People Detection with Depth Silhouettes and Convolutional Neural Networks on a Mobile Robot

Florian Spiess
Faculty of Electrical Engineering, University of Applied Sciences Wuerzburg-Schweinfurt, Schweinfurt, Germany
Email: florian.spiess@uni-wuerzburg.de

Lucas Reinhart and Norbert Strobel
Institute of Medical Engineering, University of Applied Sciences Wuerzburg-Schweinfurt, Schweinfurt, Germany
Email: {norbert.strobel, lucas.reinhart}@fhws.de

Dennis Kaiser and Samuel Kounev
Faculty of Mathematics and Computer Science, Julius-Maximilians-University of Wuerzburg, Wuerzburg, Germany
Email: {dennis.kaiser, samuel.kounev}@uni-wuerzburg.de

Tobias Kaupp
University of Applied Sciences Wuerzburg-Schweinfurt, Schweinfurt, Germany
Email: tobias.kaupp@fhws.de

***Abstract*—This paper presents a novel people detection approach for mobile robot applications based on a combination of classical computer vision techniques and a state-of-the-art neural network. Our approach involves an RGB-D camera as an environmental sensor. The depth data is used to extract silhouettes around people. The RGB images are subsequently augmented with this border information before passing it to the neural network. Under challenging lighting conditions, our system was able to outperform the neural network trained on regular RGB data alone by a factor of two.**

***Index Terms*—neural network, mobile robot**

I. Introduction

There is a growing trend to switch from rigid manufacturing and assembly processes to more flexible approaches characterized by seamless, automated manufacturing lines ("smart factories"). Automatically Guided Vehicles (AGVs) are an important component for logistical operations in smart factories, where they are still likely to encounter human workers. In these cases, reliable, real-time person recognition is required such that collisions can be avoided. Another advantage offered by automatic people detection is that AGVs could be enabled to interact with people via a Human Machine Interface (HMI). For safety reasons, people detection needs to be accurate, robust, and has to be performed in real-time using the AGV's limited onboard computing power.

Manuscript received March 2, 2021; revised July 5, 2021.

Traditional image processing techniques for people detection often involve hand-crafted features, e.g., based on Haar wavelets [1]. However in the last decade, detection approaches based on Neural Networks (NNs) have shown superior results in almost all aspects including speed and precision.

For example, a neural network for fusion of RGB with Infrared (IR) data was analyzed in [2]. Starting with individual Convolutional Neural Networks (CNNs) for RGB and IR images, they fuse these CNNs halfway in order to generate multispectral deep features for the Region Proposal Network (RPN). Vandersteegen *et al.* [3], on the other hand, augmented RGB images with thermal images in a pre-processing step. For the augmentation step, two methods were tested. In the first method, one color channel of the RGB image was replaced with the thermal data; in the other approach, the color image was weighted based on the thermal image. A CNN for the fusion of depth images and RGB pictures was designed in [4]. The information was combined in the last layer of the network. The idea of [4] was further expanded in [5]. In this work, several neural networks were created fusing the depth images with the RGB information at different stages. Best results were obtained for mid-level fusion. In [6] the fusion of RGB images and depth images was investigated further using a YOLOv2 network. Better people detection performance was observed when depth images were included for training. According to the authors, depth images enabled the network to learn that people at different distances from the camera have different scales yielding improved detection performance in occluded scenarios as well as in pictures showing groups of people [6]. In [7] a neural

doi: 10.18178/joig.9.4.135-139

network was trained on silhouettes extracted from RGB images. However, the goal of this work was human body pose detection, and no further fusion of the silhouette information with other images was performed.

The focus of this work is on the application of RGB-D cameras for people detection. RGB-D cameras provide color as well as depth information for each pixel in the image.

Our approach is based on a combination of traditional computer vision algorithms and deep-learning methods. The former are used to extract boundary information (silhouettes) around people in the depth frames. They are then used to augment the RGB color images.

The contributions of this paper are:

1) We created a dataset for people detection assuming the perspective of a modern mobile robot platform (EvoRobot, Evocortex GmbH, Nuremberg, Germany). This dataset was labeled and used to train our networks.
2) We demonstrated that a neural network trained using RGB image data, augmented with depth information, performs better compared to working on standard RGB image data, especially in scenarios characterized by poor lighting conditions.
3) We showed that our approach supports a frame rate of 5.56 fps on the Jetson TX2 board which comes with the EvoRobot mobile platform. This suggests that the technique can be applied in practice.

The remainder of the paper is organized as follows: Subsection II-A describes our method in detail. In Subsection II-B, information regarding our dataset is provided. The training of the neural network is described in Subsection II-C. Experimental results can be found in Section III. In the final section, results are discussed and conclusions are drawn.

II. People Detection Using Depth Silhouettes

The underlying rationale of our approach is that silhouettes, outlining human bodies, provide more distinct representations of people by emphasizing their shape. By offering additional features to Convolutional Neural Networks, they should be able to better learn the appearance of people, and, as a result, obtain a better people detection performance.

Extracting the silhouettes from depth images offers the advantage that this border information can also be made available under poor lighting conditions. This enables a better detection compared to pure RGB images in these situations. By design, the silhouettes in our approach are independent of the RGB image and therefore of a person's clothing, the persons themselves, and the environment. Also, silhouettes of people are quite similar to each other, yet they differ significantly from other objects. This also contributes to a better people detection performance within their surrounding. The contour around a person is found during a pre-processing stage. This outline was then integrated into the associated RGB frame and the depth image itself. In the next step, we trained four CNNs one with RGB images only, one with augmented RGB images, one with depth images only, and one with augmented depth images, respectively. However, during our initial experiments, we found that depth-based silhouette augmentation of depth data improved precision only very marginally. As a consequence, we decided to put more emphasis on the augmentation of RGB images.

The output of our image augmentation step is used as input for the YOLOv3 neural network [8]. YOLOv3 is an improved version of the neural network called "You Only Look Once" (YOLO) [9]. The architecture of this network is designed such that bounding boxes and class probabilities are predicted in one evaluation step, which is considered state-of-the-art for fast detection.

A. System Architecture

This section explains our approach in detail. As illustrated in Fig. 1, five pre-processing steps are carried out to obtain an augmented RGB image to be fed into the CNN.

1) In the first step, missing depth pixels (shadow regions) are interpolated. These shadow regions can be seen in the left-most picture of Fig. 1. Missing depth information in depth images results from the sensor not being able to acquire information about the region either due to occlusion or light defusing obstacles. Our strategy to correct for them is as follows [10]. For each pixel in the depth image with an undefined depth value (e.g., a depth value of 0), also referred to as depth shadow, we search through the neighboring data in the corresponding of the missing depth pixel and look on the left and on the right. If for both sides a valid value is found, we substitute the undefined value with the larger of the two. If just a single valid value is available, this one is used. The second picture of Fig. 1 shows the corrected depth image.

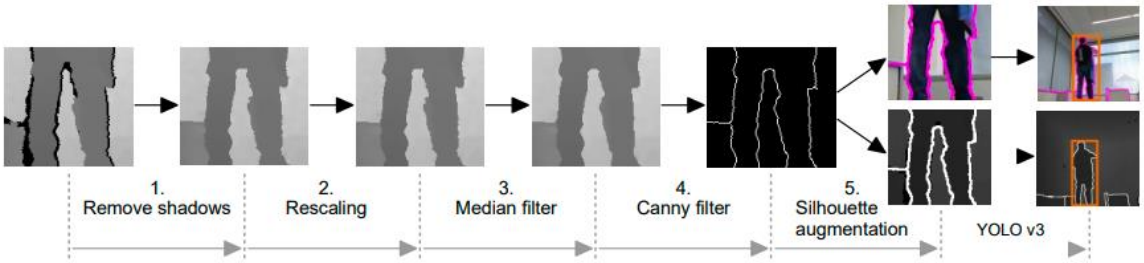

Figure 1. Overview of the proposed method (inference stage): five pre-processing steps are applied to the images before feeding them into a YOLO CNN for people detection. The first four images represent gray-scale representations of depth data with black being close and white being far from the camera, respectively.

2) The depth image is linearly scaled to a range of 0 to 255 according to [11].

$$p' = \frac{p - p_{min}}{p_{max} - p_{min}} \omega_{target} + p'_{min} \quad (1)$$

In this equation, p and p' are the depth values before and after scaling, p_{min} and p_{max} are the minimum and the maximum depth values of the image before scaling, p'_{min} is the lower bound of the new range and ω_{target} is the difference between the upper and the lower bound of the new, respectively. In our case ω_{target}=255 and p'_{min}=0.

3) Median filter is run on the rescaled depth image to remove remaining noise, in particular along the edges. The output is shown in the fourth picture of Fig. 1.

4) The Canny edge filter is applied [12].

5) The resulting contour is merged with either the RGB or the depth image. To this end, we set pixel values along the silhouette in the RGB image to the RGB-value of purple. In the depth images, the silhouette pixels were set to a depth value of 20m. In both cases, our goal was to highlight the silhouette such that it is clearly distinguishable from the background. Parts of the resulting two augmented output images can be seen in the second picture from the right in Fig. 1. On the top, we see the augmented RGB picture, and on the bottom the augmented depth image. After preprocessing, the data is fed into the trained YOLOv3 network for inference. It generates the output shown in the right-most picture of Fig. 1. When processing the depth images with the YOLOv3 network the same depth data was stored in all 3 channels of a RGB image as *uint8* values. This facilitated the use of the same network architecture as for the RGB images.

B. Dataset

The dataset for the training process was taken from the point of view of a mobile robot, i.e. from a camera mounted between 0.10-0.70 m above ground looking upwards. Images were captured using a Xbox Kinect V1 set at a matrix size of 640×480 pixel with depth and color images synchronized in time. The images were then cropped to 570×430 pixels to ensure that they contain only areas were depth and color images overlap.

Our dataset was recorded inside our university buildings. In total we had 16 different settings comprising a total of about 4300 pictures. The data was split such that 70% used for training, 10% were available for validation, and 20% remained for testing. Some example pictures are shown in Fig. 2. All pictures were taken under normal lighting conditions, i.e. a combination of daylight and artificial light. To simulate lighting conditions at evening and night, the contrast and brightness of the images was lowered by 40% and 60%, respectively. These images simulate badly lit areas such as storage rooms or a power outage. An example can be seen in Fig. 3.

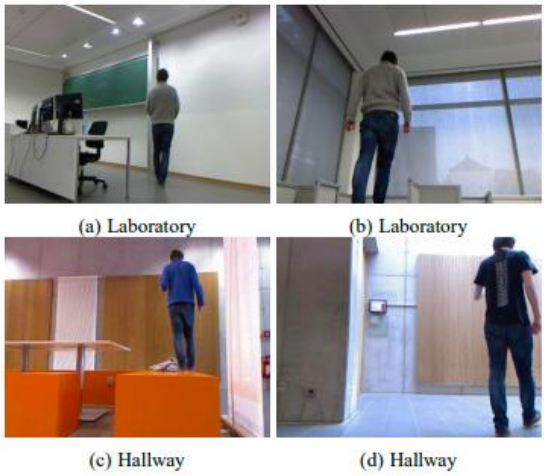

Figure 2. Example pictures from our dataset.

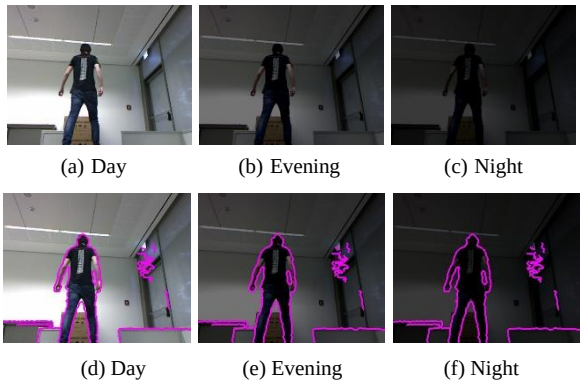

Figure 3. Example pictures with and without augmented silhouettes for different lighting conditions.

C. Training of the Neural Network

We trained and validated the network using daylight images only. By training only on daylight images, we could show that a training of the CNN with RGB images acquired at different lighting conditions is not necessary thanks to the silhouette augmentation of the RGB images. In the dataset we manually labeled only persons according to the format required by the neural network. We did not create bounding boxes for any other object class. This ensured that the network could only learn the shapes of persons. An expansion to other object classes is conceivable, as the silhouettes found during preprocessing were not limited to people. During training, the network may detect non-people objects as well, and predict bounding boxes around them, but it is only rewarded if the bounding box matches with the ground truth box of the correct class. In our case the only correct class comprises persons. We trained the CNNs with an input layer size of 416×416 over 5000 iterations. The batch size was 64, and the learning rate was set to 0.01. The YOLOv3 implementation from Bochkovskiy was used (https://github.com/AlexeyAB/darknet). We started training with a pre-trained weights file which was created by the author of YOLOv3 (https://pjreddie.com/media/files/darknet53.conv.74).

After 1000 iterations, an accuracy check based on the mean Average Precision (mAP) metric from was performed [13]. This check was repeated every 200 iterations. If the new network weights showed higher accuracy, they were selected as new weights. Otherwise the best weights remained unchanged.

For learning and evaluation, a PC with the following hardware specs was used: AMD Ryzen 9 3950X CPU, 64 GB RAM, 500 GB SSD, and two ASUS 8 GB RTX 2080 SUPER GPUs. We also implemented the algorithm on the Nvidia Jetson TX2 single-board computer which is part of the EvoRobot. We used the Robot Operation System (ROS) framework to implement the system architecture and wrote a ROS Interface based on Jung's work (https://github.com/jkjung-avt/tensorrt demos). We converted the neural network's weights to TensorRT [14] using a script written by Jung to use the full processing power of the Jetson TX2 board. This decreases the overhead of the YOLO-framework [15].

III. Results and Analysis

A. Accuracy

For the accuracy metric of the neural network, the Average Precision AP value for a single Intersection over Union (IoU) was applied. The IoU is defined as:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

with A being the size of the area of the predicted bounding box and B the size of the area of the ground truth bounding box. It describes how much a predicted bounding box overlaps with the bounding box of the ground truth [16]. If the overlapping area is bigger than a specific threshold, here set to 50%, the object is counted as correctly identified. This enables calculations of the precision as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

with *TP* being the number of true positives and FP the number of false positives [16]. The Recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

with *FN* being the number of false negatives [16]. When this metric is applied to all images, a Recall-Precision diagram is created. The area under the graph is the AP value. It is commonly approximated by summing up areas of rectangles calculated at 101 equidistant data points over the interval ranging from 0 to 1 [13].

TABLE I. AP (IoU = 0.50 %) Values of RGB, RGB+Augmentation (RGB+Aug), Depth and Depth+Augmentation (Depth+Aug) under Diff. Lighting Cond

CNN trained on: \ Lighting Cond.:	Day	Evening	Night
RGB	99.0 %	97.99 %	50.09 %
RGB+Aug	98.99 %	98.98 %	92.22 %
Depth	98.99 %	98.99 %	98.99 %
Depth+Aug	99 %	99 %	99 %

In Table I, we show AP values (IoU = 50%) for the different approaches. With RGB we refer to a network trained on the RGB data only. The label RGB+Aug indicates that the network was trained with RGB images augmented with the depth-based silhouette information. The label Depth implies that the net- work was only trained on depth data, while Depth+Aug means that the network was trained with depth images augmented with silhouette. As lighting conditions did not have any effect on the depth images, their AP values are the same for all lighting conditions. This is generally the case, if a structured light or a time-of-flight camera is used to record the depth data. Our experimental results can be summarized as follows:

1) YOLOv3 generally performed well for all four networks and lighting condition. We only encountered one false positive detection for RGB as well as for the RGB+Aug dataset. This means that the dataset was no particular challenge for the neural network. Furthermore, the high mAP values confirm this assumption.

2) Under challenging lighting conditions augmentation improved the AP (IoU = 50 %) from 50.09 % to 92.22 %.

3) Under normal lighting conditions, there is little performance difference between the RGB and depth-based CNNs.

B. Localization

TABLE II. mAP COCO Values for RGB, RGB+Augmentation (RGB+Aug), Depth and Depth+Augmentation (Depth+Aug) under Diff. Lighting Cond

CNN trained on: \ Lighting Cond.:	Day	Evening	Night
RGB	82.74 %	79.06 %	36.46 %
RGB+Aug	83.03 %	81.46 %	72.68 %
Depth	81.2 %	81.2 %	81.2 %
Depth+Aug	81.9 %	81.9 %	81.9 %

Unlike accuracy, where a detection is just counted based on a certain degree of overlap of two bounding boxes, localization tells us how well the detected bounding boxes align with the corresponding ground truth. For evaluation of the localization, the COCO metric was applied [13]. This metric calculates the mean Average Precision (mAP) of a network by averaging the AP for several IoU scores. COCO uses score values ranging from 50% to 95% with a step size of 5% applies it to all images and generates several Recall-Precision diagrams. The mAP is the average over AP for the different IoU scores.

In Table II, the mAP values for the different lighting conditions are shown. Since varying lighting conditions did not have any effect on the acquired depth images, the mAP values are again the same under all three lighting conditions. From these results, we can see that a neural network trained on depth image data can detect people with higher precision under challenging lighting conditions than a network trained on RGB data. However, under good lighting conditions, neural networks trained with RGB frames or RGB+Aug images outperform those trained on depth data alone by about 0.8 to 1.1%. In both cases of augmented image data (RGB, Depth), we see an increase in performance. Under challenging lighting

conditions the performance of the CNN trained with augmented RGB images was found to be more than twice as good as a network solely trained on standard RGB images.

C. Speed

On the PC hardware, our algorithm ran with 32.01 ± 3.84fps, while only 5.56 ± 0.10fps were achieved on the Jetson TX2 board. There, the preprocessing overhead reduced the frame rate by 0.39 fps. Note that the rate of 5.56 fps is still fast enough for people detection in our use case due to the rather low speed of AGVs when operating in areas also occupied by humans. For example, an emergency stop of the robot could be issued if a person is detected in close range of the robots path, with the robot driving 1.5 $\frac{m}{s}$. Additionally, the performance on the PC shows that with increased processing power, the detection time could be drastically reduced.

IV. Conclusions and Future Work

Our experiments suggest that the proposed people detection system can be combined with a mobile robotic platform and used in practice. Since our algorithm scales well with available hardware resources, it can also take advantage of improving hardware resources.

Our results also show that data fusion implemented as a pre-processing step outside of a neural network can increase performance for challenging use cases. In our case, the accuracy in case of poor lighting conditions was improved from 50.09% to 92.22%, while localization performance was boosted from 36.46% to 72.68%.

We also found that our depth-based silhouette augmentation approach always resulted in performance gains suggesting that it may be a promising approach for future work.

With our approach we could, for example, create a dataset to train a segmentation neural network and investigate if the performance could also be boosted.

Additionally, an in-depth study on the performance gain by augmentation of depth images with silhouettes is planned.

A more challenging dataset including occlusion of people to test the accuracy of the neural network is also part of the planned future work.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Florian Spiess and Lucas Reinhart conducted the research, analyzed the data, and wrote the paper. Norbert Strobel, Dennis Kaiser, Samuel Kounev, and Tobias Kaupp wrote the paper. All authors had approved the final version.

Acknowledgment

This work was supported by the Hans-Wilhelm Renkhoff Stiftung [17].

References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1.
[2] D. Konig, *et al.*, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.
[3] M. Vandersteegen, K. Van Beeck, and T. Goedeme, "Real-Time multispectral pedestrian detection with a single-pass deep neural network," in *Image Analysis and Recognition*, Springer, Jun. 2018, pp. 419-426.
[4] A. Eitel, *et al.*, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 681-687.
[5] T. Ophoff, K. V. Beeck, and T. Goedeme, "Improving real-time pedestrian detectors with RGB depth fusion," in *Proc. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018, pp. 1-6.
[6] T. Ophoff, K. V. Beeck, and T. Goedeme, "Exploring RGB+Depth fusion for real-time object detection," *Sensors*, vol. 19, p. 866, Feb. 2019.
[7] K. K. Luberg, "Human body poses recognition using neural networks with class based data augmentation," Master's thesis, University of Tartu, Estonia, 2018.
[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, Apr. 2018.
[9] J. Redmon, *et al.*, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779-788.
[10] G. Danciu, S. M. Banu, and A. Caliman, "Shadow removal in depth images morphology-based for Kinect cameras," in *Proc. 16th International Conference on System Theory, Control and Computing*, 2012, pp. 1-6.
[11] W. Birkfellner, *Applied Medical Image Processing: A Basic Course*, CRC Press, 2016.
[12] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679-698, 1986.
[13] T. Lin, *et al.*, "Microsoft coco: Common objects in context," in *Proc. European Conference on Computer Vision*, 2014, pp. 740-755.
[14] Nvidia tensorrt. (Jul. 7, 2020). [Online]. Available: https://developer.nvidia.com/tensorrt
[15] Nvidia deep learning platform giant leaps in performance and efficiency for ai services, from the data center to the network's edge. (Jul. 7, 2020). [Online]. Available: http://www.nextplatform.com/wp-content/uploads/2018/01/inference-technical-overview-1.pdf
[16] M. Everingham, *et al.*, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, Jun. 2010.
[17] Warema Renkhoff SE. (Jan. 26, 2020). [Online]. Available: https://www.warema-group.com/en/WAREMA_group/WAREMA_Renkhoff_SE.php

Florian Spiess was born in Schweinfurt on December 18, 1988. He received his M.Eng. in Electrical Engineering and Information Technology from University of Applied Science Wuerzburg Schweinfurt, Schweinfurt, Germany in 2016.
He is a Ph.D. student at the chair of software engineering at the University of Wuerzburg, Germany.