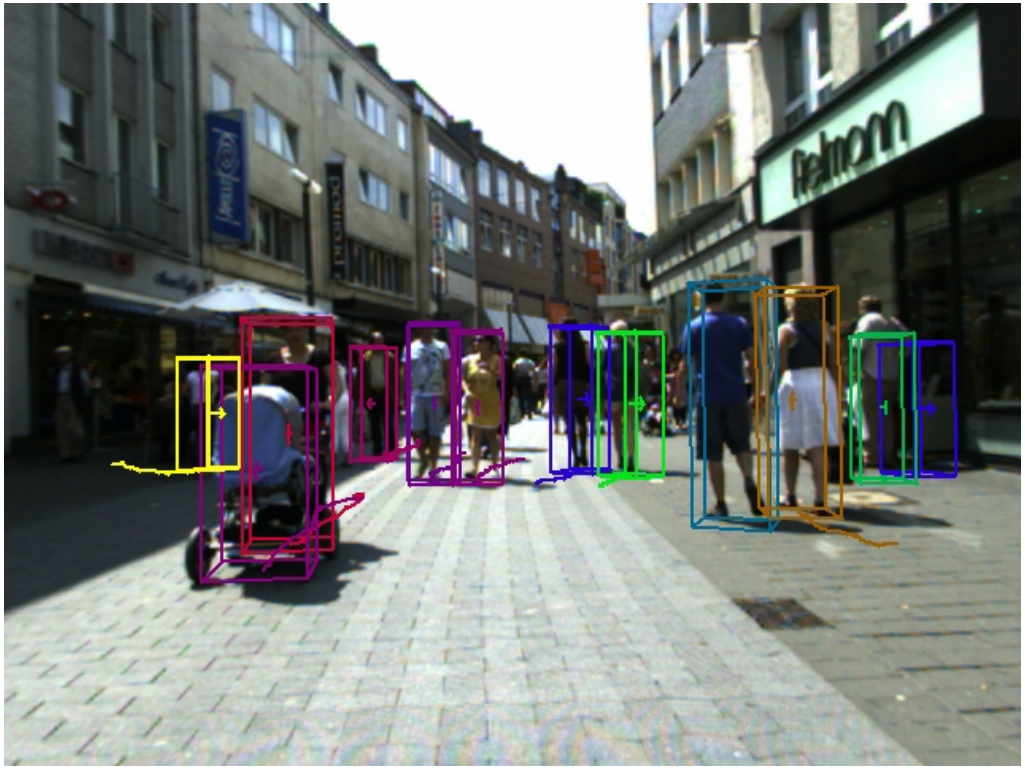
$$\psi_i(y_i) = -log(p(y_i|r_i))$$

  - $p(y_i|r_i) = bhatta(r_{i,object}, r_{i,model})s(r_{i,object})$, where:
    - $bhatta(*)$ is the Bhattacharyya distance between the distance histograms of the on-line tracked and learned model rays.
    - $s(*)$ is a sigmoid weighting function applied to the component of ray distance $r_{i,object}$ that is orthogonal to the walking direction. Serves as a prior that carried items are usually not in the leg area.

- For pairwise potentials a contrast-sensitive Potts model based on image colors was used:

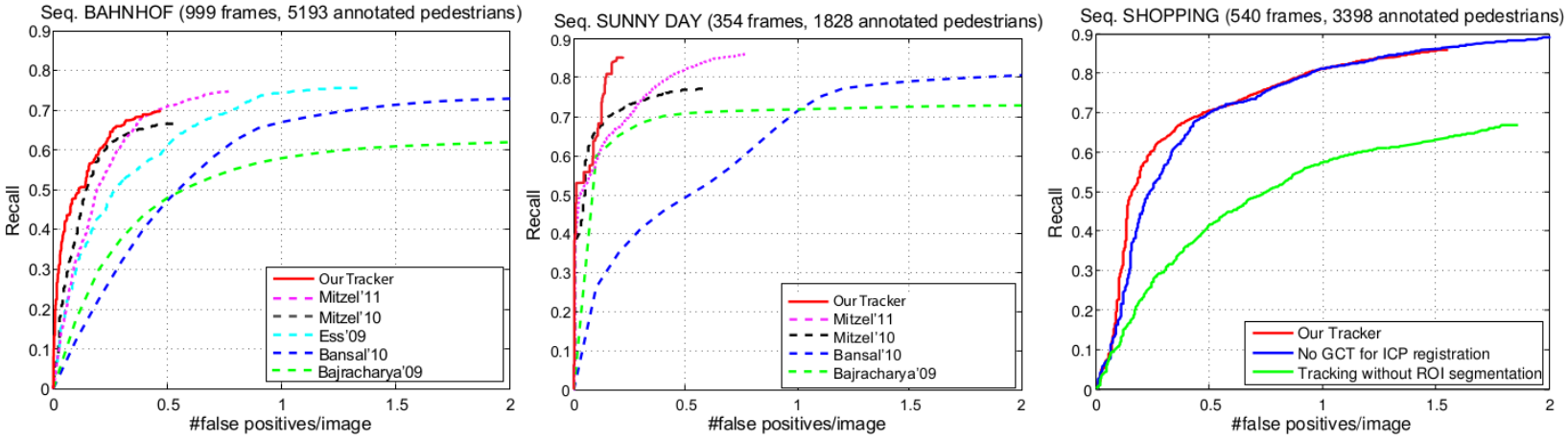
$$\psi_{ij}(y_i, y_j) = \theta_{i,j} \exp(-\beta ||x_i - x_j||^2) \delta(y_i \neq y_j)$$

Datasets



- Three different datasets captured from a stereo camera setup were used:
  - BAHNHOF: 999 frames with 5193 pedestrian annotations.
  - SUNNY DAY: 354 frames with 1867 pedestrian annotations.
  - SHOPPING: over 540 frames with 3398 pedestrians annotations.

Pedestrian Tracking Performance



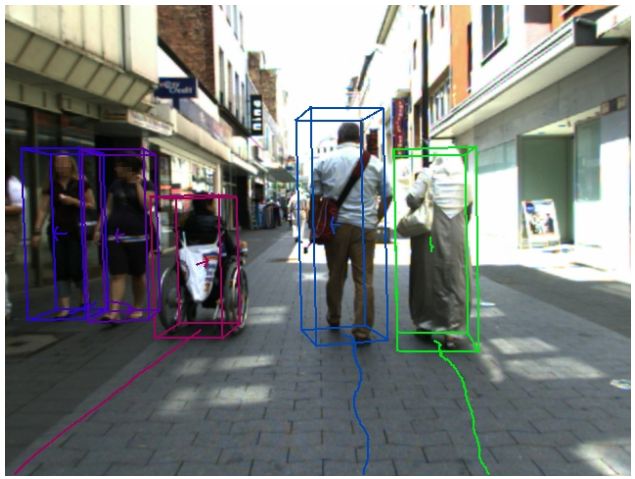
- In every frame we measure the intersection-over-union of tracked person bounding boxes and annotations.
- Detections with an overlap larger than 0.5 are accepted as correct.
- The results are presented in terms of recall vs. false positives per image (fppi).

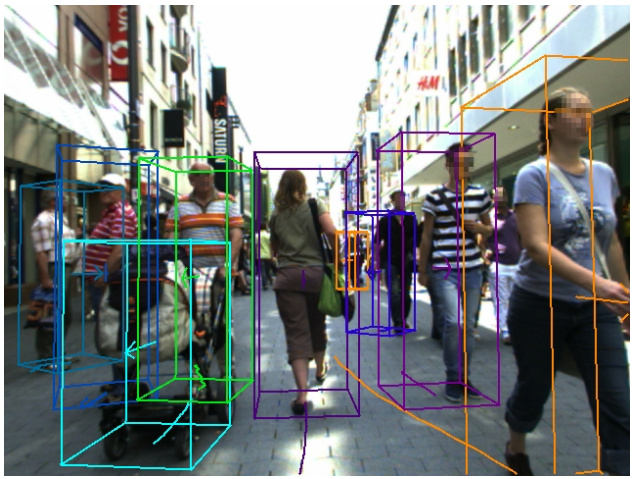
Object Tracking Performance

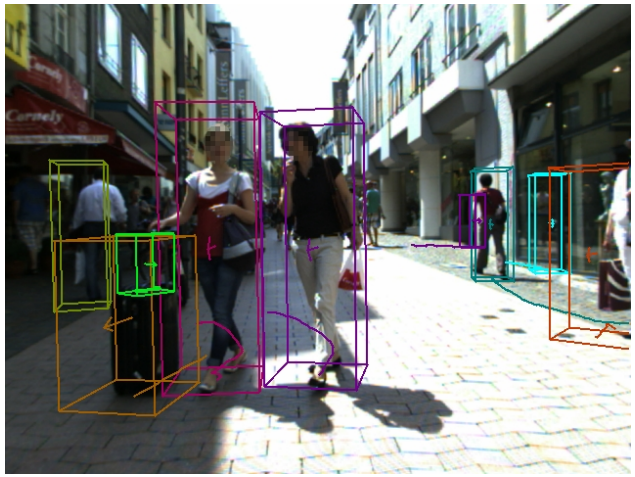
Object types	mostly tracked ($> 80\%$)	nearly tracked ($> 50\%$)	not tracked ($< 20\%$)
Child stroller	5	0	0
Walking aid	0	1	0
Suitcase	1	1	0
Wheelchair	2	0	0
Garbage bin	6	1	0
Bicycle	2	3	4
Advertising rack	0	2	3
Ticket dispenser	1	1	0

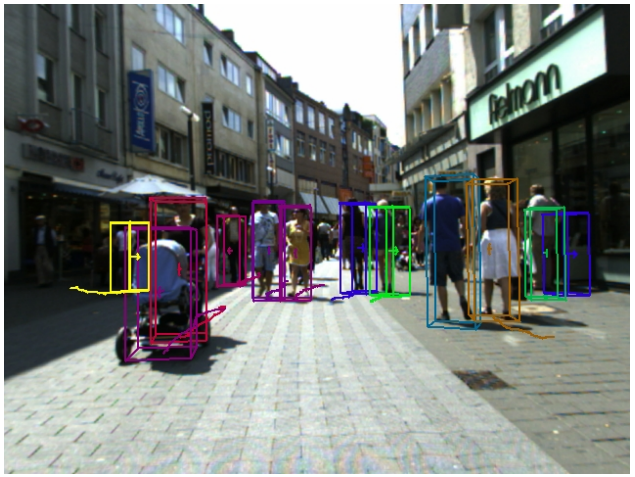
- For assessing the performance of our approach for tracking unknown objects, we have processed 6060 frames of video material, which contained the unknown objects listed in Tab. 8

Object Tracking Examples



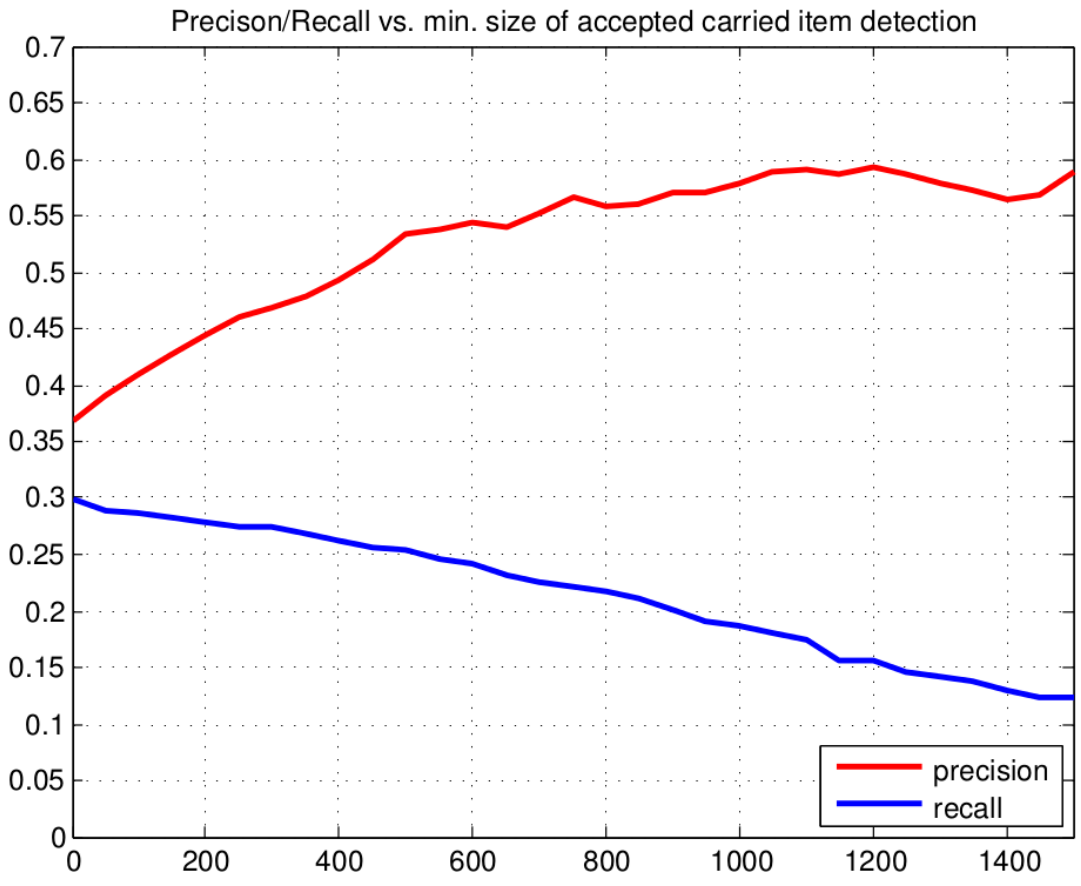




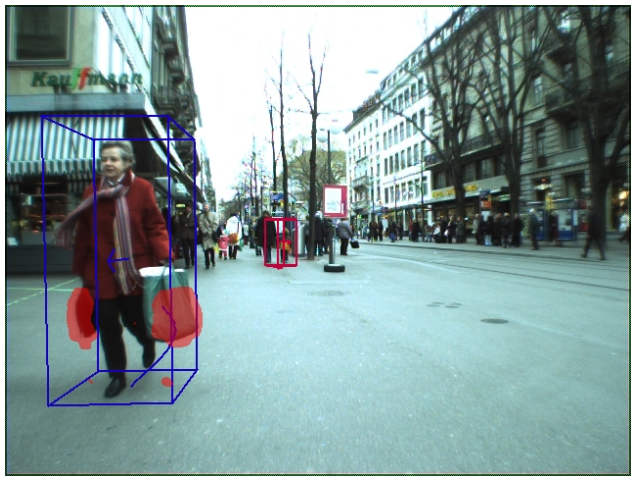


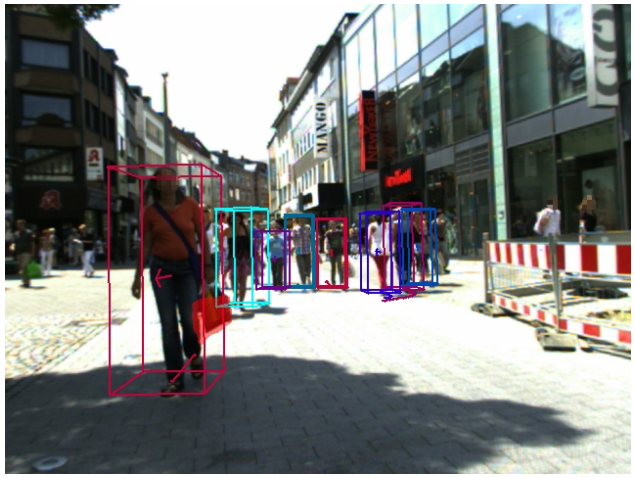
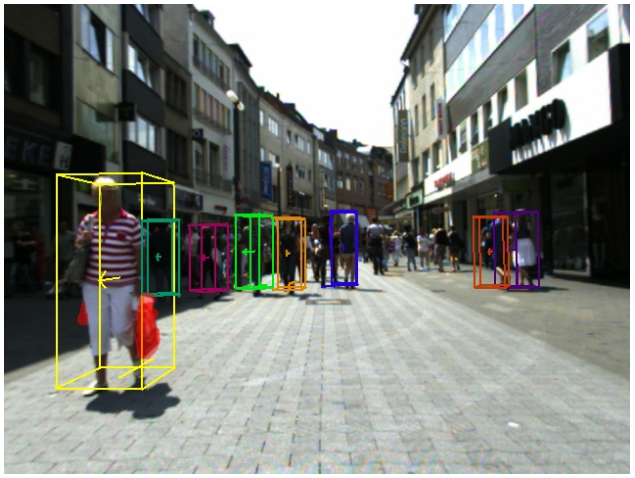
Carried Item Detection Performance

- Pixel-wise comparison of our hypothesized segmentation results with the labeled data, resulting in the precision/recall plot.



Carried Item Detection Examples



Conclusion

- The approach is based on tracking-before-detection paradigm as opposed to older detection-before-tracking one.
- This leads to the possibility to detect and track unknown objects.
- The presented novel 3D model allows not only achieving state-of-the-art tracking performance but also analyzing the shape of tracked person in more detail while running the system on-line.
- This allows to hypothesize the parts of the shape that likely to be carried items by comparing the shape with a learned probabilistic shape.

Questions?