# Ablation Study of a Person Re-Identification on a Mobile Robot Using a Depth Camera

Sebastian Flores
*AI and Autonomous Systems*
*Fraunhofer Institute for Material Flow and Logistics*
Dortmund, Germany
sebastian.flores@iml.fraunhofer.de

Jana Jost
*Robotics and Cognitive Systems*
*Fraunhofer Institute for Material Flow and Logistics*
Dortmund, Germany
jana.jost@iml.fraunhofer.de

*Abstract*—**In this paper, we describe an ablation study for a person re-identification API on a mobile robot, for a closed-world setting, using only the IR gray value image of a depth camera. Previously, we have trained the state-of-the-art neural network for person re-identification with common parameters and methods. The resulting real-time application reached as closed-world setting a rank-1-accuracy of 94.78% and a mAP of 68.16%. Now, we focused on increasing the accuracy by removing and adjusting the image processing pipeline of our dataset. By these adjustments, we have reached a rank-1-accuracy of 98.56% and a mAP of 79.05%.**

*Index Terms*—**person re-identification, neural network, mobile robot**

## I. Introduction

Nowadays, robots are common in many production and logistics facilities as well as other areas e.g., hospitals where humans are present and need to work with those systems. Unfortunately due to the shortage of skilled workers, most of them are not used to interacting with robots. Further, the acceptance of the human towards the robot only increases if the human feels comfortable and the interaction with such a system is designed in an intuitive and user-individual way. Therefore, the machines have to adapt to the individual person. Once a robot takes into account the characteristics and skills of the workers e.g., operating speed, height or process know-how, the processes can become more efficient, the interaction can be designed in an ergonomic way and the process times can be reduced.

In this paper, we address the use-case of a mobile social robot transporting e.g., bins and packages in warehouses as well as production facilities or food trays and medical equipment in hospitals. The design of the robot keeps in mind various ergonomic aspects. It can lift the transported goods up to an ergonomic height (see Fig. 1b) and is equipped with a display to directly inform the worker about its current state. Further, our robot [1] behaves according to social norms adapted from the human-human interaction. It can change its path or adapt its speed depending on the situation and the individual human (cf. Proxemics) as well as follow the worker through the warehouse. To realize such a human-robot interaction, the first step is to identify – and later on re-identify – the human worker as an individual.

Common approaches for re-identification (re-ID) either use hardware dependent solutions e.g., transponders at the human or implement it by using anthropometric data [2], gait analyzes [3] or face recognition [4]. However, the state-of-the-art solutions do not suit our use-case and robot. Existing solutions e.g., [5]–[8] need to work on data with a higher refresh rate, RGB-data and resolution than the one our camera is offering or from another perspective than our mobile robot allows.

Therefore, we collected images of 31 persons in a realistic warehouse environment from the frog's eye perspective. Further in our previous research, we set up an image processing pipeline to extract a dataset. Then we loaded, trained, validated and evaluated a model based on the state-of-the-art neural network (NN) from [9]. The impact of common training tricks was analyzed. To improve the accuracy of the model, we present in this paper an ongoing ablation study.

The study focuses on the image pipeline, which was designed based on solutions for person re-identification [9] using RGB-images from surveillance cameras, such as the Market1501 [10] dataset, combined with our created mask using the pose of the person related to [11]. Contrary to the previous work, we applied it for image processing and not in the person re-identification model itself. In addition, our image normalization is specified for the application on IR gray value images. The normalization in previous work cannot be applied.

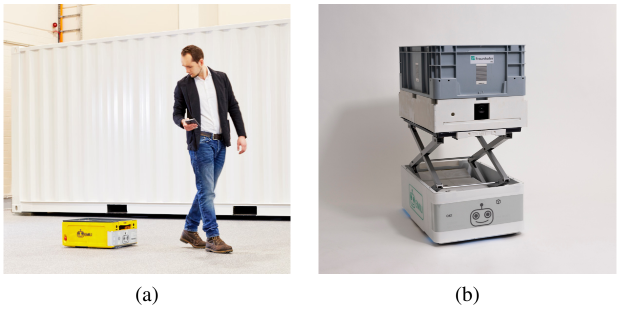
The main reason for our image processing pipeline is to only extract the persons key-relevant information out of the image and then use it as an input for the re-ID model. Thus the model can faster reach the global minimum and lead to a higher score than without any pre-processing steps. Therefore, the ablation study investigates the influence of the different steps of the image processing pipeline (see Fig. 2) on the model.

## II. Technical Description

This ablation study is based on our previous research, in which we discovered the best fitting model for our use-case, model-no. 9. In this study, we focus on the influence of our image pipeline (see Fig. 2).

This work is financed by the German Ministry of Education, Culture and Research under grant number 16MEE0213 (IMOCO4.E). In addition, this research has received funding from the ECSEL Joint Undertaking from the European Union's Horizon 2020 research and innovation program under grant agreement 101007311 (IMOCO4.E).

DOI: 978-1-6654-8240-0/22/$31.00 ©2022 IEEE Electronic Preprint



(a) (b)

Fig. 1. Our mobile robot in yellow and white, following a person in (a) and carrying a small load carrier in (b).

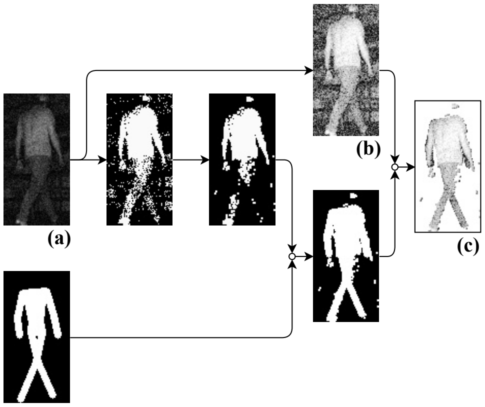


Fig. 2. Image processing for the person re-ID, based on paper [12]. (a) the original image. (b) the original histogram image. (c) the pose image.

### A. Model

Our model is based on a ResNet18 [13] NN with a last stride set to one and the random erasing augmentation training method is used, both is based on the paper [9]. Since we are using a gray value image the channels are reduced to one. Additionally, for a possible (re-)training on the robots hardware (a Jetson TX2 board), the batch-size is decreased to 16. Further, the number of epochs is halved to 60 epochs, which cuts the training time in half. The reduction of the epochs also leads to a decrease of the milestones to 20 and 35.

Since the warm-up learning rate and label smoothing [9] decreased the scope of our model, they are no longer applied. Further, the BNNeck and the center loss from [9] were also not used, because the loss of our base model, model-no. 1, converged very quickly and the classes were already separated very pronounced.

### B. Image Processing Pipeline

In our previous research, we used the pose image Fig. 2c to create the dataset. To do this, we utilized the cut out gray value image (see Fig. 2a), which depends on the detected person in the depth image followed by the calculated region of interest (ROI). This image was binarized and then an opening was applied (erosion and dilation) to reduce the noise in the image. After that the image was combined to a mask using the drawn pose of the person. This mask was finally utilized with Fig. 2b, which is the cut out gray value image with an applied histogram equalization.

The histogram equalization was used, so that the person in the image has the same brightness no matter how far or close to the camera the person is. This is important because a IR gray value image created by a depth camera was used as input. It is similar to the histogram equalization done by the authors of [14]. They analyzed the intensity normalization using T1-weighted magnetic resonances imaging data of the cerebellum to train a model for segmentation of tumors and lesions [14]. In contrast to their approach, our histogram equalization is calculated for every new image again and not for all images or a batch at ones. This was necessary because we are using a IR gray value image and so our histogram equalization has to be dynamically.

In this work, we are also using the original image Fig. 2a and the original histogram image Fig. 2b to create datasets and we will look at the influence of our image processing pipeline for our model, model-no. 9. The detection of persons in the depth camera images and cropping them out as described in our previous research [12], stays untouched.

Further, it is important that our resulting images, used for the model, were converted to a float tensor in a range between 0 and 1. Then they were normalized between -1 and 1, using a standard deviation and a mean of 0.5. Since we are using the IR gray value image created by a depth camera, we could not calculate the mean and standard deviation as it is usually done, neither for the whole dataset nor the batches.

## III. Preliminary Results

Each of the created datasets – original-, original histogram- and pose-dataset – is split into a training/validation- and an evaluation dataset. Every dataset involves the same 31 persons as IDs. The evaluation dataset contains the IDs 00, 01, 02, 03, 09, 10, 11, 21, 24, 28 and 30. The 20 persons left are included in the training dataset. The IDs are not consecutive, because the division is based on the visual characteristics – short or long trousers or a skirt, their gender and if they were wearing a mask because of the corona pandemic.

### A. Training

The results of training and validating the model-no. 9 which is optimized for the dataset using the pose images (see Fig. 2c), on the datasets using the original and original histogram image are shown in Fig. 3 and Fig. 4. At the beginning, both figures show that using less image processing techniques reaches a higher starting accuracy and a lower starting loss. The course of the curves are similar.

For training and validation the accuracy when using dataset no. 3, pose images, increases in contrast to the accuracy when using dataset no. 2, original histogram images, by 1.03 % and 0.56 %. The loss decreases by 0.0355 and 0.0244 (see Table I). Compared to the dataset no. 2 as input, with the dataset no. 3, pose images, as input, the accuracy increases further 0.1 %

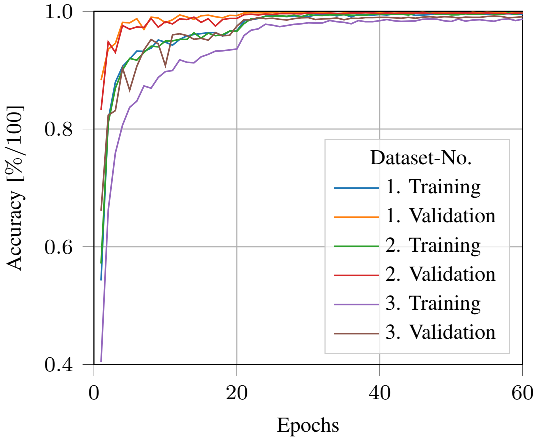


Fig. 3. Accuracy comparison between the three datasets, using original, original histogram or pose images.

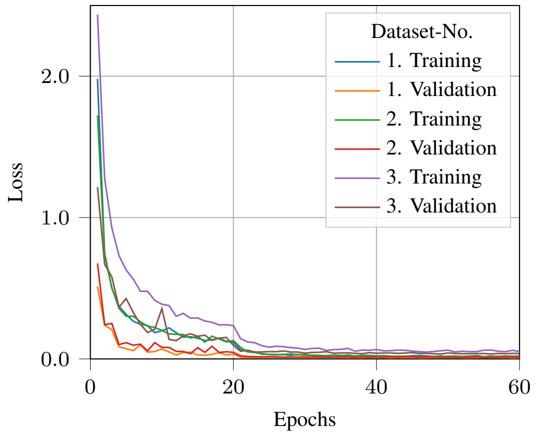


Fig. 4. Loss comparison between the three datasets, using original, original histogram or pose images.

and 0.07 % and the loss decreases by 0.0017 and 0.0026, for training and validation.

### B. Evaluation

During the evaluation the rank-1-accuracy (r1) and rank-5-accuracy (r5) increases when using dataset-no. 3, pose images, in contrast to dataset-no. 2, original histogram images, by 2.98 % and 0.46 % (see Fig. 5 and Table I). The mAP increases by 8.96 %. Furthermore, the r1, r5 and mAP increases form dataset-no. 2 to dataset-no. 1 by 0.8 %, 0.17 % and 1.93 %. The course of the mAP of each ID, (see Fig. 6), shows that the the processing pipeline, using the pose and additional techniques, results in a significant lower mAP for the IDs 02, 03, 10 and 30. For ID 24 the datasets orhi and pose score lower.

### C. Image Pipeline

The duration of the image pipeline sped up analogous to the increase of the training/validation and evaluation scores (see Table II). The duration of the original image represents the needed time to crop the image out of the camera image. This is done in 16 $\mu$s as mean. Using the additional histogram equalization for the orhi image, the time increases to 211 $\mu$s. The most time, with a mean of 3322 $\mu$s, requires the pose images which represents the whole image pipeline (see Fig. 2).

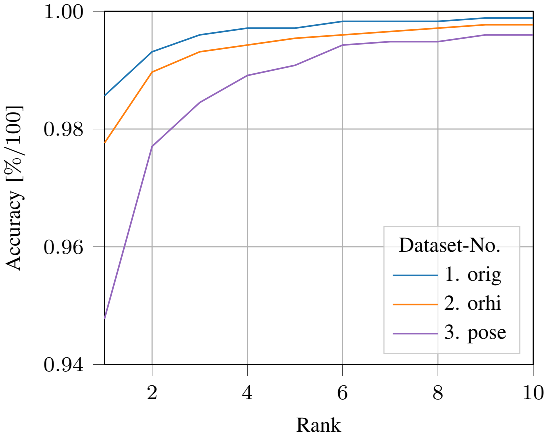


Fig. 5. CMC-curves of the three datasets, using original, original histogram or pose images.

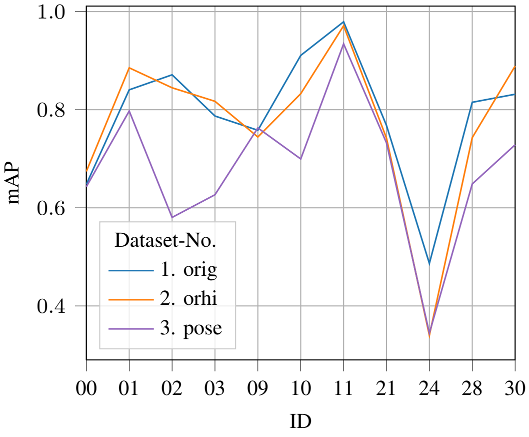


Fig. 6. The IDs represent the person, which the evaluation dataset contains.

## IV. Discussion

The conducted ablation study showed the best results without any steps of the image processing pipeline except the detection of region of interest.

The model accuracy increased for training and validation by 0.93 % and 0.63 % and also the loss decreased by 0.0338 and 0.027. These values are marginal, but we have almost reached the end of refinement. More interesting is the improvement during the evaluation, which increases the r1, r5 and mAP by 3.78 %, 0.63 % and 10.89 %. Further, the individual mAP of the IDs 02, 03, 10, 24 and 30 were higher (see Fig. 6).

The reduction of the duration of the image pipeline is marginal in context of the Live-API, which runs between 10 and 16 FPS, for one up to three persons in an image at once. This means, we do not have to shorten our image pipeline for a time improvement.

TABLE I
COMPARISON OF THE RESULTS BETWEEN THE DATASETS.

| | Training | | Validation | | Evaluation [%] | | |
|---|---|---|---|---|---|---|---|
| Dataset-No. | Acc [%] | Loss | Acc [%] | Loss | r1 | r5 | mAP |
| 1. orig | 99.64 | 0.0145 | 99.81 | 0.0071 | 98.56 | 99.71 | 79.05 |
| 2. orhi | 99.74 | 0.0128 | 99.74 | 0.0097 | 97.76 | 99.54 | 77.12 |
| 3. pose | 98.71 | 0.0483 | 99.18 | 0.0341 | 94.78 | 99.08 | 68.16 |

TABLE II
COMPARISON OF THE NEEDED TIME FOR THE IMAGE PIPELINE, EACH CALCULATED USING 1000 CYCLES.

| | Duration [$\mu$s] | | |
|---|---|---|---|
| Image Pipeline | Min | Max | Mean |
| 1. orig | 9 | 180 | 16 |
| 2. orhi | 132 | 1327 | 211 |
| 3. pose | 1953 | 9729 | 3322 |

Nevertheless, the used model has been initially applied to the dataset Market1501. Therefore, the different parameters as well as the additionally ones e. g., the BNNeck and the center loss from [9] need to be adjusted. Especially, the influence of the WU and the LSm should be examined since [9] does not include the CMC-curve.

## V. PLANNED NEXT STEPS

The next steps will be focusing further on the images themselves. On the one hand we want to invert the images Fig. 2a-c. The preliminary results of the ablation study have shown that the Fig. 2a reaches the best scores. One reason might be that the background of image Fig. 2c was not set to black. Therefore, promising results for our use-case might be achieved when setting the background of image Fig. 2c to black (0). Additional to this, the influence of inverting this resulting image could be interesting too. Since the usage of image Fig. 2a leads to higher scores compared to the usage of image Fig. 2b it might be useful to cancel the histogram equalization. This can be validated once the described background change to black reaches better results. On the other hand we want to investigate how a background of gray might influence the overall results. The applied normalization had a range from -1 to 1 with a standard deviation and a mean of 0.5 which leads to the usage of gray for the value 0.

In addition, we want to swap the cosine similarity with the Euclidean distance. The authors of [9] observed that the cosine distance led to a better performance than the usage of the Euclidean distance. However, there dataset consisted of RGB-images. Therefore, it is possible to reach better results for our IR gray value dataset when applying the Euclidean distance.

Once the image processing pipeline is adjusted, we need to optimize the model for the final image input in the same way as it is done for the image Fig. 2c in our previous research [12].

Overall, we want to extend our dataset after the corona pandemic and include more realistic scenes in logistics or production facilities as well as hospitals.

## REFERENCES

[1] T. Kirks and J. Jost, "Mensch-Technik-Interaktion in Industrie 4.0 Umgebungen am Beispiel von EMILI," in *Handbuch Industrie 4.0*, ser. Springer NachschlageWissen, B. Vogel-Heuser, T. Bauernhansl, and M. ten Hompel, Eds. Springer, 2019.

[2] D. Liciotti, M. Paolanti, E. Frontoni, A. Mancini, and P. Zingaretti, "Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration," in *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, ser. Lecture Notes in Computer Science Ser, K. Nasrollahi, Ed. Springer International Publishing AG, 2017, vol. 10165, pp. 1–11.

[3] A. Virginia and A. Ricardo, "Person Identification Using Anthropometric and Gait Data from Kinect Sensor," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence and the Twenty-Seventh Innovative Applications of Artificial Intelligence Conference, 25 - 30 January, Austin, Texas, USA ; [along with the AAAI/SIGAI doctoral consortium]*. AAAI Press, 2015, pp. 425–431.

[4] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face Recognition with Learning-Based Descriptor," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*. IEEE, 2010, pp. 2707–2714.

[5] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust Depth-Based Person Re-Identification," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 26, no. 6, pp. 2588–2603, 2017.

[6] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. van Gool, "One-Shot Person Re-identification with a Consumer Depth Camera," in *Person re-identification*, ser. Advances in in computer vision and pattern recognition, S. Gong, M. Cristani, S. Yan, C. C. Loy, and M. Christani, Eds. Springer, 2014, pp. 161–181.

[7] E. Bondi, P. Pala, L. Seidenari, S. Berretti, and A. Del Bimbo, "Long Term Person Re-identification from Depth Cameras Using Facial and Skeleton Data," in *Understanding Human Activities Through 3D Sensors*, ser. Lecture Notes in Computer Science, H. Wannous, P. Pala, M. Daoudi, and F. Flórez-Revuelta, Eds. Springer International Publishing, 2018, vol. 10188, pp. 29–41.

[8] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti, "A Feature-Based Approach to People Re-Identification using Skeleton Keypoints," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5/31/2014 - 6/7/2014, pp. 5644–5651.

[9] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of Tricks and a Strong Baseline for Deep Person Re-Identification," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*. IEEE, 2019, pp. 1487–1495.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 12/7/2015 - 12/13/2015, pp. 1116–1124.

[11] A. Bhuiyan, Y. Liu, P. Siva, M. Javan, I. B. Ayed, and E. Granger, "Pose Guided Gated Fusion for Person Re-identification," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 3/1/2020 - 3/5/2020, pp. 2664–2673.

[12] S. Flores and J. Jost, "Person Re-Identification on a Mobile Robot Using a Depth Camera," in *press*.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6/27/2016 - 6/30/2016, pp. 770–778.

[14] N. Jacobsen, A. Deistung, D. Timmann, S. L. Goericke, J. R. Reichenbach, and D. Güllmar, "Analysis of Intensity Normalization for Optimal Segmentation Performance of a Fully Convolutional Neural Network," *Zeitschrift fur medizinische Physik*, vol. 29, no. 2, pp. 128–138, 2019.