Int. Symposium on Robotics (ISR), Munich, Germany, pp. 141-148, VDE Verlag 2016

May I be your Personal Coach? Bringing Together Person Tracking and Visual Re-identification on a Mobile Robot

Tim Wengefeld,
Markus Eisenbach,
Thanh Q. Trinh,
Horst-Michael Gross [1]

Ilmenau University of Technology,
Neuroinformatics and Cognitive Robotics Lab,
PF 100565, 98684 Ilmenau, Germany.
fax: +493677691665, e-mail: tim.wengefeld@tu-ilmenau.de

Abstract

Mobile robots following and guiding stroke patients during their rehabilitation program are in the focus of our research in rehabilitation robotics. To be able to act autonomously, it is crucial for the robot to extract long and precise movement trajectories of the patients. But already keeping track on one specific person in a crowded dynamic environment is inherently hard, since multi-sensor tracking as well as appearance-based re-identification are challenging tasks in real-world environments. Therefore, we aim for developing a coupled person tracking system that combines user tracking by spatial proximity with appearance-based user recognition. We analyzed all subcomponents of such a system and identified four essential parts, that significantly influence the overall performance. We show, that it is essential, to (1) accurately detect all persons in scene, (2) track people as long as no ambiguities occur, (3) visually re-identify the user otherwise, and (4) reduce the search space for re-identification to just relevant hypotheses using spatial proximity as criterion. In our experiments, we show, that by addressing all these aspects, our system significantly outperforms each approach, that excludes just one of these important parts.

1 Introduction

A recent trend in late stages of stroke patients' rehabilitation is the so called self-training. It includes the unrestricted exploration of patients in the clinic, re-training both their physical and cognitive skills. Our robot-assisted training approach aims at supporting this process by guiding the patients to pre-defined memorable places in the rehab center. The robot shall also serve as companion, to address the patients' insecurity and anxiety ("Am I able to do that", "Will I find my way back?") which are possible reasons for not performing or neglecting self-training.

In the following, a sketch of a typical walking training session described in [1] is given. The training session is initiated either by the robot by sending a text message to the phone in the patient's room or optionally by the patient, who can call the robot by telephone. The robot then autonomously drives to the patient's room and takes a non-blocking waiting position at the door. It observes the corridor for a person emerging from the respective room door and starts a verbal greeting. Then the patient logs in via touching a start button on the screen as asked by the robot. After that, the re-identification module learns the patient's current appearance.

Based on the training progress in preceding training sessions, the patient can choose the path for the upcoming walking training. Then, the robot follows the patient in a polite distance. On the way along the walking session, the robot points out possibilities for having a rest and also remarks orientation features (e.g. pictures on the wall, plants, etc.) which are helpful for finding the way back on longer tours. Thus, the patient can either go on or take a seat to revive. When the user appears to be exhausted, the robot suggests to finish the training and going back to the patient's room.

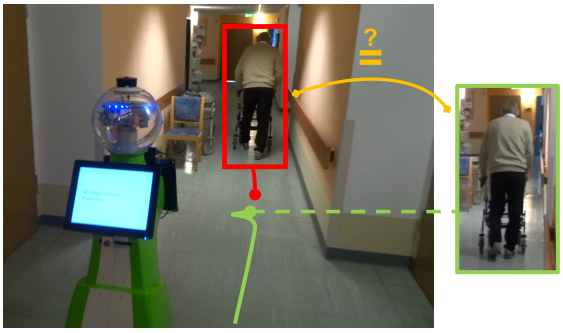

Figure 1: Typical situation of a mobile robot following a patient, where a track is cut. The robot has to decide, if the new found track still corresponds with the user.

In this scenario, to be able to act autonomously, it is crucial for the robot to keep track of the patient as precise and long as possible. In the field of mobile robotics, this problem is often treated as tracking approach. Different cues are used to enable long person tracks. However, sometimes tracks have to be cut due to temporal fully occlusions, or to avoid ID switches, which would result in a wrong navigation behavior of following wrong persons. In these situations, in recent work, tracks are either connected by spatial distance of track hypotheses or by vi-

[1] This work has received funding from the German Federal Ministry of Education and Research as part of the projects ROREAS under grant agreement no. 16SV6133, SYMPARTNER (16SV7218), and 3D-PersA (03ZZ0408).

sual re-identification. In this paper, we show, that none of them is capable of creating long person tracks on its own, due to limitations of both approaches: The assignment by geometrical proximity cannot avoid ID switches, in case of other persons nearby. On the other hand, vision-based re-identification often faces problems in case of changing illuminations. Therefore, we propose a coupled system, that decides in each situation, which merging approach is appropriate, and thus, tries to overcome the drawbacks of each approach alone.

2 System Design for Tracking

Person tracking on a mobile robot as well as visual re-identification are challenging tasks in crowded real-world environments. Our System aims to bring both fields of research together and shows that they can benefit from each other.

2.1 System Overview

Our target platform is a SCITOS G3 which was designed for the clinical training scenario. Two of our experimental platforms, named Roreas and Ringo, are shown in Fig. 2. The relatively small footprint size of 45×55 cm with a total height of 1.5 meters allows them to interact safely under limited space conditions. For navigation tasks, like path planning and localization, the robots are both equipped with two 270° SICK S300 laser range finders, mounted 20 cm above ground. Additionally, three Asus RGB-D cameras (two in driving direction and one backwards) are used to avoid obstacles outside the lasers' field of view. The sensors used for tracking and re-identification are the laser range finders and an omni-directional color vision system. While the system on Roreas consist of four μEye122xLE-C cameras, Ringo possess an experimental system with six cameras and fisheye lenses. The cameras were mounted on top of the robots' heads, to ensure a large field of view for the detection and tracking task. Both robots are equipped with an i7 quad core processor with 3.2 GHz exclusively for person detection and tracking. Therefore, even the highly

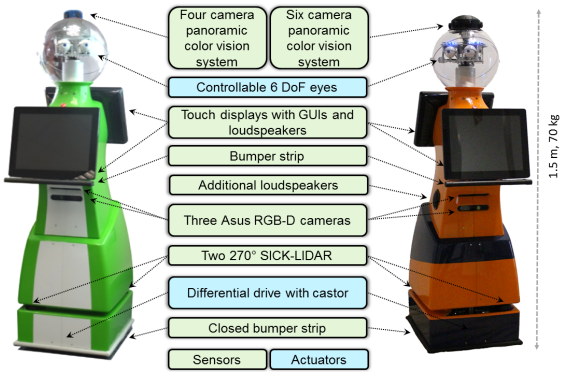

Figure 2: Both robot platforms used as walking assistants including their equipment used for person tracking, person re-identification, navigation and HRI.

computationally expensive approaches of our system can be computed in real time.

Fig. 3 gives an overview of the coupled person detection and tracking system. All sensor data are processed asynchronously by person detection modules, referred to as detection cues. The output generated by these cues is projected into a global world-coordinate system and fused by the tracking module to hypotheses which correspond with persons in the local surroundings of the robot. The tracklet association module decides which hypothesis corresponds with the person to guide or follow, which is than passed to the HRI module. In ambiguous cases, the visual re-identification module is asked to decide, which hypothesis corresponds with the current user.

2.2 Sub-Modules

In the following, we briefly describe the sub-modules of our tracking system.

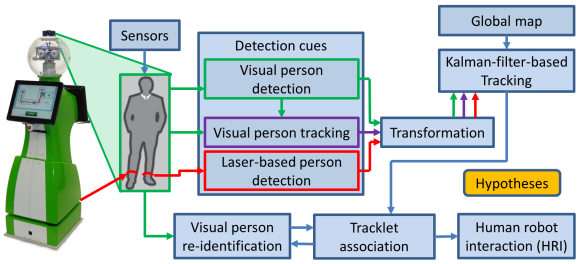

Figure 3: System overview of all sub-modules and their connections.

Human robot interaction (HRI): The scenario-specific HRI tasks are guiding and following stroke patients to different destinations in the clinic. These tasks are implemented as part of the navigation system, consisting of a Dynamic Window Approach (DWA) [2] and an E^* planner [3]. Therefore, the module receives the current user track and navigates the robot, taking the user's position into account. The robot adapts its speed according to its current user to keep a comfortable distance to the user. To realize a polite accompanying behavior, different tasks respecting people's personal space and detection of deadlock situations [4] are implemented. Therefore, a reliable detection [5] and tracking [6] of all people in the robot's vicinity is necessary. For more details on the navigation system, we refer to [7].

Laser-based person detection: Person detection in laser scans usually is based on the idea of detecting legs [8, 9]. In [5, 10], we have shown that it is possible to increase the detection performance significantly by using Binary Decision Trees as weak classifier instead of the previously used Decision Stumps. Furthermore, it is also possible to detect persons using walking aids like crutches or walkers and even people sitting in wheelchairs. Since the operational setting remains the same as in previous work, we use the detector of [5] and refer to [5, 10] for evaluation results.

Visual person detection: In addition to the laser detections, our system supports different replaceable visual de-

tection approaches. In this particular walking training scenario, persons occur in diverse ranges of appearance, e.g., diverse poses or occlusions at near scale where the legs cannot be perceived. Therefore, we integrated different detection approaches into our system which are able to detect full views of persons or parts of them. We evaluate the scenario-specific performance of all of these approaches in the experimental section of this paper.

1) VJ detector: The detector of Viola & Jones [11] utilizes haar-like features in a cascade of separate Adaboost classifiers. For this detector, we make use of three different models trained for faces, full-, and upper-body detections.

2) HOG detector: To detect people by the shape of their body, we use a full body detector based on Histograms of Oriented Gradients (HOG) as features with a linear Support Vector Machine (SVM) as classifier [12]. For this detector we just use a full-body model.

3) Multi-Class-HOG detector: To detect the upper-body of a person with HOG features, we make use of the detector described in [13]. This detector uses a tree of linear SVMs to detect persons while estimating their orientations.

4) Part-HOG detector: The Deformable Part Model (DPM) described in [14] uses HOG features with a root-filter similar to [12] in combination with smaller patches ("parts") arranged in a star-structure around the root-filter. Therefore, it is able to use the high discriminability of the full-body shape, while handling occlusions sufficiently. In favor of its real-time capabilities, we use the fast implementation of [15].

Visual person tracking: For tracking directly within image space, we apply a template-based visual tracker [16]. Its purpose is to handle video frames, that had to be skipped by the time-consuming visual detectors due to real-time requirements. This tracker processes frames much faster than in real-time at low computational cost, and thus can additionally process those frames, that could not be analyzed by a visual detector.

Kalman-filter-based tracking: Our tracking approach, is based on [6, 17]. For hypotheses tracking, we use a 7D linear Kalman filter with three dimensions for the global position, three dimensions for the velocity and one dimension for the person's upper body orientation, which is visually detected [13]. Additionally, we use information from the model of the environment for hypotheses pruning, so implausible hypotheses inside walls or obstacles are removed.

Tracklet association: To get a long user-specific track, multiple shorter tracklets need to be connected. For this tracklet association, we advanced our conservative approach from previous work [18]. Conservative refers to the fact, that the tracklets had been connected by their spatio-temporal proximity, only if the association was unambiguous without the risk of ID switches. Otherwise, tracklets of nearby persons were cut, and new tracklets were initialized by the Kalman-filter-based tracking. In spite of this conservative ID association, certain events are likely to cause ID switches or wrong tracks, as for example persons entering or leaving the robot's field of view, people walking close to each other, people not being observed due to occlusions, and so on. These events need to be handled with great care, since they cannot be resolved later on. To avoid ID switches in these cases, appearance-based visual person re-identification is applied to resolve the ambiguity. To support this, the search space for re-identification is reduced to only those persons who could have caused the ambiguity (see Fig. 4). To this end, a decision tree is applied to consider track IDs, re-identification matching scores, and spatial distances to the user's predicted position to decide which of the candidates is the current user of the robotic coach.

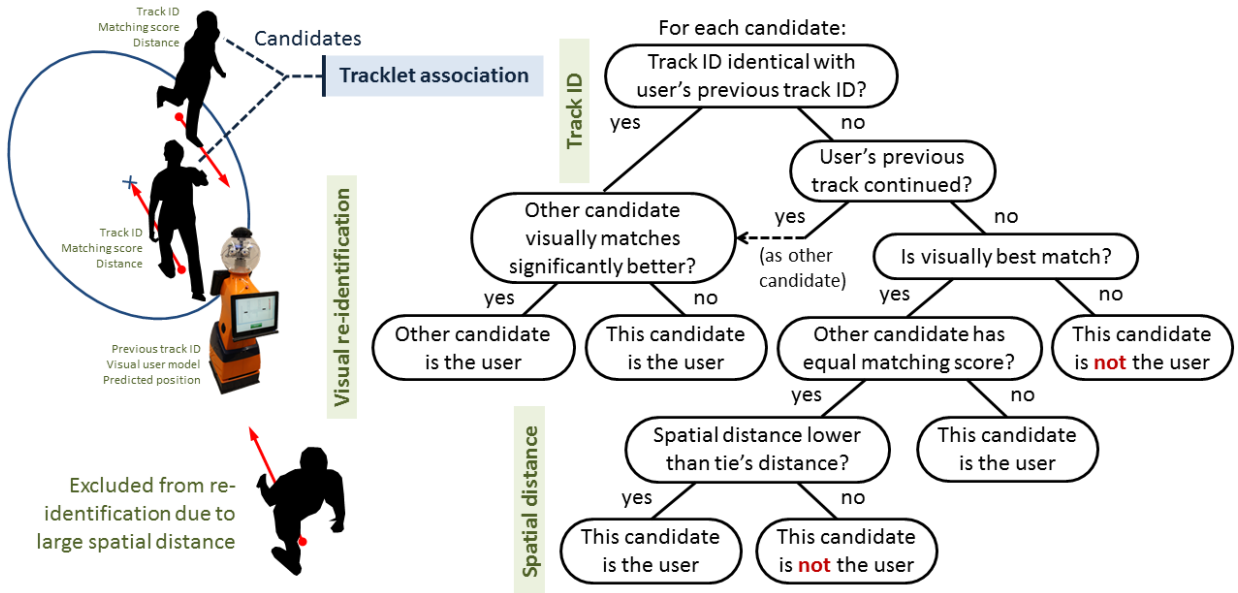

Figure 4: For tracklet association a decision tree is applied to identify the user from possible candidates in the robot's surroundings.

Additionally, tracklets of people that are confidently identified as non-user (due to a large difference in the matching score of the visual re-identification in comparison to the correct match) are excluded from future comparisons, to speed up processing and avoid later mis-matches in case of changing illumination or other environmental influences that affect the person's appearance.

Person re-identification: For identifying the user in various poses and from different viewpoints, we have decided in favor of a non-biometric, appearance-based re-identification approach [19]. Weighted color histograms from upper and lower body (wHSV) and Maximum Stable Color Regions (MSCR), both features of the SDALF approach [20], are used to describe the user's current appearance. The user template, learned during registration (see Sect. 1), is compared with reasonable hypotheses by applying a learned distance metric, that compensates for changing illumination and partial occlusions. Therefore, we decided in favor of the kernel-LFDA distance metric learning method (with modifications described in [19]), as it showed very good performance on many datasets in the extensive evaluation of Xiong *et al.* [21]. Then, multiple features are fused at score-level using the PROPER approach [22]. Finally, the decision, which person hypothesis represents the current user, is made by a probabilistic voting, considering distance scores of multiple observations per track and rankings in comparison to other tracks. Fur further details, we refer to [19].

Communication and data exchange: As communication infrastructure between the modules, we use the robotics middleware framework MIRA [23]. To handle components' different processing times, this framework allows the modules to dynamically exchange data of already past and present points in time as needed. Therefore, we use the callbacks, where every module receives a notification when new data arrives, e.g., when a new image is ready for processing.

Figure 5: Typical examples from our dataset with exemplary ground truth and detection shown. Top: the robot follows a patient walking with a crutch. Bottom: two scientific staff members walk behind the robot to observe its behavior. Ground truth is shown as a green bounding box, a detection as a blue one, and a ground truth which was set to be ignored due to heavy occlusions as red one. Detections matching ground truth labeled to be ignored neither count as true positives nor false positives. If none of the detections matches this kind of ground truth, it is also not count as false negative (see [24]).

3 Experiments

To evaluate the proposed person tracking system, we performed numerous tests with the robot in the rehab clinic, where the patients are to be coached during their walking training.

3.1 Experimental Setting

To evaluate the visual detection cues of our system, we manually labeled data of two complete walking exercises with real patients by annotating tracks in the global coordinate system. The statistics of this dataset can be seen in Tab. 1. Our detection dataset contains 4,985 images and 4,434 labeled persons. Therefore, it is comparable with the standard pedestrian detection benchmark data set Caltech, which includes 4,024 images and 1,002 person labels in reasonable size (see below) in the test set. During these exercises, 30 different persons were present in the surroundings of the robot. While four of these persons were scientific staff members and occur in both datasets, the rest were patients and clinical staff members.

The ground truth data for the visual detections were automatically generated by back-projecting the hypotheses to the image plane. When persons were occluded from walls or other persons, and less than 75% of their appearance was visible, the corresponding ground truth was manually set to be ignored (see Fig.5). Ground truth boxes labeled to be ignores are handled as defined in the pedestrian detection benchmarking protocol [24] of the widely used Caltech dataset. So detections in this critical areas of the images neither count as true positive nor as false positive. If none of the detections matches this kind of ground truth, it is also not count as false negative. In those cases, where the robot's localization was too inaccurate, the corresponding frames were also discarded manually in order to avoid projection errors. The dataset includes images from all four cameras of the panoramic camera system shown in Fig. 2 left, recorded with a frame rate of 2 Hz. To decide which of the different visual detection approaches is sufficient for our application, we used the evaluation protocol defined in [24]. Detections are considered a match with ground truth if the ratio of intersec-

tion to union of the two bounding boxes is 0.5 or above. This means, their overlap needs to be about 66%. If multiple detections match the same ground truth, the decision is made by the detections' confidence scores. Every detection can be assigned to one ground truth at most and vice versa. Detections and ground truth boxes that are not assigned, are counted as false positives and false negatives respectively. In [24], all bounding boxes that exceed image borders are excluded from evaluation. We follow this protocol, except we do not ignore persons which exceed the lower image border, since these detections result from persons standing near the robot and interacting with it. So this is an important part of our scenario.

We use two subsets for the evaluation. Following the protocol, the "reasonable" subset as defined in [24] includes ground truths with a height of 50 pixels and larger while at least 65% of their bounding box is unoccluded. All other bounding boxes are labeled as "ignore", and thus, they are excluded from evaluation (see above). This subset is used to evaluate the overall detection capabilities, since it contains a good cross-section of person appearances during the whole training exercise. The "partially occluded" subset as defined in [24] includes just ground truths where 1% – 35% of their bounding box is occluded. Unoccluded ground truth bounding boxes are ignored. Since partially occluded ground truths are mainly caused by persons standing in close proximity to the robot, where the legs were out of the camera's field of view, this subset is a benchmark for the tracking capabilities during Human Robot Interaction. In the following, this dataset including two exercises is referred to as detection dataset.

	exercise 1	exercise 2	overall
duration	4:11 min	7:42 min	11:53 min
distance	98 m	101 m	199 m
other persons	10	20	30
walking aid	walker	crutch	various
images	1850	3135	4985
reasonable gt	1747	2687	4434
unoccluded gt	1356	1919	3275
partially occl. gt	391	768	1159

Table 1: Statistics of our detection benchmark dataset. gt = number of ground truth labels

For evaluation of the complete tracking system, we extend this dataset by eight additional sequences. Therefore, it contains walking trainings of four different patients using various walking aids. The overall runtime is 52 minutes, in which the robot drove a distance of 929 meters and encountered 141 other people, including technical and clinical staff, as well as other patients.

3.2 Visual Person Detection

For adequate user recognition, robust person detection is essential (see [19]). Therefore, we have to determine the best person detectors for the addresed scenario first.

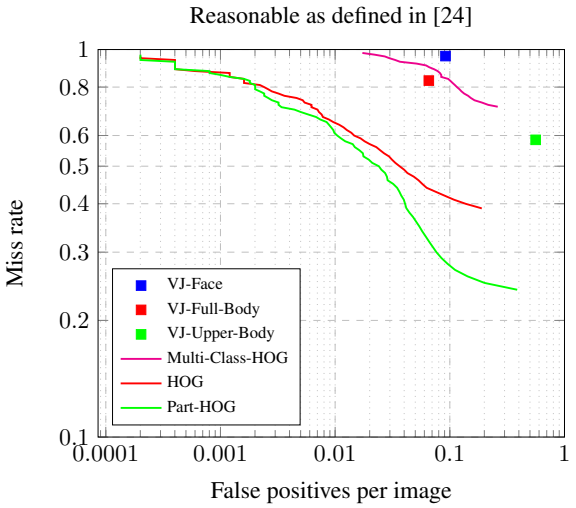

Figure 6: Detector performances on the reasonable image subset (person height > 50 pixels, occlusion max. 35%), we use as benchmark for the detection performance of the whole walking training.

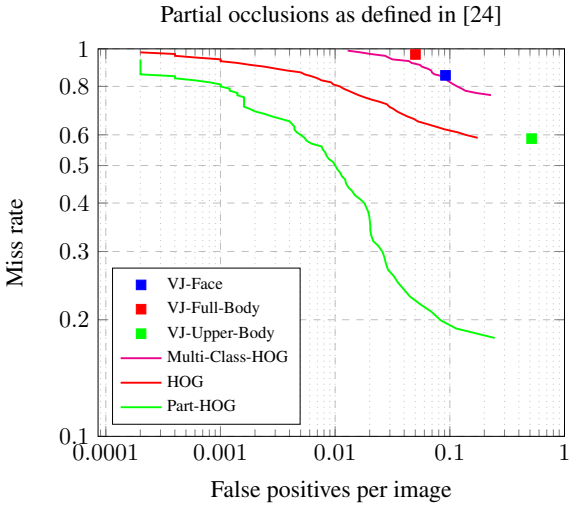

Figure 7: Detector performances on the occluded subset, which we use as a benchmark for the detection performance when Human Robot Interaction is done.

A comparison of the detection results is given in Fig. 6 and Fig. 7. We plot the average miss rate against the average false positives per image on logarithmic axes.

The Viola-Jones detector equipped with the three different models for face, full-, and upper-body detections (*VJ-Face*, *VJ-Full-Body*, *VJ-Upper-Body*) under-performs on our dataset compared to the other approaches. The performance of the face detector increases on the occluded dataset, while the full-body detector performs worse. The upper-body detector has a lower miss rate, but one false positive detection per second in a single camera, which is not acceptable for this scenario. Please note that, since this is an cascaded approach, the results are just plotted as a single point in the diagrams.

The *Multi-Class-HOG* detector for detection and orientation estimation under-performs as well on both of our

datasets. This might be caused by the relatively small training dataset, since some persons get detected very well and others do not. Even on the occluded dataset, where an upper-body detector should perform better, this detector yields no superior results.

The *HOG* detector performs relatively well on the reasonable dataset (person height > 50 pixels, occlusion max. 35%) with 41.3% missed detections at an average of 1 false positive per 10 images, which is a common breakpoint to compare detectors. On the other hand, on the occluded subset the miss rate increases to 61.3% at the same false positves rate. This is unacceptable for our application, since the interaction with the robot is a critical part for user acceptance and, therefore, a robust person detection is crucial.

The *Part-HOG* yields the best results on the reasonable dataset (person height > 50 pixels, occlusion max. 35%), where just 27.6% of the persons were missed at an average of 1 false positive per 10 images. The performance is even better on the occluded subset, where the miss rate drops to 19.4%. This can be explained with the larger height of persons in close proximity to the robot. Therefore, the *Part-HOG* is our choice for the application and further experiments.

The runtime of the different detection approaches can be seen in Tab. 2. Since the four camera system we use in this paper is restricted by hardware to 2 Hz each, all of the detection approaches can compute in real-time without frame skips, using the exclusive PC for visual detections. Therefore, our visual person tracking approach is currently not used in the four camera setup at 2 Hz. However, the experimental panoramic camera system depicted in Fig. 2 right, consists of six cameras with frame rates of up to 15 Hz each. Therefore, visual tracking and experiments regarding the tracking quality at higher frame rates will become an important part in our future work.

	runtime	frame rate
VJ-Face	159 ms	~6 Hz
VJ-Full-Body	62 ms	~16 Hz
VJ-Upper-Body	265 ms	~4 Hz
Multi-Class-HOG	92 ms	~11 Hz
HOG	552 ms	~2 Hz
Part-HOG	495 ms	~2 Hz

Table 2: Runtime comparison on an Intel core i7 with 3.2 GHz. Please note that the runtime of the VJ detector strongly depends on the number of features used in the different models.

3.3 Person Re-Identification

To benchmark the scenario-specific re-identification performance, we recorded a new ROREAS dataset in the rehab clinic. During rush-hour times of two days, the robot frequently drove through the corridor where patient's walking exercises took place. Images of nearby persons were automatically detected and saved. Therefore, a total of 11,034 images, showing 207 different people, was collected. The person IDs were semi-manually labeled. To make the dataset more realistic, we eliminated similar appearances for each person automatically by clustering. Similar appearances most often result from person observations of the same track. These can be handled by tracking. In real-world applications, re-identification only handles more complicated cases where observations of different tracks are compared. After elimination, each person, for whom at least two different views were available, was added to the ROREAS dataset. It consists of 776 images showing 192 different persons with $2-10$ views each.

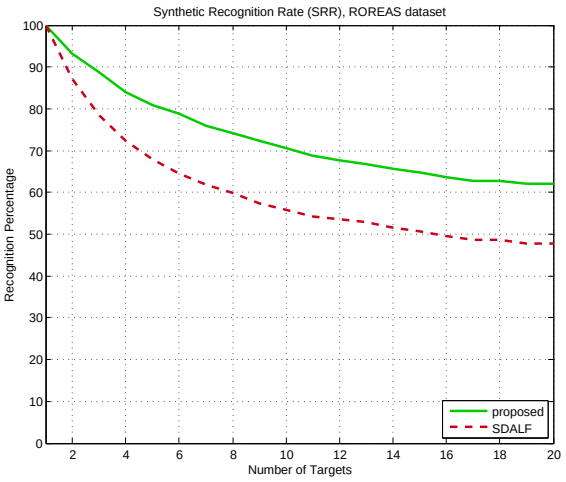

Figure 8: Synthetic recognition rate on scenario-specific dataset. Our re-identification approach [19] outperforms the popular method SDALF [20], using identical features.

Fig. 8 shows the SRR curve (Synthetic recognition rate) of our re-identification component in comparison to the SDALF approach [20], that extracts the same features. As described in [19], in comparison to [20], our approach applies metric learning to compare feature vectors. Therefore, the training set is preprocessed to ensure to learn a distance metric that compensates for changing illumination. Additionally score level fusion is used to combine the features instead of a manually designed fusion scheme. It is visible, that these modifications lead to a significant improvement of the re-identification rate. However, the SRR curve of our approach on this dataset is considerably lower than on other standard re-identification benchmark datasets. That means, appearance-based person re-identification on a mobile robot (ROREAS) is far more difficult than re-identification of pedestrians in multiple static cameras with disjoint views (standard benchmarks). The average recognition rate for two persons in the scene was 93 %, which is a typical number of persons for ambiguities in tracking. For six persons in the surroundings of the robot, the recognition rate drops below 80 %. Hence, situations where the user model has to be compared with lots of hypotheses should be avoided. Therefore, the search space should be reduced by the tracklet association module to only the persons that are observed near the provious observation of the user. In our scenario the search space is almost always reduced to two possible candidates. This is

why tracklet association is essential to accomplish robust autonomous tracking by combining different cues.

3.4 User Following and Guiding

Next, we evaluated the skills of our coupled system. Over a period of 52 minutes, the robot followed and guided four subjects. During the exercise, the robot drove a distance of 929 meters. 32 times the track was lost. Using the tracklet association and person re-identification modules, the amount of lost tracks could be reduced to 15. While using normal tracklets (without association) for user tests, just one complete exercise could be finished without manual intervention. With the coupled system, this amount could be increased to four out of ten. The average track length was extended from 29 meters to 61 meters.

Assigning cut tracklets to new hypotheses only by means of spatial proximity fails in the addressed scenario since erroneous assignments occur to often. Using re-identification without a tracklet association component performs much worse due to much more ambiguities and more people to compare with. Therefore, the combination of spatial tracklet association and visual re-identification is essential. Additionally, people wearing similar clothes, e.g., only slightly different gray tones with a large variance in multiple illuminations, can be distinguished much better using the tracklet association component, since they are compared only if they are nearby each other, and thus, are observed under similar lighting conditions.

The remaining issues are mainly caused by lighting changes between daylight near windows and artificial illumination within corridors. Thus, the patients' appearances change significantly, which complicates re-identification. A better performance would only be possible by using more advanced features or distance metrics that are applied in these situations to compensate for the differences in illumination.

4 Conclusion

We presented a coupled person tracking system that unites tracking by spatial proximity with appearance-based user recognition. The tracking aims to help the robot to keep track of patients performing their walking training in order to act as a personal coach. To be able to act autonomously, it is crucial for the robot, to extract long and precise movement trajectories of the patients. To achieve this objective, it is essential to:

- accurately detect all persons in scene in one ore more sensor cues,
- track people as long as no ambiguities occur,
- robustly re-identify the user out of a group of people if necessary,
- reduce the search space for re-identification to just relevant hypotheses using spatial proximity as criterion.

By addressing all these aspects, our system performs significantly better and increased fully autonomous exercises from 10% to 40%. The average track length has been doubled. Nevertheless, fully autonomy is still out of range, due to difficult lightings and a very crowded environment.

Acknowledgments

The authors wish to thank their partners of the ROREAS research consortium for their trustful cooperation allowing them to do robotics and HRI research in a challenging real-world scenario and environment, the "m&i Fachklinik" Rehabilitation Center Bad Liebenstein, the SIBIS Institute for Social Research in Berlin, and the health insurance fund Barmer GEK Wuppertal.

References

[1] H.-M. Gross, A. Scheidig, K. Debes, E. Einhorn, M. Eisenbach, St. Mueller, Th. Schmiedel, T. Q. Trinh, Ch. Weinrich, T. Wengefeld, A. Bley, and Ch. Martin. Roreas: robot coach for walking and orientation training in clinical post-stroke rehabilitation: Prototype implementation and evaluation in field trials. *Autonomous Robots*, pages 1–20, 2016.

[2] D. Fox, W. Burgard, and S. Thrun. The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 4(1):23–33, 1997.

[3] R. Philippsen and R. Siegwart. An interpolated dynamic navigation function. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3782–3789, 2005.

[4] T. Q. Trinh, Ch. Schroeter, and H.-M. Gross. "go ahead, please": Recognition and resolution of conflict situations in narrow passages for polite mobile robot navigation. In *Int. Conf. on Social Robotics (ICSR)*, pages 643–653. Springer, 2015.

[5] Ch. Weinrich, T. Wengefeld, M. Volkhardt, A. Scheidig, and H.-M. Gross. Generic distance-invariant features for detecting people with walking aid in 2D laser range data. In *Int. Conf. on Intelligent Autonomous Systems (IAS)*, 2014.

[6] M. Volkhardt, Ch. Weinrich, and H.-M. Gross. Multi-modal people tracking on a mobile companion robot. In *Europ. Conf. on Mobile Robots (ECMR)*, 2013.

[7] H.-M. Gross, K. Debes, E. Einhorn, St. Mueller, A. Scheidig, Ch. Weinrich, A. Bley, and Ch. Martin. Mobile robotic rehabilitation assistant for walking and orientation training of stroke patients: A report on work in progress. In *IEEE Int. Conf. on*

Systems, Man, and Cybernetics (SMC), pages 1880–1887. IEEE, 2014.

[8] K. O. Arras, O. M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2D range data. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3402–3407. IEEE, 2007.

[9] L. Spinello and R. Siegwart. Human detection using multimodal and multidimensional features. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 3264–3269. IEEE, 2008.

[10] Ch. Weinrich, T. Wengefeld, Ch. Schroeter, and H.-M. Gross. People detection and distinction of their walking aids in 2D laser range data based on generic distance-invariant features. In *IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, pages 767–773. IEEE, 2014.

[11] P. Viola and M. Jones. Robust real-time object detection. *Int. Journal of Computer Vision*, 4:51–52, 2001.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[13] Ch. Weinrich, Ch. Vollmer, and H.-M. Gross. Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2147–2152. IEEE, 2012.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[15] C. Dubout and F. Fleuret. Exact acceleration of linear object detectors. In *Europ. Conf. on Computer Vision (ECCV)*, pages 301–311. Springer, 2012.

[16] A. Kolarow, M. Brauckmann, M. Eisenbach, K. Schenk, E. Einhorn, K. Debes, and H.-M. Gross. Vision-based hyper-real-time object tracker for robotic applications. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2108–2115. IEEE, 2012.

[17] M. Volkhardt, Ch. Weinrich, and H.-M. Gross. People tracking on a mobile companion robot. In *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, pages 4354–4359. IEEE, 2013.

[18] A. Kolarow, K. Schenk, M. Eisenbach, M. Dose, M. Brauckmann, K. Debes, and H.-M. Gross. APFel: The intelligent video analysis and surveillance system for assisting human operators. In *IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 195–201. IEEE, 2013.

[19] M. Eisenbach, A. Vorndran, S. Sorge, and H.-M. Gross. User recognition for guiding and following people with a mobile robot in a clinical environment. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3600–3607. IEEE, 2015.

[20] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2360–2367, 2010.

[21] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Europ. Conf. on Computer Vision (ECCV)*, pages 1–16, 2014.

[22] M. Eisenbach, A. Kolarow, A. Vorndran, J. Niebling, and H.-M. Gross. Evaluation of multi feature fusion at score-level for appearance-based person re-identification. In *Int. Joint Conf. on Neural Networks (IJCNN)*, pages 469–476. IEEE, 2015.

[23] E. Einhorn, T. Langner, R. Stricker, Ch. Martin, and H.-M. Gross. MIRA – middleware for robotic applications. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 2591–2598. IEEE, 2012.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):743–761, 2012.