Towards Accurate 3D Person Detection and Localization from RGB-D in Cluttered Environments

Timm Linder[1], Dennis Griesser[2], Narunas Vaskevicius[3] and Kai O. Arras[3]

***Abstract*— We focus on the problem of accurately detecting and localizing 3D centroids of persons in RGB-D scenes with frequent heavy occlusions, as often encountered in industrial and service robotics use-cases. While recently, enormous progress has been made in 2D object detection, which is often evaluated in terms of bounding box overlap in image space, robotics systems often rely on metric 3D world coordinates for applications such as human tracking across sensor boundaries, socially aware motion planning or safety and collision avoidance. Starting with a state-of-the-art 2D single-stage detector, we examine how we can robustly lift the coordinates into 3D to outperform the state-of-the-art in RGB-D person detection at 50 frames per second. Evaluation on our Kinect v2 dataset from an intralogistics warehouse indicates that there might be better intermediate representations for this purpose than 2D bounding boxes, such as instance segmentation masks or keypoint estimates. As an alternative strategy, we also compare our method against a recently proposed bottom-up 3D human pose estimation approach. We find that our 2D top-down person detector achieves higher maximum recall, while the bottom-up 3D human pose estimation method can reach higher precision.**

I. INTRODUCTION

Robust tracking of surrounding humans is important for vehicles in intralogistics to operate safely and efficiently. This requires robust person detectors that output their detections in metric 3D space [1]. We focus on person centroids, as they are easier to annotate than full 3D bounding boxes and sufficient for many use-cases where no direct interaction between the robot and the person occurs, which would make 3D joint keypoints a better choice. Time-of-flight RGB-D sensors like the Kinect v2 are well-suited for many indoor environments as they are cheap, provide better depth estimates than comparable stereo camera systems, and denser information than e. g. lidar or 2D laser.

Computer vision has made significant advances in 2D image-based object detection using modern two-stage detectors, e. g. based on Faster R-CNN [2]–[4]. Variants thereof have also been applied to person detection in RGB [5] and RGB-D data [6]–[8]. On the other hand, single-stage methods like SSD [9] or YOLO [10] have become significantly more robust in their later versions [11], [12] while remaining real-time capable at high frame rates. This makes them attractive for mobile robotics applications with limited computational resources. The challenge when applying these methods to RGB-D data is how to fully exploit the additional depth modality, while still being able to profit from pre-training on large-scale annotated RGB datasets. Most so-called 2.5D approaches [13] start with feature representations in 2D color/depth image space [8], [13]–[17] in order to benefit from the dense color image to achieve high recall, while only few methods begin with a 3D representations of the point cloud [7]. More recently, real-time RGB(-D) 3D human pose estimation methods have been proposed [17].

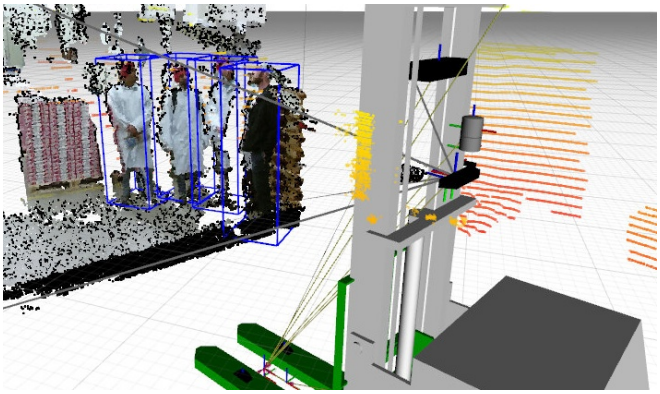


Fig. 1: We want to robustly detect and localize 3D person *centroids* in RGB-D data from real-world robotics scenarios (bounding boxes are just for illustration).

II. DATASET

Existing available Kinect v2 datasets for person detection have been recorded in indoor and outdoor environments at a university campus [6], [18] or a hospital [7]. Some lack person annotations when depth is not available [6], or are destined for multi-class detection involving persons with walking aids [7]. To obtain a better estimate of detection accuracy more relevant to our target domain of intralogistics, we recorded our own dataset using a Kinect v2 sensor mounted horizontally at 1.5 meter height on an AGV in driving direction. The dataset consists of a sequence recorded in an actual warehouse, spanning three days, and a shorter 10-minute sequence from a robotics lab equipped with typical warehouse shelves and pallet trucks. Several people are wearing protective clothing as often found in the food industry, or reflective safety vests. We thoroughly annotated selected parts of both sequences with 2D bounding boxes in around 1200 image frames, and 3 minutes of 3D centroid trajectories. Additional sensors such as lidar were used to aid in the annotation process.

[1]Timm Linder is with Robert Bosch GmbH, Corporate Research, Stuttgart, Germany and with the University of Freiburg, Germany. first.last@de.bosch.com

[2]Dennis Griesser is with HTWG Hochschule Konstanz, Germany. first.last@htwg-konstanz.de

[3]Narunas Vaskevicius and Kai O. Arras are with Robert Bosch GmbH, Corporate Research, Stuttgart, Germany.

Average precision at threshold:	0.25m	0.50m	Hz
MobilityAids (RGB-CNN) [7]	0.580	0.717	15
YOLO v3 [12], naive depth	0.718	0.809	50
+ instance masks [4]	0.738	0.829	4
Mask R-CNN [4] + masks	0.795	0.888	4
RGB-D Pose 3D [17]	0.751	0.802	2

Fig. 2: Initial quantitative results for person detection on our annotated 3D test set (without fine-tuning) at two thresholds.

Fig. 3: Detections of our YOLO v3-based method in blue, Mobility Aids in magenta, RGB-D Pose 3D in red. Upper row shows false alarms of [7], [17], lower row a false depth estimate of [17], which we avoid via instance segmentation.

III. EXPERIMENTS & INITIAL RESULTS

As a simple but efficient baseline (running at 50 Hz on our AGV platform), we use a YOLO v3 RGB detector [12] trained on MS COCO [19]. For our experiments, we refrain from fine-tuning on our data since we want to analyze how well the examined methods transfer to new scenarios. For evaluation in 3D world coordinates, we perform naive lifting into 3D by computing the median of the depth values in the 2D bounding box. Without otherwise fusing depth information as e. g. proposed in [6], [8], this already surpasses the performance of [7], probably due to a better-performing 2D detector. We integrate 2D instance segmentation masks, obtained offline [4] but potentially real-time [20], for more robust depth estimation under heavy occlusion, boosting AP by 2 percent points as shown in Figure 2.

As segmentation masks seem beneficial, we were wondering if articulated 3D human pose estimation would provide similar or better results for our detection task. After tuning parameters to improve the recall of [17], we found out that in fact their approach can reach higher maximum precision and AP at 0.25m evaluation threshold, while our YOLO-based method can attain slightly higher recall and outperforms [17] in AP at 0.5m; more details on our experiments in our poster.

Qualitatively, the main challenges that all examined methods struggle with on our dataset are heavy occlusion (by persons, AGVs or static warehouse objects), variations in lighting and partially low contrast, motion blur, and reflections in glass and metallic surfaces, which could partly be improved upon by incorporating depth at the detection stage. To increase recall under heavy occlusion, we think that synthetic occlusion augmentation [21] could help. However, we also observe that reflective safety vests occasionally lead to entirely wrong depth readings of the Kinect v2 sensor. Figure 3 shows example detections of the evaluated methods.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732737 (ILIAD).

REFERENCES

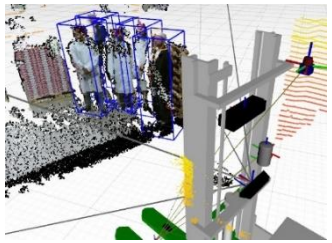
[1] T. Linder, S. Breuers, B. Leibe, and K. O. Arras, "On multi-modal people tracking from mobile platforms in very crowded and dynamic environments," in *ICRA*, 2016.
[2] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.
[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *PAMI*, vol. 39, no. 6, June 2017.
[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.
[5] L. Zhang, L. Lin, X. Liang, and K. He, "Is Faster R-CNN doing well for pedestrian detection?" in *ECCV*, 2016.
[6] O. Mees, A. Eitel, and W. Burgard, "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments," in *IROS*, 2016.
[7] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard, "Deep detection of people and their mobility aids for a hospital robot," in *ECMR*, 2017.
[8] J. Guerry, B. L. Saux, and D. Filliat, ""Look at this one": Detection sharing between modality-independent classifiers for robotic discovery of people," in *ECMR*, 2017.
[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.
[10] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.
[11] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, arXiv:1701.06659.
[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, arXiv:1804.02767.
[13] Z. Deng and L. J. Latecki, "Amodal detection of 3D objects: Inferring 3D bounding boxes from 2D ones in RGB-D images," in *CVPR*, 2017.
[14] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *IROS*, 2015.
[15] Q. Luo, H. Ma, Y. Wang, L. Tang, and R. Xiong, "3D-SSD: Learning hierarchical features from RGB-D images for amodal 3D object detection," 2017, arXiv:1711.00238.
[16] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *CVPR*, 2018.
[17] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D human pose estimation in RGBD images for robotic task learning," in *ICRA*, 2018.
[18] N. Wojke, R. Memmesheimer, and D. Paulus, "Joint operator detection and tracking for person following from mobile platforms," in *Proc. International Conference on Information Fusion (FUSION)*, 2017.
[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.
[20] J. Uhrig, E. Rehder, B. Fröhlich, U. Franke, and T. Brox, "Box2Pix: Single-shot instance segmentation by assigning pixels to object boxes," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018.
[21] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3D human pose estimation to occlusion?" in *IROS - Workshop on Robotic Co-workers 4.0: Human Safety and Comfort in Human-Robot Interactive Social Environments*, 2018.

Towards Accurate 3D Person Detection and Localization from RGB-D in Cluttered Environments

Timm Linder, Dennis Griesser, Narunas Vaskevicius and Kai O. Arras

Goal

- **Real-time** detection of **3D person centroids** from onboard RGB-D sensor (Kinect v2) in intralogistics



Research questions

- Do existing 2D real-time detection methods (e.g. YOLOv3 pretrained on COCO) **generalize well** to intralogistics use-cases (protective clothing)?
- How to transfer RGB knowledge to depth modality?
- Are 2D bounding boxes a **good intermediate representation** if we ultimately want to estimate 3D centroids?
- Does articulated **3D human pose estimation** outperform a purely detection-based approach?

Dataset

- Recorded in different **warehouse** environments from **mobile forklift** over several days
- Around 1100 frames annotated with **2D boxes**
- 3 minutes of **3D/2.5D birds-eye view centroid** trajectory annotations so far, working on more

Method

- Available RGB-D datasets are limited in size, thus training from scratch leads to inferior results
- **Compositional, modular** approach
  - Leverage existing YOLO v3 RGB detector + large-scale RGB datasets like COCO
  - Combine with depth-based detector/classifier to improve accuracy under challenging conditions
- **Variant 1: Naïve depth estimation** via median of depth values in 2D bbox
- **Variant 2:** Use **instance segmentation mask** from Mask R-CNN to focus region for depth estimation
- **Variant 3:** Mask R-CNN instead of YOLO v3 for detection

Results

- Comparison against **two strong recent baselines**
- **No fine-tuning** of any method on our dataset to examine how well they generalize.

AP @ threshold:	0.25m	0.50m	Hz
Mobility Aids [1]	0.580	0.717	15
V1: YOLO v3, naïve depth	**0.718**	**0.809**	**50**
V2: + instance masks	0.738	0.829	4
V3: Mask R-CNN + masks	**0.795**	**0.888**	**4**
RGB-D Pose 3D [2]	0.751	0.802	2

Not real-time

Insights: 2D detection

- 2D detection failures in RGB mainly due to **heavy occlusion** (>70%) and **reflections** in glass/metal
  - Some reflections can be filtered out via depth
  - Use training-time occlusion augmentation [3] to improve performance under heavy occlusion?
- **Protective clothing** causes no issues in RGB, but:
- Kinect v2 sensor sometimes yields **wrong raw depth values for reflective safety vests** (3m off!)

Insights: 3D localization

- **Accurate 3D localization is a big issue**. Often fails if there is just light occlusion, because wrong region of depth image is chosen to estimate depth
- **Instance segmentation masks help**
- **Bottom-up 3D human pose estimation** outperforms 2D detector in max. attainable precision, but not AP
- Results indicate we should maybe focus directly on **3D articulated pose estimation**, instead of detection.
- But large-scale multi-person **datasets** for training of pose estimation are missing.

References

[1] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard, *"Deep detection of people and their mobility aids for a hospital robot"* in ECMR, 2017.
[2] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, *"3D human pose estimation in RGBD images for robotic task learning"* in ICRA, 2018.
[3] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, *"How robust is 3D human pose estimation to occlusion?"* in IROS 2018 Workshop on Robotic Co-workers 4.0: Human Safety and Comfort [...].

Towards Accurate 3D Person Detection and Localization from RGB-D in Cluttered Environments

Timm Linder, Dennis Griesser, Narunas Vaskevicius and Kai O. Arras

Fusion of depth and RGB for detection

- **Variant 1**: Separate detector networks for RGB + D
  - Slow (double effort), more modular (separate training on distinct datasets possible)
  - Training via **supervision transfer: distillation loss**
  - Very late fusion at NMS stage ("U-Fusion")
  - YOLO v3 detection results on our validation set, jet colormap depth encoding [4]:

AP @ IoU threshold:	0.5	0.75	[0.5;0.95]
COCO training, no finetuning	0.17	0.14	0.10
Finetuned on InOutDoor [5]	0.40	0.24	0.22
Supervision transfer [6] on MobilityAids dataset [1]	0.38	0.17	0.18
Supervision transfer [6] on smaller InOutDoor dataset [1]	**0.61**	**0.41**	**0.36**

- **Variant 2**: RGB detector network + subsequent depth classifier
  - Can improve precision, but not recall: e.g. to fix false positives due to reflections
  - Can be trained with weak supervision by using RGB classifier as teacher network / supervision transfer

- **Variant 3**: Early fusion of RGB and D (e.g. as extra channel), single detector network
  - Difficult to train in end-to-end fashion due to lack of large-scale RGB-D datasets
  - Would probably give best results in in AP + Hz

Failure cases in 2D image-based detection

reflective vests (depth)

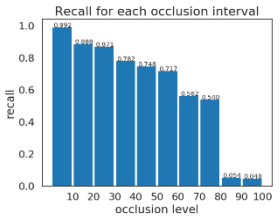
low contrast (RGB)

motion blur (RGB)

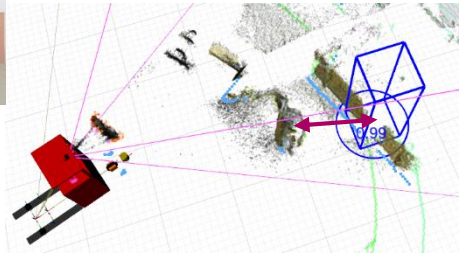
heavy occlusion (both)

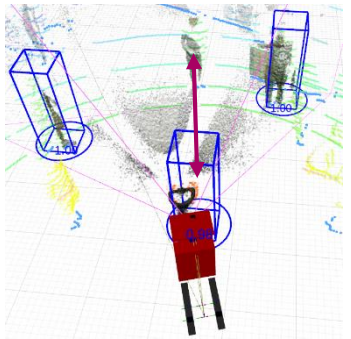
reflection (RGB)

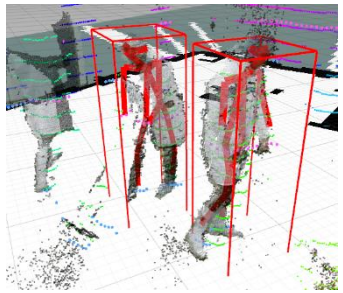
unusual poses (both)



Failure cases in 3D pose estimation







References (continued)

[4] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, *"Multimodal deep learning for robust RGB-D object recognition"* in IROS, 2015.
[5] O. Mees, A. Eitel, and W. Burgard, *"Choosing smartly: Adaptive multimodal fusion for object detection in changing environments"* in IROS, 2016.
[6] S. Gupta, J. Hoffman, J. Malik, *"Cross Modal Distillation for Supervision Transfer"* in CVPR, 2016.

Towards Accurate 3D Person Detection and Localization from RGB-D in Cluttered Environments

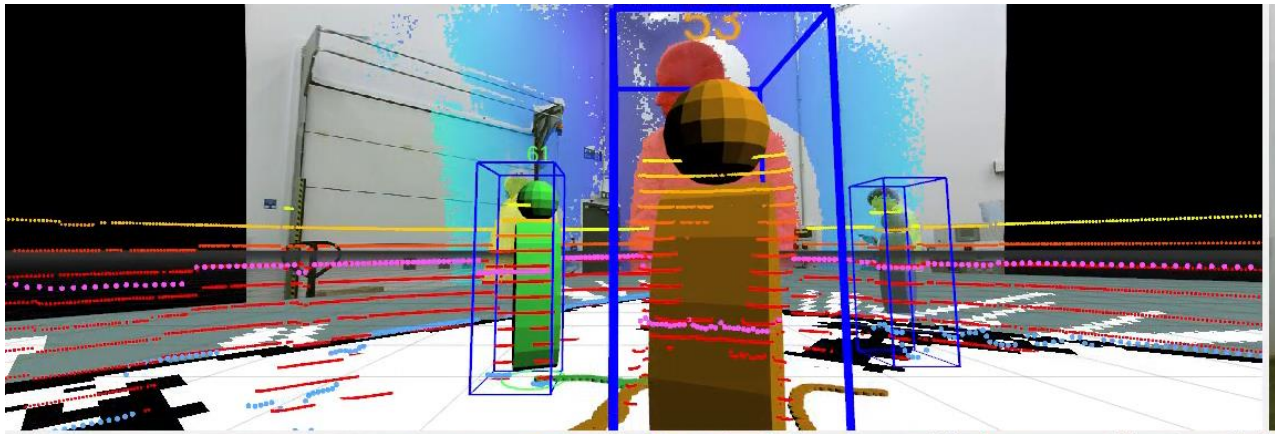
Timm Linder, Dennis Griesser, Narunas Vaskevicius and Kai O. Arras

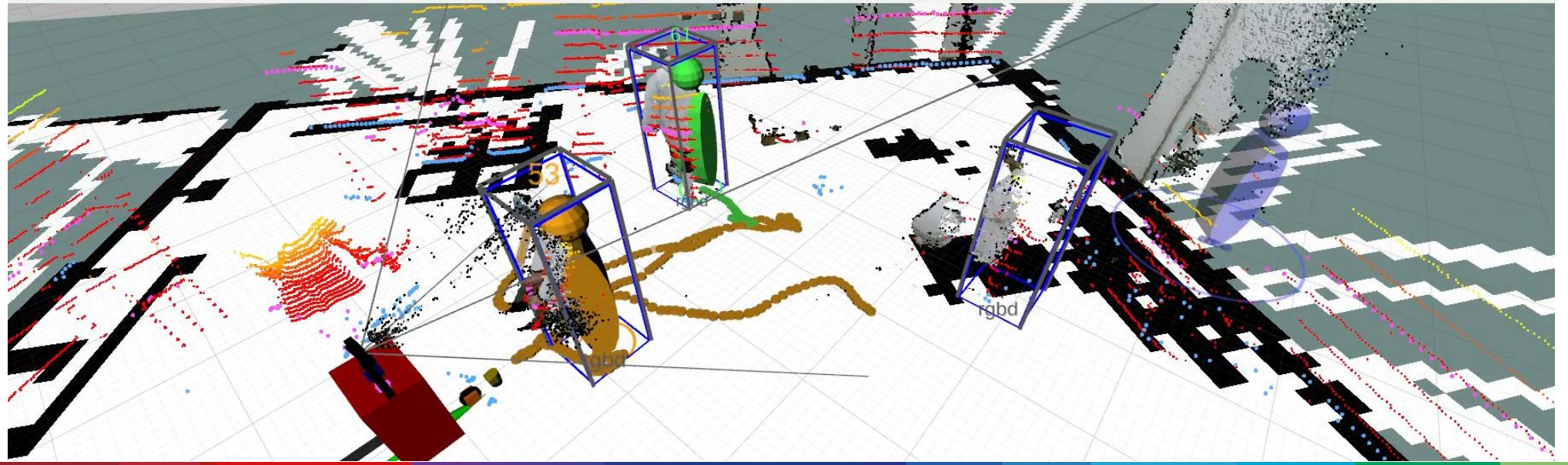
BOSCH

ILIAD

HTWG Hochschule Konstanz Technik, Wirtschaft und Gestaltung

UNI FREIBURG





BOSCH