Person Re-Identification on a Mobile Robot Using Only the IR Gray Value Image of 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 designed and implemented a real-time person re-identification API on a mobile robot, for a closed- and open-world setting, using only the IR gray value image of a depth camera. Since common datasets are not usable we created our own dataset using the IR gray value images, the pose and image processing techniques. Then we trained the state-of-the-art neural network for person re-identification with common parameters and methods. For running it in real-time, we sped up the model as well as the application. It is possible to re-identify three persons at once, at around 10 FPS. Our model reaches as closed-world setting a rank-1-accuracy of 95.5%. With an additional threshold, coming from rank-1-accuracy of closed-world setting, our real-time application reaches as open-world setting a f1-score of 79.44% and a recall of 68.44%.**

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

I. Introduction

The interaction between humans and machines increased over the past years. In order to make this interaction accessible to everyone, regardless of their knowledge in technology, the machines have to adapt to the individual person. This adaptation leads to an intuitive usability, e.g. mobile robots that follow employees in hospitals as well as represent a source of information and help in public facilities. In facilities, the characteristics and skills of employees, such as working height, workflow or operating speed can be saved as profiles in the robot so that it knows how to react accordingly. This can result in an improved process execution, an ergonomic posture of the worker and in a reduction of the process time. To do this, as a first step persons need to be re-identified.

II. Background

This paper is divided into eleven parts. In the first part, Sec. I, the importance of the interaction between humans and machines which leads to the necessity of re-identification (re-ID) is described. Sec. II, presents an overview of related work according to interaction-modality and person re-ID. The third part, Sec. III, deals with use cases and our robot. In the next part, Sec. IV, our concept is described based on related work. To realize this concept, a dataset will be generated in Sec. V using the pose of the person and further image processing techniques. In the sixth part, Sec. VI, the state-of-the-art neural network (NN) will be loaded, trained, validated and evaluated. The impact of common training tricks will be also analyzed. Sec. VII, is based on improving the NN in order to speed it up and make it suitable for our robot. In Sec. VIII, the application will be created to run in real-time. The last three parts, Sec. IX-XI are the conclusion, the discussion and the outlook. These parts deal with the advantages and borders of this work.

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).

A. Human-Technology Interaction

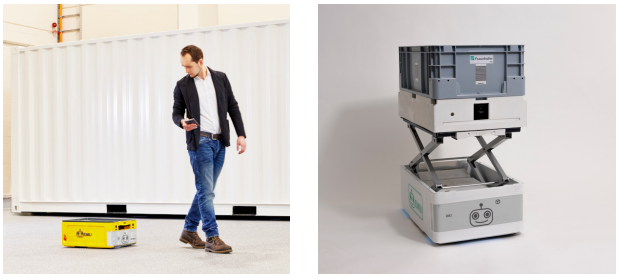
Any collaboration between humans and technology is viewed as human-technology interaction (HTI). The human-computer interaction is part of it. The HTI must be able to communicate over at least one information channel, e.g. visually. Examples for HTI are mobile robots that can follow a person [1], gesture control [2] or person re-ID [3]. To include the human better into the factory of the future and therefore, to retain more efficient processes, one has to consider not only safety requirements, but also aspects to reduce physical as well as mental workload. Here human-human concepts e.g. proxemics have to be adapted for HTI. Further, the technology has to perceive the worker as an individual. The re-ID can be implemented by using the anthropometric data [4], gait analyzes [5] or face recognition [6]. There are also hardware dependent solutions e.g. transponders.

B. Person Re-Identification

In this section, we present an overview of datasets, solutions for person re-ID and already existing robots using these. Algorithms for person re-ID using Deep Neural Networks (DNNs) have made good progress in recent years [7].

1) Datasets: These DNNs are trained on publicly accessible large RGB-image datasets such as MARKET1501 [8] and CUHK03 [9] or video datasets such as iLIDS-VID [10], which contains sequences of persons saved as images. The datasets are generated using surveillance cameras from elevated positions. There are also datasets captured with an RGB-D camera, like BIWI RGBD-ID [11], IAS-Lab RGBD-ID [12] or RGBD-ID [13] which include RGB-, depth-data and skeleton coordinates. An additional dataset is the Florence 3D

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).

Actions Dataset [14], which only contains RGB-videos and skeleton coordinates, but no depth data.

2) Works in Re-Identification: One of the state-of-the-art models for RGB-datasets [7] is the Bag of Tricks (BoT) Baseline [3], which is using the ResNet50 [15] model as backbone and applied training tricks and customized parameters. Using the ResNet50 model for person re-ID is a common approach. Another method uses the RGB- and depth images for a cross-modal re-ID in which the model learns a shared feature representation space of the person in both images [16]. The network is called cross-modal distillation network.
There are several further soft biometric cues, like the previous described anthropometric person re-ID (see Sec. II-A) using the point cloud data [17]–[20]. Those are calculating a descriptor to match persons. This descriptor contains, either face and body part geometry- [19], tracked joints- [20], all joints- [18], [20], depth voxel covariance or locally rotation invariant depth shape- [17] data. There is also a solution matching the whole point cloud of the subject/person [18]. The goal of these works are trying to reach a higher accuracy using publicly available datasets to compare their work.

3) Works in Robotics: The works in robotics are either implementing the previous solutions on mobile robots or trying to create their own solution customized for their robots and use cases. The existing works are based on DNNs using the RGB-Image for the re-ID. One existing project [21] uses an RGB camera and several LIDAR sensors. For the person re-ID the Convolutional Channel Feature model with online boosting is implemented. It uses OpenPose [22] for person detection and an unscented Kalman filter for tracking. Another project [23] uses the re-ID models PCB [24] and IID [25], a face recognition model [26] and a person tracking model [27].
The last interesting project [28] locates persons using the Yolo V3 model [29]. The re-ID is based on their own created HENet model with the ResNet50 as a backbone. The calculation is running on a server, where the images are send to.

III. Our Mobile Robot and Use Case

To identify a possible solution for our mobile robot, we first describe the use case and introduce the robot.

Our mobile robot [30] works in warehouses together with order pickers and in hospitals with nurses and caregivers. It transports bins and packages in logistics or food trays in hospitals and can lift them up to an ergonomic height (see Fig. 1b). Further, it can follow the human and shall recognize the individual person. Because of technical aspects e.g. higher speeds, the lift is retracted while the robot follows a person (see Fig. 1a). The already existing "follow-me" function uses the low resolution PicoFlexx depth camera [31] in 10 FPS mode to identify a person from a frog's eye view. A person re-ID solution should therefore setup on the given hardware and result in an input for the "follow-me" function. The mode was chosen because it has to run in real-time and still offers a long range, 1.0 to 4.0 meters, at a high exposure time, 1000 μs.

The mobile robot uses the Jetson TX2 Board [32] for data processing. The PicoFlexx depth camera is attached to the robot at a height of 335 mm and tilted 4° backwards for an increased view starting 1 meter in front of the robot. Furthermore, the construction of the robot leads to the camera position, so it cannot be changed. This camera outputs a point cloud, a depth image and an IR gray value image. The later represents the signal strength of the active illumination [31] and contains more information of the person as the depth-image (see Fig. 2a and Fig. 2b with respect to Fig. 2c). The resolution is 224 x 171 pixel and the camera has several modes starting from 5 FPS up to 45 FPS. The increasing FPS lead to less brightness of the IR gray value image and therefore a decrease in exposure. Further, with higher FPS modes the range declines too.

IV. Conceptual Design

To setup on existing work for an implementation of person re-ID for our mobile robot, the related works from Sec. II-B will be discussed and a conceptual design will be described.

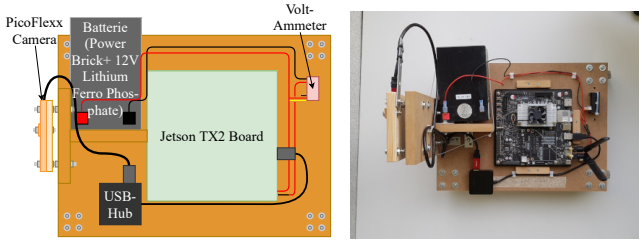
A. Discussion of Related Works

From the works in re-ID, seems only the DNNs usable for our use case. The existing solutions, soft biometrics, e. g. gait analyses, and anthropometrics, the measurement of body lengths, need to work on data with a higher refresh rate and resolution than the one our camera is offering. Therefore, the faces are rarely in the field of view (see Fig. 1).

The three presented methods in works in robotics (see Sec. II-B3) are using RGB-images, face recognition, not the state-of-the-art models or a different perspective. Therefore, they cannot be used on our robot. Further, the faces are rarely in the filed of view and even if they are the resolution is very low. However, the model in [28] uses the state-of-the-art ResNet50 model as a backbone and is the only one which is working from the frog's eye view, similar to our robot. Nevertheless, the model cannot be applied to our use case. On the one hand, the used models for person re-ID, PCB and IID, score to low in rank-1-accuracy and mAP and on the other hand, the NN is running on a server. In addition, the persons are identified with Yolo V3 [29], which outputs the bounding box but not the pose of persons. In our use case, the pose is needed for the already existing "follow-me" function.

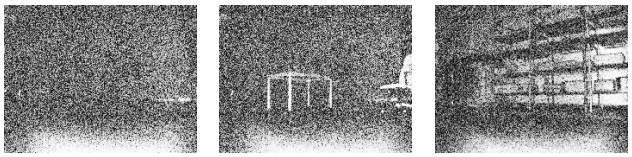
(a) (b) (c) (d)

Fig. 2. Comparison between (a) BIWI depth image and (b) PicoFlexx IR gray value image created from (c) PicoFlexx point cloud. The BIWI depth image has an increased brightness for a better representation. In (d) is the scene from (b) and (c) captured as RGB-image, for logistic context in front of shelves.



(a) (b)

Fig. 3. Experimental bare-bone setup, as sketch in (a) and assembled in (b).



(a) (b) (c)

Fig. 4. A histogram equalization was used in this figure for a better representation. (a) The open space. (b) The chair and the table. (c) The shelf.

Also the previous described datasets (see Sec. II-B1) do not match the requirements of our robot (see Sec. III). On the one hand, the first three RGB-datasets [8]–[10] only include RGB-images which are taken from elevated position. On the other hand, the four RGB-D-datasets [11]–[14] consist of RGB-images, depth images and point clouds and not of IR gray value images. Further, the conversion of an RGB-image to a gray image does not result in an IR gray value image.

B. Our Concept

Our concept is divided into four parts. The first part is the creation of a dataset, which is needed because publicly available sets are not usable for our robot. In the images the persons need to be detected with as much information as possible. The datasets is divided into a training/validation dataset and in an evaluation dataset. The persons are evenly distributed on both sets. In the second part, a person re-ID model is developed. This person re-ID sets up on the ResNet50 model with additional tricks from the *BoT* paper [3]. The third part contains additional optimizations, e.g. to speed up the model, so that the model can be trained and evaluated on the robot. In the last part the application for a real-time purpose is created. This application uses the customized, optimized and accelerated model for the re-ID task.

V. Data Set Creation

To create a data set, an experimental setup was designed and built (see Fig. 3). We only used the bare-bone parts of our mobile robot which are the PicoFlexx depth camera and the Jetson TX2 Board. The camera is 4° tilted backwards at a height of 335 mm, as described in Sec. III.

31 persons were recorded as rosbag files using the ROS-framework, each one in three scenarios with two orientations and for 30 seconds. These scenarios were in front of an open space, a chair with a table (for the hospital use case) and a shelf (for the logistic scenario, see Fig. 4). During the recording the persons ran an eight, so that every site of them was saved. The persons were between 20 and 50 years old.

The persons from these recordings (see Fig. 5a) are detected (see Fig. 5b) using the trt_pose detection [33] based on [22] and cut out by the keypoints generated from them as a bounding box (see Fig. 5c and Fig. 5d). However, at least twelve keypoints of a person must be recognized, otherwise the image will not provide enough information of the person, which would lead to a false re-ID. In addition, the person's pose is drawn in a new black image of the same size (see Fig. 5e). This is similar to the authors' approach in [34], but instead of using it in the NN, we use it as a pre-processing step. The pose in Fig. 5e is drawn in another way than the trt_pose (see Fig. 5b). The torso is based on four filled triangles, so that if one keypoint is missing, it can be still be partially drawn.

The cut out gray value image is binarized and then the opening is applied (erosion and dilation) to reduce the noise, respectively to eliminate singular pixel errors (see Fig. 5f and Fig. 5g). Next, this image is combined with the image of the drawn pose to obtain a mask (see Fig. 5h). Thus, there is as many information of the person as possible. In this mask, the previous poorly visible legs are now represented correctly. Next, a histogram equalization is applied (see Fig. 5i), so that images of the same person have the same gray value level. In the last step, this image is multiplied with the mask, which results in an image containing only the person (see Fig. 5j) and is saved.

This procedure is applied to all 186 rosbag files of the

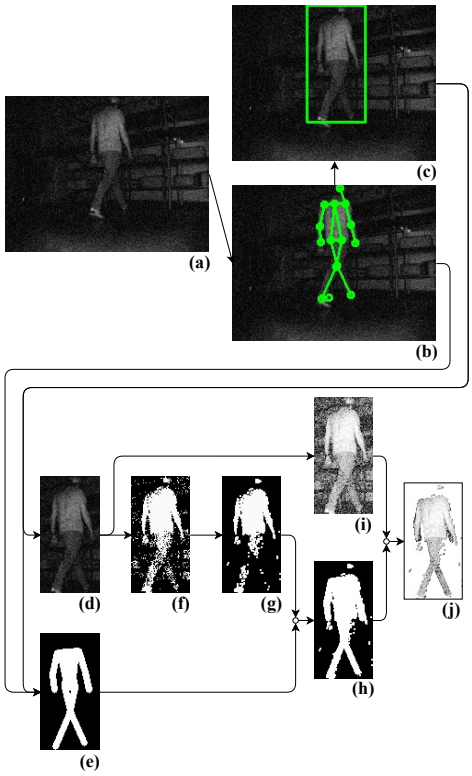


Fig. 5. Image processing for the person re-ID.

31 persons with 30 seconds 10 Hz image material each. After that, the images with at least twelve keypoints of the persons are evenly divided by their visual characteristics into two datasets. These are different types of clothes – short or long trousers or a skirt, their gender and if they are wearing a mask because of the corona pandemic. The first dataset contains 20 persons for training with validation and the others are in the evaluation dataset (see Table I).

VI. Person Re-Identification

For the person re-ID the state-of-the-art model needs to be trained and evaluated. With our created dataset and common tricks.

A. State-of-the-Art Model

The model ist based on the state-of-the-art model from [3] which includes a warm-up learning rate (WU), to boost the performance of the model, a random erasing augmentation (REA), to produce artificial occlusions, a label smoothing (LSm), to prevent an overfitting and the last stride (LSt), which can lead to a larger output of the feature maps of ResNet's last convolutional layer. Then the model was trained/validated and evaluated. In addition, the effects of deactivating these four methods separately were considered for our purpose.

B. Model Evaluation

The result of this consideration is shown in Table II. Furthermore, the characteristics of each model are listed. They are separated in four parts. The first part is Model-No. 1 and

TABLE I
Numbers of images per person-ID for both datasets.

Training and Validation			Evaluation		
Person-ID	Training	Validation	Person-ID	Gallery[a]	Query[b]
4	366	182	0	432	215
5	197	98	1	176	88
6	547	272	2	201	100
7	165	82	3	245	122
8	39	18	9	531	264
12	32	16	10	362	180
13	314	156	11	83	40
14	161	80	21	588	294
15	44	22	24	170	84
16	312	156	28	436	218
17	154	76	30	328	163
18	98	48			
19	288	142			
20	302	150			
22	362	180			
23	395	196			
25	240	118			
26	432	214			
27	435	216			
29	431	214			
$\sum$	5314	2636	$\sum$	3552	1768

[a]The gallery contains of known persons with their images.
[b]The images in the query are for the re-ID with persons from gallery.

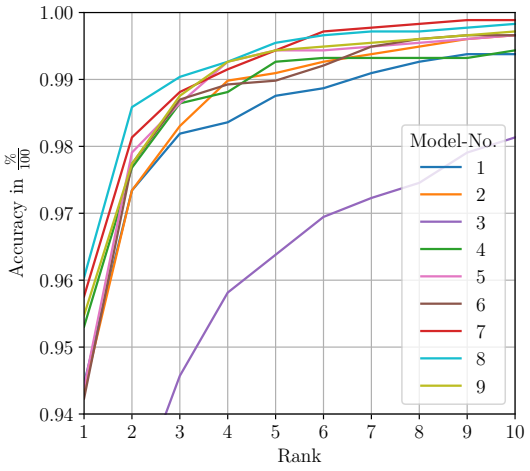


Fig. 6. CMC-curves of Model-No. 1 - 9.

this is the SOTA model as previously described in Sec. VI-A. In the second part, the methods are individually deactivated. For further optimization (see Sec. VII), the part three and four are listed. The table lists the WU, REA, LSm and LSt as well as the number of channels (Ch), the batch size (Ba), the epochs (E) and the training time (Duration). Furthermore for the evaluation, the table contains the rank-1-accuracy (r1), rank-5-accuracy (r5) and the mean average precision (mAP).

The WU improves the r1 as shown in [3], even in our model (see Table II), but they did not look at the course of the CMC-curve or the rank-n-accuracy (rn), although these values score better starting from rank $n = 3$ (see Fig. 6). Therefore, it will not be used.

The LSm stays deactivated too. Contrary to [3], it leads to

TABLE II
COMPARISON OF THE RESULTS BETWEEN ALL MODELS.

	Training									Evaluation		
Model-No.	ResNet	WU	REA	LSm	LSt	Ch	Ba	E	Duration	r1	r5	mAP
1	50	yes	yes	yes	1	3	64	120	118	0.945	0.988	0.656
2	50	no	yes	yes	1	3	64	120	118	0.943	0.991	0.646
(3)[a]	50	no	no	yes	1	3	64	120	118	0.876	0.964	0.554
4	50	no	yes	no	1	3	64	120	118	0.953	0.993	0.672
(5)[a]	50	no	yes	no	2	3	64	120	96	0.944	0.994	0.632
6	50	no	yes	no	1	1	64	120	112	0.942	0.990	0.683
7	50	no	yes	no	1	1	16	120	126	0.958	0.994	0.676
8	50	no	yes	no	1	1	16	60	63	0.960	0.995	0.684
9	18	no	yes	no	1	1	16	60	23	0.955	0.994	0.674

[a]These models are worse and skipped during the iterative adjustment.

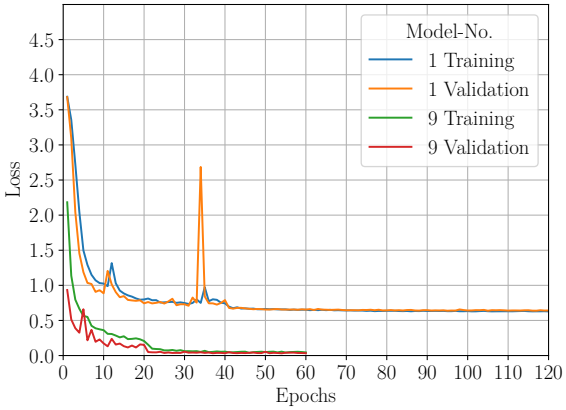


Fig. 7. Loss comparison between Model-No. 1 and 9. After Model-No. 4, there are only marginal changes between the average values of the curves, so Model-No. 9 was chosen for this representation.

TABLE III
COMPARISON OF THE INFLUENCE OF THE TRAINING TRICKS BETWEEN THE *BoT* PAPER AND OUR MODELS.

	BoT		Model-No. 2 - 5 [a]	
Influence	r1	mAP	r1	mAP
WU	+0.010	+0.012	+0.002	+0.010
REA	+0.026	+0.041	+0.067	+0.092
LSm	+0.001	+0.010	-0.010	-0.026
LSt	+0.006	+0.014	+0.009	+0.040

[a]Our models with same methods, listed in Table II.

a lower r1 and mAP value. The reason for this could be the smaller dataset, despite that the authors promise a significant increase for smaller datasets (without any proof). Another reason for the LSm to stay deactivated is that the loss improves by a factor of 14 from 0.63 to 0.045 during training and by a factor 18.88 from 0.642 to 0.034 during validation (see Fig. 7). A comparison for each method is shown in Table III.

VII. REAL-TIME RE-IDENTIFICATION

In order to make the model run in real-time on the Jetson TX2 Board, it needs further optimization and acceleration.

A. Model with further Optimization

The state-of-the-art model with deactivated WU and LSm will be further optimized. Therefore, the input channel is reduced to one, because only IR gray value images are used.

The Jetson TX2 Board is limited to 8 GB of RAM, therefore the batch size is reduced from 64 to 16. Now it is possible to (re-)train on the Jetson TX2 Board. The training with the batch size of 16 needs 6.7 GB RAM.

Next, the number of epochs and the milestones are reduced. This can be done because the accuracy changes only marginal after epoch 50. The epochs are reduced to 60 and contain a buffer of 10 epochs. The milestones will be reduced as well, from 40 and 70 to 20 and 35. This halves the training time.

The last step is to change the model architecture from ResNet50 to ResNet18. This increases the speed of the re-ID.

B. Model Evaluation

Reducing the input channel to one results in a worse r1 and r5. They decrease by 0.01 and 0.003, but the rn, respectively CMC-curve, increases at rank $n = 2, 3, 4, 7, 8, 9, 10$. Furthermore, the mAP improves by 0.011 (see Table II Model-No. 6).

Changing the batch size from 64 to 16 improves the rn, respectively the CMC-curve. The r1 and r5 increase by 0.016 and 0.004. The mAP gets worse by 0.007 and the training time increases by 14 minutes, but this is negligible (see Table II Model-No. 7). In addition, the advantage of being able to train on the Jetson TX2 Board is important.

Then the epochs and milestones are halved which lead to an improvement in the r1 and r5 of 0.002 and 0.001 as well as the mAP by 0.008 (see Table II Model-No. 8). The course of the CMC-curve is improved up to rank $n = 5$ and has only marginal deviations up to rank $n = 10$ (see Fig. 6 Model-No. 8). The results are better by a halved training time.

The change of the model architecture reduces the accuracies marginal, r1 and r5 by 0.005 and 0.001 and the mAP by 0.01, but leads to a faster model. The model speeds up between 2.5 and 2.9 times. Furthermore, the training time is reduced by additional 40 minutes (see Table II Model-No. 9).

C. Model Acceleration

The model will be further optimized and accelerated by the use of the Python API torch2trt [35] from TensorRT (trt) [36]. It makes the application of the model, loading the image + throughput, 2.33 times faster and the throughput of the model between 6.1 and 15.3 times faster (see Table IV).

TABLE IV
TORCH2TRT COMPARISON FOR MODEL-NO. 9 TO SPEED UP.

Model-No. 9	Method	Duration	FPS
w/o trt	Throughput[a]	0.019 - 0.023	43.5 - 52.6
w/o trt	+ Image loading[b]	0.028	35.7
w/ trt	Throughput[a]	0.0015 - 0.0031	322.6 - 666.7
w/ trt	+ Image loading[b]	0.012	83.3

[a]The throughput is calculated as min and max values of 100 loops.
[b]The + Image loading is computed using 3552 images.

VIII. Real-Time Application

In this section, the application, which runs in real-time using the re-ID model from Sec. VII, is described and evaluated.

A. Application

First of all, the application for the person re-ID loads all persons who shall be recognized from a gallery. Then they are run through our evaluated re-ID model and the penultimate layer of the model, the GAP-layer, outputs for each image a 1 x 512 vector. These vectors are stored for every image of every person in the gallery as matrix $\mathbf{H}_g$ and has the form m x 512 with m being the number of all images in the gallery. This only happens once in the beginning (see Fig. 8 red frame).

Next, the persons in the current image of the depth camera are extracted. This is done the same way as the dataset creation, but without saving the image (see Fig. 5). Up to three persons with at least twelve keypoints are now processed individually (see Fig. 8 green frame). First, the extracted image, respectively query image, is run through the re-ID model and outputs the vector $\mathbf{h}_q$ of length 512, the same way as every gallery image. In the next step, the similarities are calculated. This is done using the CosineSimilarity [37] which returns a vector $\mathbf{d}_m$ with $m =$ as the number of all images in the gallery. The values in the vector are between zero and one. These are the probabilities of a match with the query image. The vector $\mathbf{d}_m$ is now sorted from best fit to worst fit. Now d_1, the closest match between the query image and a gallery image, has to be greater than or equal to the r1 value of 0.95. This value depends on the achieved r1 of the evaluated model (see Table II Model-No. 9). If this is the case, the ID-Name of this gallery image is sent together with the coordinates of the bounding box to a rosnode. If it is smaller, the ID-Name *unknown* is sent with the coordinates of the bounding box.

B. Application Evaluation

Unlike the model evaluation, in which the rn and mAP value for verification were taken, the evaluation of the application is done by calculating the precision, the recall and the f1-score of the classification-task because of the newly added class *unknown*.

The application was evaluated with three groups of three recordings, resulting in 30 seconds 10 Hz rosbag files in front of the open space scenario. The persons ran around to provoke overlaps. The three groups of three recordings are classified as positive, for persons who are in gallery, and negative, for persons who are not in gallery which should be classified by the application as *unknown*.

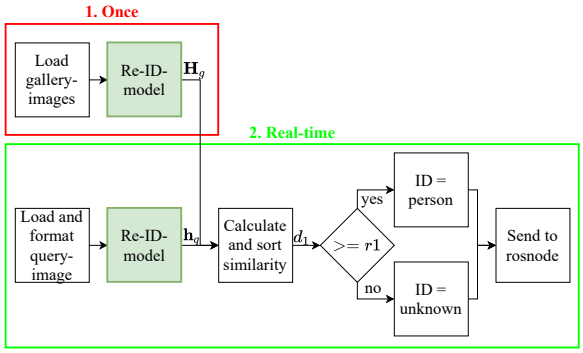


Fig. 8. Real-time application process.

First, the application calculates for every image from the recording, the recognized person and saves it with the ID-Name. This happens for all three recordings as seen in Table V. The values in bold correspond to the IDs to be recognized. In the case of the negative test it is the ID unknown. Cells with the value zero are empty for a better representation. In the negative test, IDs 07, 08 and 16 are added together as ID unknown, because these persons do not appear in the gallery. The precision, recall and f1-score are calculated using these values. The value of the recall and f1-score are similar in the positive and negative test, with the exception of the outlier of ID 02. Without the outlier, they are between 0.6 and 0.88 for recall and 0.75 and 0.94 for f1-score. The reason for this is the threshold with the value 0.95, which depends on the r1 from Model-No. 9 (see Table II). The threshold is not changed, because the values of the positive and negative test are, despite ID 02, similar. Furthermore, the outlier, ID 02, is recognized 37 times as unknown. This means that this ID is not assigned to any other person.

The two positive as well as the one negative test are shown in Fig. 9, Fig. 10 and Fig. 11.

IX. Conclusion

This paper proposes a real-time person re-ID API on a mobile robot, for a closed- and open-world setting, using only the IR gray value image of the PicoFlexx depth camera. Our API can re-identify one person around 16 FPS, two at around 13 FPS and three at around 10 FPS, at the same time. Therefore, we created a dataset by applying image processing techniques e. g. binarization to the extracted pose of a person from the IR gray value image. Further, we adjusted the model with common methods and also sped it up.

The NNs were evaluated by the metrics for a closed-world problem [38], as usual. However, the evaluation of the real-time application, with an added threshold, was done using classification metrics. With this threshold the model can be used for an open-world problem [38], which will differentiate between known and unknown persons.

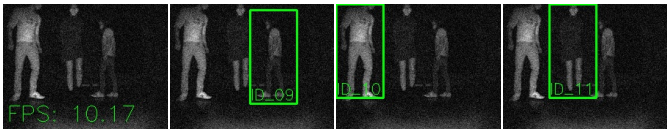
All models were trained on one Tesla V100 GPU of the NVIDIA DGX2-Server [39]. The model acceleration and real-

TABLE V
COMPARISON OF THE EVALUATION RESULTS OF THE THREE GROUPS OF THREE RECORDINGS. EMPTY SPACES ARE ZERO FOR A CLEANER OVERVIEW.

	Positive IDs [image_cnt/recording]								Negative IDs [image_cnt/recording]				
Person-ID	01	02	03	Precision	09	10	11	Precision	07	08	16	$\sum$	Precision
Unknown	19	37	8		20	8	11		**40**	**29**	**29**	**98**	1.0
00											1	1	
01	**36**			1.0						1		1	
02		**10**		1.0	1					1	10	11	
03			**81**	1.0									
09		1			**102**			1.0		1	6	7	
10			1			**82**		1.0					
11							**21**	1.0					
21						1			5			5	
24									2			2	
28			12			3							
30		1							12	1	2	15	
$\sum$	55	49	102		123	94	32		59	33	48	140	
Recall	0.65	0.20	0.79		0.83	0.87	0.66		0.68	0.88	0.6	0.7	
F1-score	0.79	0.34	0.89		0.91	0.93	0.79		0.81	0.94	0.75	0.82	

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

Fig. 9. The positive recording of ID 01, 02 & 03. (a) is the raw image with the FPS for the API. (b) to (d) are the detected persons with drawn ID.



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

Fig. 10. The positive recording of ID 09, 10 & 11. (a) is the raw image with the FPS for the API. (b) to (d) are the detected persons with drawn ID.

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

Fig. 11. The negative recording of ID 07, 08 & 16. Two of three identified correctly as unknown. ID 07 is falsely re-identified as ID 30 and should be identified as unknown. (a) is the raw image with the FPS for the API. (b) to (d) are the detected persons with drawn ID.

time application were evaluated on the mobile robot, respectively on the Jetson TX2 Board. Furthermore, this work was using the Robot Operating System (ROS) [40] to subscribe to the camera topic and to publish the results to a rosnode.

X. DISCUSSION

The results of this paper are not limited to the self created dataset, they can be transferred to other groups of persons as well as other mobile robots. However, due to the COVID-19 pandemic, the dataset with 31 persons is quite small and not divers. It should be extended to an amount of people similar as in the MARKET1501 dataset. The variance of persons could be increased using persons with a wide range of clothing, age, skin color, gender as well as healthy and physically impaired persons. The last mentioned group of persons is important because the postures of them can vary greatly. The skin color is of greater importance, because the state-of-the-art models for object recognition have a higher precision on lower Fitzpatrick skin types than higher types [41]. Additionally, the dataset should contain persons wearing the same clothes, e.g. work coat, to test the usability for hospitals.

Further, our work can be converted to other depth cameras without any problems. RGB cameras can also be used, but then two minor adjustments must be made. On the one hand, during the extraction of persons the binarization and the opening have to be deleted. On the other hand, the RGB image has to be transformed into a gray value image. The hardware can also be changed, but should correspond to the Jetson TX2 Board, otherwise the real-time (10 FPS) could not longer be achieved.

XI. OUTLOOK

We are still working on an ablation study focusing on our model and the techniques for the dataset creation.

In addition, we will evaluate the existing model for other camera perspectives. Further improvements to the re-ID model could result from the addition of the BNNeck and the center loss from [3]. These were not used in this work because in the basic model, Model-No. 1, the loss converged very quickly and the classes were already separated very pronounced. The influence of the WU and the LSm should be examined more closely, because [3] does not include the CMC-curve and the LSm has a negative influence of our model.

Further, we are planning to extend our dataset in research projects, which will be done in a hospital and logistics.

ACKNOWLEDGMENT

This work is based on the results of a master thesis. The authors would like to thank all people who made themselves

available for the dataset, despite the corona pandemic. Attention was paid to compliance with the corona protective measures.

REFERENCES

[1] S. Kautsar, B. Widiawan, B. Etikasari, S. Anwar, R. D. Yunita, and M. Syai'in, "A simple algorithm for person-following robot control with differential wheeled based on depth camera," in *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, 2019, pp. 114–117.

[2] P. Iyer, S. Tarekar, and S. Dixit, "Hand gesture controlled robot," in *2019 9th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-19)*, 2019, pp. 1–5.

[3] 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 (CVPRW)*, 2019, pp. 1487–1495.

[4] 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*, K. Nasrollahi, C. Distante, G. Hua, A. Cavallaro, T. B. Moeslund, S. Battiato, and Q. Ji, Eds. Cham: Springer International Publishing, 2017, pp. 1–11.

[5] V. O. Andersson and R. M. Araujo, "Person identification using anthropometric and gait data from kinect sensor," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. Austin, Texas: AAAI Press, 2015, p. 425–431.

[6] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2707–2714.

[7] paperswithcode. (2021) Person re-identification on market-1501. [Online]. Available: https://paperswithcode.com/sota/person-re-identification-on-market-1501

[8] 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)*, 2015, pp. 1116–1124.

[9] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 152–159.

[10] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 688–703.

[11] 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*, S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds. London: Springer London, 2014, pp. 161–181.

[12] 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)*, 2014, pp. 5644–5651.

[13] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," in *Computer Vision – ECCV 2012. Workshops and Demonstrations*, A. Fusiello, V. Murino, and R. Cucchiara, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 433–442.

[14] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 479–485.

[15] 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)*, 2016, pp. 770–778.

[16] F. M. Hafner, A. Bhuiyan, J. F. P. Kooij, and E. Granger, "Rgb-depth cross-modal person re-identification," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019, pp. 1–8.

[17] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2588–2603, 2017.

[18] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool, *One-Shot Person Re-Identification with a Consumer Depth Camera*, 01 2014, pp. 161–181.

[19] 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*, 05 2018, pp. 29–41.

[20] 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)*, 2014, pp. 5644–5651.

[21] K. Koide, J. Miura, and E. Menegatti, "Monocular person tracking and identification with on-line deep feature selection for person following robots," *Robotics and Autonomous Systems*, 02 2020.

[22] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1302–1310.

[23] Z. Sha, Z. Zeng, Z. Wang, Y. Natori, Y. Taniguchi, and S. Satoh, "Progressive domain adaptation for robot vision person re-identification," 10 2020.

[24] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," 11 2017.

[25] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3064–3074, 2020.

[26] J. Chen, J. Chen, Z. Wang, C. Liang, and C.-W. Lin, "Identity-aware face super-resolution for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 645–649, 2020.

[27] W. Ruan, W. Liu, Q. Bao, J. Chen, Y. Cheng, and T. Mei, "Poinet: Pose-guided ovonic insight network for multi-person pose tracking," in *Proceedings of the 27th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2019, p. 284–292.

[28] J. Zhang, C. Xu, X. Zhao, L. Liu, Y. Liu, J. Yao, and Z. Pan, "Learning hierarchical and efficient person re-identification for robotic navigation," *International Journal of Intelligent Robotics and Applications*, vol. 5, no. 2, pp. 104–118, apr 2021.

[29] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[30] T. Kirks and J. Jost, "Mensch-Technik-Interaktion in Industrie 4.0 Umgebungen am Beispiel von EMILI," in *Handbuch Industrie 4.0*, B. Vogel-Heuser, T. Bauernhansl, and M. Ten Hompel, Eds. Wiesbaden: Springer, 2019, pp. 1–11.

[31] pmdtechnologies. (2021) Get the most flexxible 3d time-of-flight development kit. [Online]. Available: https://pmdtec.com/picofamily/flexx/

[32] NVIDIA. (2021) Harness ai at the edge with the jetson tx2 developer kit. [Online]. Available: https://developer.nvidia.com/embedded/jetson-tx2-developer-kit

[33] NVIDIA. (2021) trt_pose. [Online]. Available: https://github.com/NVIDIA-AI-IOT/trt_pose

[34] 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)*, 2020, pp. 2664–2673.

[35] NVIDIA. (2021) torch2trt. [Online]. Available: https://github.com/NVIDIA-AI-IOT/torch2trt

[36] NVIDIA . (2021) TensorRT developer guide, NVIDIA docs. [Online]. Available: https://docs.nvidia.com/deeplearning/tensorrt/developer-guide/index.html

[37] Torch Contributors. (2021) COSINESIMILARITY. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.CosineSimilarity.html?highlight=cosine%20sim

[38] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[39] NVIDIA. (2021) Nvidia dgx-2. [Online]. Available: https://www.nvidia.com/de-de/data-center/dgx-2/

[40] ROS Contributors. (2021) ROS. [Online]. Available: https://www.ros.org/

[41] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," *CoRR*, vol. abs/1902.11097, 2019. [Online]. Available: http://arxiv.org/abs/1902.11097