IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Hamburg, Germany, pp. 3600-3607, 2015

User Recognition for Guiding and Following People with a Mobile Robot in a Clinical Environment

Markus Eisenbach, Alexander Vorndran, Sven Sorge, and Horst-Michael Gross*

***Abstract*— Rehabilitative follow-up care is important for stroke patients to regain their motor and cognitive skills. We aim to develop a robotic rehabilitation assistant for walking exercises in late stages of rehabilitation. The robotic rehab assistant is to accompany inpatients during their self-training, practicing both mobility and spatial orientation skills. To hold contact to the patient, even after temporally full occlusions, robust user re-identification is essential. Therefore, we implemented a person re-identification module that continuously re-identifies the patient, using only few amount of the robot's processing resources. It is robust to varying illumination and occlusions. State-of-the-art performance is confirmed on a standard benchmark dataset, as well as on a recorded scenario-specific dataset. Additionally, the benefit of using a visual re-identification component is verified by live-tests with the robot in a stroke rehab clinic.**

I. Introduction

About 2-5% of all health related costs in the western developed nations originate from stroke disease patterns. Due to demographic change, the rate of stroke occurrences is expected to increase, while at the same time family structures are changing and cohabitation of different generations, providing possibilities for informal care, is receding. In effect, demand for rehabilitative follow-up care for stroke patients is increasing. As motor and cognitive learning are not passive processes, patients recovering from a stroke must play an active role in the rehabilitation process if improvement is to occur [1]. Against this background, a new trend in rehabilitation care is promising vast medical as well as economic potential – the so-called self-training. This finding is the context and the motivation for the research project ROREAS [8], which aims at developing a robotic rehabilitation assistant for walking and orientation exercising in self-training during clinical stroke follow-up care. The robotic rehab assistant is to accompany inpatients during their walking and orientation exercises, practicing both mobility and spatial orientation skills. It shall also 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.

The task of the robot is to follow patients during their walking exercises and, if necessary, guide them back to their room. This self-training is performed in the corridor of the rehab clinic. Therefore, many other people are present in the surroundings of the robot. To hold contact with the user, the robot must continuously track him. Furthermore, in cases of temporally full occlusions or in the case of ambiguities in tracking, the robot must be able to re-identify the user by its visual appearance.

Fig. 1. Person following in a clinical environment needs a re-identification component to succeed during rush-hour times.

II. Requirements and Challenges of Person Re-Identification on a Mobile Robot

Re-Identification of a person by a mobile robot is very challenging. This is due the need for a good user recognition in real-time while using only few computing capacity, leaving enough time for security related tasks, like collision avoidance, and computationally expensive tasks, like path planning. While meeting these real-time requirements is hard for visual algorithms, the task of re-identification becomes even more complicated due to lots of motion of the robot, which blurs the images, a very dynamical environment with many different lighting, and many objects, that trick state-of-the-art visual person detectors to false detections (e.g. wand lamps between two doors). Person detection is also complicated by many people using walking aids in our scenario. The narrow corridors of the building often lead to partially and temporal fully occlusions of the user by other people. The user may also be only partially visible, if he is near the robot due to the robot's head mounted camera. Additionally, the size of the image regions containing the user (and thus the resolution) varies a lot with the distance to the robot. Therefore, the person re-identification module has to be robust to image blur, varying resolution and illumination, occlusions, people with walking aids, and false detections.

Summarized, to be suitable for daily use, the person re-identification module has to satisfy the following require-

*This work has received funding from the German Federal Ministry of Education and Research as part of the ROREAS project under grant agreement no. 16SV6133.

M. Eisenbach, A. Vorndran, S. Sorge, and H.-M. Gross are with Neuroinformatics and Cognitive Robotics Lab, Ilmenau University of Technology, 98694 Ilmenau, Germany. markus.eisenbach@tu-ilmenau.de

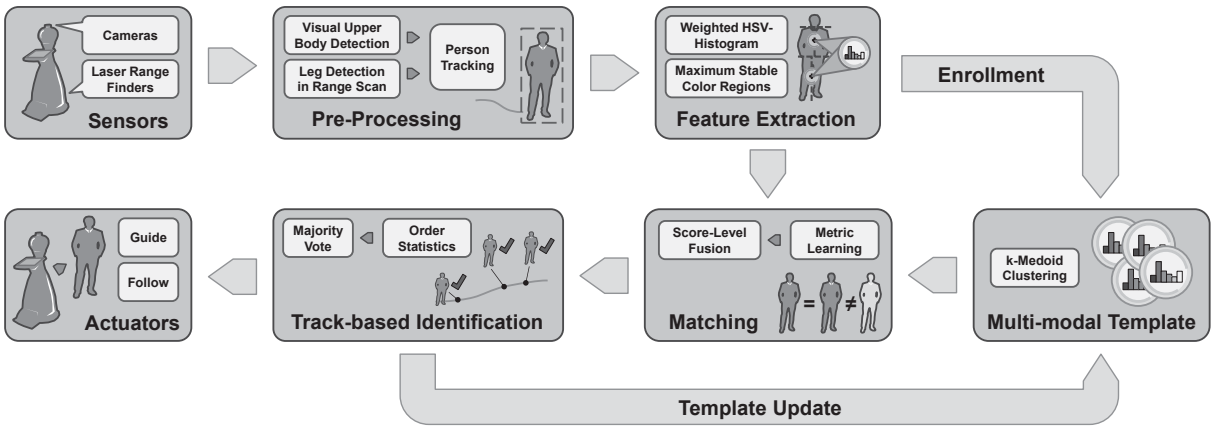


Fig. 2. Processing chain for fast person re-identification on a mobile robot.

ments: It has to start immediately when the user logs in (maximum 100ms for model training) and must then frequently compare each of the observed persons with the user model twice a second, using a maximum of ten milliseconds per person and 10% CPU at most. Despite the challenges listed above, the user should be recognized in at least 95% of all cases where re-identification is necessary. When a decision between two nearby hypotheses is hard to make, the robot should stop and ask the user to make itself felt. The robot should also stop when it loses contact to the user and cannot re-detect him or her.

III. Sub-Modules for User Re-Identification

To meet the requirements described in the previous section, we have chosen a re-identification workflow that is optimized regarding processing speed, but does not decrease recognition accuracy. Fig. 2 shows all sub-modules: First, all persons in the image have to be detected and tracked. Then, we describe their appearance by multiple complementary features. The current user is represented by a multi-modal template composed of features observed in the enrollment phase[1]. To reduce the size of the template, similar appearances are removed by clustering. To accurately compare persons in the scene with the user template, a distance metric that has been trained on a scenario-specific dataset is applied. To compose the matching results for the different features, score-level fusion is performed. Afterwards, the follow or guide hypothesis is chosen by a track-based decision considering multiple observations over time. Additionally, if the person can be identified securely during tracking as the current user, the template is updated.

A. Pre-Processing

Pre-Processing in terms of person re-identification is everything needed to get person-centered cropped sub-images from the camera output. This includes modules for person detection and tracking. For person detection, we implement two modalities: A laser-based leg detector (10 Hz) and a visual upper body detector (2 Hz). As laser-based method, we implemented the GDIF detector [17], which finds pairs of legs in the 2D range scan to robustly detect persons, even in situations where they use walking aids, like in our scenario. As visual detector we chose an orientation-based decision tree of upper body HOGs [16]. We decided in favor of an upper body detector, since the users are often close to the robot to interact via the touch-screen, and therefore, only their upper body is visible in the head-mounted camera. Since visual person detection is computationally very expensive, it is executed on a second on-board PC, with an energy saving CPU and only at 2 Hz. For data exchange between the two on-board PCs, we use the robotics middleware MIRA [3]. The results of both detectors are merged by a person tracker [15] based on Covariance Intersection and Kalman-filters for temporal tracking. New detections are checked to be valid hypotheses by being either detected by both modalities, or by one modality and moving through the scene (no-static criterion). Additionally, a global occupancy map is used to verify that the hypotheses are visible by the robot. All valid person hypotheses with visual reference are then passed to the re-identification module twice a second. This includes situations where the user may not be visually detected but still be tracked by the laser. Therefore, the re-identification module also needs to consider older visual observations for its (track-based) decision.

B. Feature Extraction

After all persons have been detected and tracked, the user can be identified. Therefore, we follow the idea of our previous work [10] and trust the person tracker to securely recognize the user as long as no ambiguous situations (e.g. low spatial distance to other persons) appear. Whenever ambiguities occur, the re-identification module has to decide which tracks should be connected to share the same person ID. Because the trackers' decisions are based on spatio-temporal proximity, the decision of the re-identification module has to apply a complementary modality to resolve ambiguities. Therefore, we use a visual approach to compare the appearance of people in the scene. Fig. 3 shows a listing of potential visual features to describe a person. The features used in our approach are highlighted.

Since persons are often observed in low resolution due to a greater distance to the robot, or the user can only be observed from behind (in the user following mode), we cannot use biometric features, like face, iris or ear, for re-identification.

[1]The enrollment phase begins with the login and lasts as long as the current user can be tracked securely.

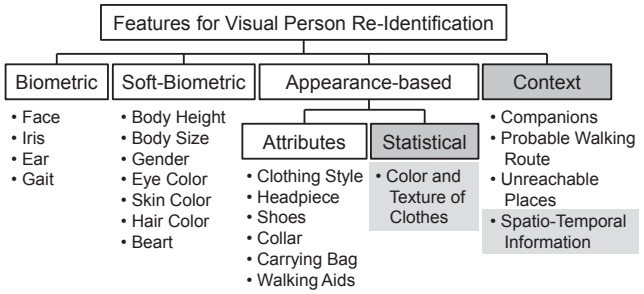


Fig. 3. Systematization of visual features for person re-identification. Features used in the proposed approach are highlighted.

Using gait recognition would be possible, but is computationally expensive and performs worse than appearance-based approaches (see [4] for an evaluation). The robust extraction of soft-biometric features and semantic attributes is also time consuming and may not be discriminatively enough (only in combination). Therefore, we decided in favor of features describing the color and texture of the clothes, that are fast to extract with low computationally cost. Additionally, we use spatio-temporal information of all persons to be seen in the scene as a kind of contextual information.

To describe the clothes' color, we utilize a weighted HSV-color-histogram (wHSV) and Maximum Stable Color Regions (MSCR), both components of the SDALF approach [6]. The wHSV features describe the appearance of upper and lower body by a histogram in HSV color space, where the weight of a single pixel is defined by a Gaussian kernel centered at symmetry lines found in the upper and lower body. MSCR describes the appearance of a person with stable color blobs in CIEL*a*b* color space. To describe the clothes' texture, we explored the widely used histogram of local binary patterns (LBP) [13]. The uniform LBP encode darker and lighter pixels around a center point, which is shifted to every possible position in the source image. Non-uniform LBP are represented by an extra histogram bin. However, we found, that in our scenario texture does not help to describe a person's appearance, since most of the extracted images are heavily blurred due to ego motion of the robot, and nearly all patients wear homogeneous colored clothes. Automatic feature weighting by the score-level fusion module has validated this hypothesis (see Sect. III-D).

Other appearance-based features, that show good re-identification performance, are BiCov [12], LDFV [11], and SDC [19]. We did not utilize these features, since none of them can be extracted in real-time.

The original implementation of the SDALF features (MatLab) does not fulfill the real-time restrictions described in Sect. II. Therefore, we implemented these features in C++ (optimized for fast extraction) and further examined each processing step for potential performance gains and improvements of re-identification rate. We found some approximations that speed up feature extraction and do not harm re-identification performance: The body partitioning and symmetry lines of the wHSV feature do not need to be very accurate and can be static. Also the foreground mask used by MSCR can be replaced by a static average person mask. Additionally, we reverted some approximations of the original implementation to improve recognition rates. Therefore, we use full tri-linear interpolated histograms for wHSV instead of marginalized ones. Finally, we performed cross-validated parameter tuning.

To compare extracted histogram features of two person hypotheses, we learned a distance metric. MSCR, however, does not suit metric learning due to its varying descriptor size. Therefore, we used the improved comparison method of Cheng *et al.* [2] for this feature. The features are combined by score-level fusion.

C. Metric Learning

Beside a good feature representation, we also need a proper scenario-specific distance metric to compare feature vectors, allowing for compensation of differences in illumination, image resolution, and other challenges listed in Sect. II. Therefore, we decided in favor of the kernel-LFDA method for distance metric learning, as it showed very good performance on many datasets in the extensive evaluation of Xiong *et al.* [18].

To get a proper metric, we do not just apply kernel-LFDA on the raw scenario-specific training dataset, but pre-process the data set. This improves the recognition performance significantly. The pre-processing is done in three steps: (1) We add an additional dataset with many persons to get a more generic metric. Therefore, we add half of co-training data to the scenario-specific training data set and use the other half as validation dataset. (2) We reduce the number of samples per person by k-medoid clustering, using only the cluster centers ($k \approx 5-8$). This is necessary to increase the inner-class variance by removing very similar samples that could be matched by any simple metric easily and would prevent the metric learning approach from spotting more advanced interrelationships. (3) We balance the training dataset. Therefore, we apply k-medoid clustering on samples of all persons and select only the cluster centers ($k = 500$ achieves best results). This groups similar outfits and thus reduces overrepresented clothing combinations, like black jackets with jeans or white outfits of clinical staff that would otherwise dominate the training dataset.

Done this, we choose the remaining 500 samples as kernel vectors, transform all samples into kernel space using a $\chi^2 - RBF$-kernel, and apply LFDA.

To compare two samples described by the same features, both feature vectors have to be transformed into kernel space and then projected to the k-dimensional LFDA-subspace (evaluations show $k = 40$ performs best). These projections can then be compared using the (squared) Euclidean distance, resulting in a single distance score.

D. Score-Level Fusion

Score-level fusion allows to fuse information at an abstract level. Therefore, it combines distance scores from different matched feature vectors. The goal is to get a fused score that is suitable to calculate a ranking. This is done in three steps: First, the scores for all features are normalized to make them comparable. The second step is to calculate the weight for each feature. In the third step, the fused score is calculated as

weighted sum. All these steps include only few and simple calculations. Therefore, score-level fusion can be performed very fast.

In order to apply score-level fusion, all scores have to be in the same value domain. This is usually achieved by normalizing the value range. In [5], we comparatively evaluated eight state-of-the-art normalization approaches. Most approaches need a huge amount of training data. This data must be distinct from the data for metric learning (see [5]). But to learn an adequate metric, we already used all of the training data. Thus, we decided in favor of a simple transformation-based approach, the z-normalization. Therefore, after normalization, scores have zero mean and standard variance in relation to the training data.

After all scores have been normalized to the same value domain, they can be combined. In order to calculate the fused score as weighted sum, a weight w is computed for each feature by using a test set of normalized distance scores. Common ways to calculate the weights are either weighting all features equally, or calculating weights as function of a performance measure derived from the genuine-impostor distribution (see Fig. 4). Re-identification errors are minimized, when the overlap of the two distributions decreases.

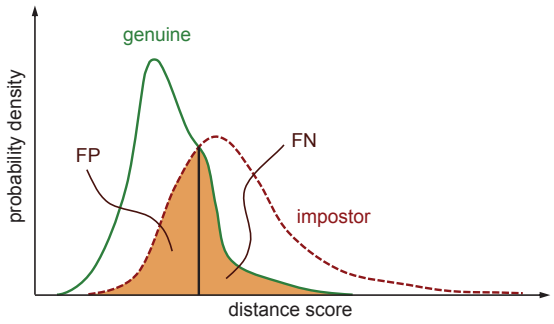


Fig. 4. Exemplary genuine-impostor score distribution (MSCR feature [6] on VIPeR dataset [7]). The highlighted area, where genuine scores (distance scores for image pairs that represent the same person) and impostor scores (distance scores for image pairs showing different persons) overlap, will produce errors (false positives (FP) and false negatives (FN)) when a threshold is chosen at the intersection point of genuine and impostor scores as marked.

In [5] we showed, that calculating weights for each feature separately is suboptimal. In our scenario, all features are extracted from the same image. Therefore, we have additional information about joint genuine-impostor distributions. To make use of this information, we decided to formulate the computation of weights as a pairwise optimization problem: The weights w_1 and w_2 for two features define a vector on which the scores of two features are projected to get the fused score. W.l.o.g. these weights can be expressed as $k \cdot w_1 = \cos(\phi)$ and $k \cdot w_2 = \sin(\phi)$, with ϕ being the angle between the x-axis and the projection vector (see Fig. 5 for visualization).

Then the fused genuine and impostor distributions are a function of the marginal distributions (normalized scores) and the angle of the projection vector ϕ. Therefore, finding good weights is the task to find ϕ, where the overlap of

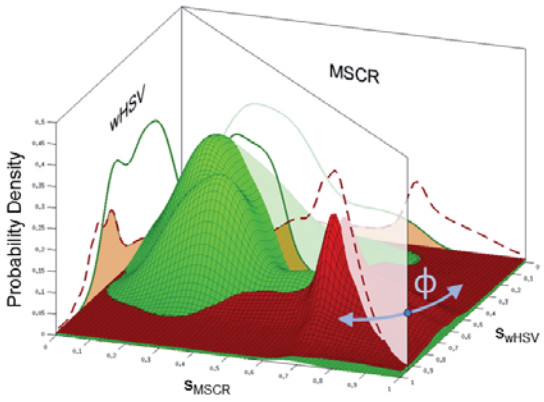


Fig. 5. Weighting formulated as optimization problem. The projection vector (displayed as semi-transparent plane) depends only on ϕ. Notice, that marginal probability densities are scaled at z-axis to visually highlight the relationship to the joint probability density distributions.

the projected genuine-impostor score distribution is minimized. For further details and experimental evaluation of our method named PROPER (Pairwise optimization of projected genuine-impostor overlap), we refer to [5].

Using a scenario-specific dataset captured with the robot, PROPER assigned weights of 0.8657 to wHSV (with learned metric), 0.1343 to MSCR, and eliminated LPB (with learned metric) by assigning weight 0.0.

E. Track-based Identification

To decide which of the hypotheses represents the user, we consider multiple observations. This reduces the amount of re-identification errors drastically, due to low influence of outlying low genuine and high imposter scores.

To achieve a multi-sample re-identification, we collect the latest matching scores for each track and apply a probabilistic framework to make the decision which track represents the user. Therefore, we extract several indicators which each vote with a probability p_i that the track should be assigned to the user and with $(p_i - 1)$ that it should not. For each track, we calculate the probability that the majority of the indicators vote for the assignment (e.g. for three indicators the track score would be $p_{track} = p_1 \cdot p_2 \cdot (1-p_3) + p_1 \cdot (1-p_2) \cdot p_3 + (1-p_1) \cdot p_2 \cdot p_3 + p_1 \cdot p_2 \cdot p_3$, i.e. either two indicators vote for the assignment and one against it or all three indicators vote for the assignment). Finally, we assign the user's person ID to the track with the highest score, if it is above a threshold. In the case that this probability-score is near 1.0, we additionally update the user template to consider changed environmental conditions (see Fig. 2).

One kind of indicators verifies, if the observed scores are similar to the user template. Therefore, for each score we determine the probability that it is a genuine score. The probability that a score is genuine, and thus represents a match with the user template, can be calculated by

$$P(gen|s_i) = \frac{P(s_i|gen)}{P(s_i|gen) + P(s_i|imp)}, \qquad (1)$$

where $P(s_i|gen)$ and $P(s_i|imp)$ are the probability densities of the genuine and imposter distribution (see Fig. 4) at s_i. For fast calculation, both probability density functions (PDF) are

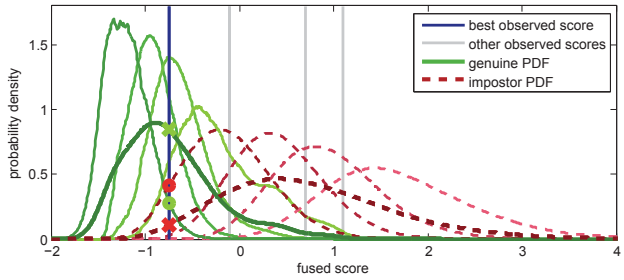


Fig. 6. Rank-corrected genuine and impostor PDFs for four sort scores of a track, calculated from the particular original PDFs (darker, bold). The four rank-corrected genuine PDFs specify in which value range (from left to right) the best, 2nd-best, 3rd-best and worst of the four scores would be expected if they were genuine scores. The four rank-corrected impostor PDFs specify the value range in which the scores would be expected if they were impostors. Example: The best of the four scores s_1 is marked as blue line, the others gray. If the PDFs were not corrected, Eq. 1 would suggest that s_1 is genuine with probability $p = \frac{0.84}{0.84+0.11} = 0.88$ (crosses). However, the knowledge that this score is the best of four observations changes the expectations. A genuine score will likely be observed within the range specified by the leftmost genuine PDF and an impostor score by the range of the leftmost impostor PDF. Therefore, the correct probability that this score is genuine, calculated by Eq. 1, is only $p = \frac{0.275}{0.275+0.415} = 0.40$ (circles).

computed beforehand on a training dataset. This calculation assumes each of the observations, and thus the scores, to be independent, which does not hold for scores of the same track, since they all belong to the same person. Therefore, for each track we must consider the order of scores (see Fig. 6 for a descriptive example). To correct the PDFs for the sorted scores $s_1, \ldots, s_k, \ldots, s_n$, we calculate order statistics:

$$f^{(m)}_{(k)}(s_k) = n \cdot f^{(m)}(s_k) \cdot \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} \cdot F^{(m)}(s_k)^{k-1} \cdot (1 - F^{(m)}(s_k))^{n-k}, \quad (2)$$

where $f^{(m)}_{(k)}(s_k)$ is the corrected probability density $P(s_k|m)$ for the k-th best score s_k with m being genuine or impostor, $f^{(m)}$ is the original PDF, and $F^{(m)}$ is the original cumulative probability function (CDF). The probability, that a sorted score is genuine, can then be calculated by Eq. 1 using the corrected genuine and impostor PDF (Fig. 6 exemplarily visualizes the procedure for a track with four observations).

Beside the score-related indicators, we also add a rank-related indicator. The probability that this indicator votes for an assignment of the track to the user is reduced whenever a score of this track is not the best match compared to simultaneously observed scores of other tracks. Therefore, for each track n score-related indicators and one rank-related indicator are included in the voting.

IV. Experiments

The evaluation of the developed person re-identification approach is tripartite: First, we report the recognition rate on a standard benchmark dataset to compare with state-of-the-art approaches. Second, we evaluate re-identification performance in the attended domain, by benchmarking on a dataset that we recorded in a stroke rehab clinic within the ROREAS project [8]. Last, we report results from live tests with three probands in a clinic to evaluate the benefit of the re-identification module for following and guiding users with a robotic walking coach.

A. Benchmarking on VIPeR Dataset

To evaluate the performance of our approach, we first examined the performance of all sub-components and then compared the proposed person re-identification system to state-of-the-art approaches.

First, we wanted to evaluate the performance of the proposed algorithm on a standard benchmark dataset. Therefore, we utilized the widely used and very challenging VIPeR dataset [7]. It consists of 632 persons, with two images each, taken from disjoint camera views, showing them under very different angles and lighting conditions (see Fig. 7(a)). The images are all normalized to a size of 128×48 pixels. To obtain comparable results, we followed the 10-fold cross-validation protocol of [6]. For each of the ten folds, 316 of the 632 available persons were chosen for testing. The images of the 316 remaining persons were used for training. Images of persons in test set from camera A represent the gallery, while camera B provided the corresponding probe images.

We made several modifications to the SDALF features (see Sect. III-B) to speed up feature extraction, increase recognition rates, and eliminate processing steps that cannot be implemented on a mobile robot that simple (e.g. background subtraction). Tab. I shows the effect of evaluated modifications. The listing shows, that

- the wHSV feature benefits from a partitioning in upper and lower body as well as a consideration of symmetry. However, static partitioning and symmetry lines do not decrease performance significantly.
- a tri-linear interpolated full wHSV-histogram increases recognition rates.
- the MSCR feature benefits from a foreground mask. However, a static average person mask is even more beneficial, because errors during extraction of the mask that have a large influence on feature extraction, are avoided.
- parameter tuning as well as metric learning improves re-identification performance significantly.

TABLE I
Effect of Modifications on Features

modification	nAUC (VIPeR)
none (original SDALF)	0.922
w/o partitioning, w/o symmetry (wHSV)	0.808
w/ partitioning, w/o symmetry (wHSV)	0.906
static partitioning and symmetry (wHSV)	0.917
histogram with interpolation (wHSV) (static partitioning and symmetry)	0.925
no mask (MSCR)	0.921
static mask (MSCR)	0.927
cross-validated parameter tuning	
marginal wHSV-histogram + MSCR	0.942
full wHSV-histogram + MSCR	0.944
additional metric learning (full tri-linear interpolated histogram, static partitioning, symmetry and mask, cross-validated parameter tuning)	0.963

nAUC: normalized area under CMC curve

This shows, that the modifications help to improve re-identification and eliminate the need for a foreground mask,

TABLE II

COMPUTATION TIME

processing step	time*
one-time offline training (metric learning)	$1.19\,s$ on VIPeR dataset [7]
feature extraction wHSV feature	$2.881\,ms$ per person
feature extraction MSCR feature	$7.775\,ms$ per person
matching wHSV	$5.382\,\mu s$ per comparison
matching MSCR	$62.726\,\mu s$ per comparison
score-level fusion	$< 1\,\mu s$ per comparison
sum 10 runs on VIPeR	
$10\times$ training (316 persons)	$11.9\,s$
$10 \times 2 \times 316$ feature extractions	$67.3\,s$
$10 \times 316 \times 316$ matchings	$68.0\,s$
	$147.2\,s$
baseline SDALF [6]	
10 runs on VIPeR	$43\,min$
proposed approach with extended feature set (48.440 dimensions)	
10 runs on VIPeR	$25\,min$

* CPU: Intel Core i7-620 (2.66GHz)

which is favorably for a mobile robotic application. Additionally, Tab. II proves that the developed re-identification module fulfills the real-time requirements listed in Sect. II.

Next, we wanted to compare the proposed method with state-of-the-art approaches. Tab. III shows the performance of the proposed method in comparison to state-of-the-art methods listed in [12], [18], and [19] (for all methods that were evaluated with different configurations, we only report the best result). Additionally, Fig. 7(b) and 7(c) show the Cumulative Match Characteristic (CMC) and Synthetic Recognition Rate (SRR). It can be seen, that the proposed real-time capable version of our re-identification algorithm keeps up with the best state-of-the-art approaches. The SRR shows, that the user will be identified with 95% probability for up to six targets (= persons in front of the robot). This performance suffices for our scenario of following or guiding users through narrow corridors.

Note, that none of the state-of-the-art approaches achieves real-time performance as defined in Sect. II. This is due the feature extraction step. Calculating histograms on multiple parts of the image in multiple color spaces, and extraction of some of the textural features is very time consuming. The performance closest to real-time is achieved by R_{χ^2}-LFDA, which approximately needs 20 ms for feature extraction for each person. This is twice the time our approach needs. The performance gain of the non real-time capable version of the proposed approach in comparison to R_{χ^2}-LFDA and R_{χ^2}-MFA comes with full instead of marginal histograms and the preprocessing of the training data for metric learning (see Sect. III-C).

B. Benchmarking on a Clinical Dataset

To benchmark the scenario-specific re-identification performance, we recorded a new 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 22,807 images, showing 207 different people, was collected. We manually eliminated false and not properly aligned person detections. The remaining

TABLE III

COMPARISON TO STATE OF THE ART (VIPER DATASET)

CMC at rank	1	5	10	20
Proposed + additional features[1]	**34.9**	**67.4**	**81.3**	**91.2**
R_{χ^2}-MFA[2] [18]	32.2	66.0	79.7	90.6
R_{χ^2}-LFDA[2] [18]	32.3	65.8	79.7	90.9
SVMML[3] [18]	30.1	63.2	77.4	88.1
Proposed (real-time capable)[4]	**27.5**	**56.7**	**70.0**	**82.8**
sLDFV[5] [12]	26.5	56.4	70.9	84.6
KISSME[6] [18]	25.8	56.2	70.1	82.9
R_{χ^2}-rPCCA[2] [18]	22.0	54.8	71.0	85.3
R_{χ^2}-PCCA[2] [18]	19.6	51.5	68.2	82.9
eSDC[7] [19]	26.7	50.7	62.4	76.4
LFDA[6] [18]	21.4	49.6	65.2	79.5
CPS[8] [2]	21.8	45.0	57.2	71.0
eBiCov[9] [11]	20.7	42.0	56.2	68.0
MCC[10] [20]	15.2	41.8	57.6	73.4
SDALF[11] [6]	19.9	38.9	49.4	65.7
PRDC[10] [20]	15.7	38.4	53.9	70.1
PRSVM[10] [14]	13.0	37.0	51.0	68.0
ITML[10] [20]	11.6	31.4	45.8	63.9
ELF[10] [7]	12.0	31.0	41.0	58.0
LMNN[10] [20]	6.2	19.7	32.6	52.3

Methods sort by rank 5 performance.

Features:

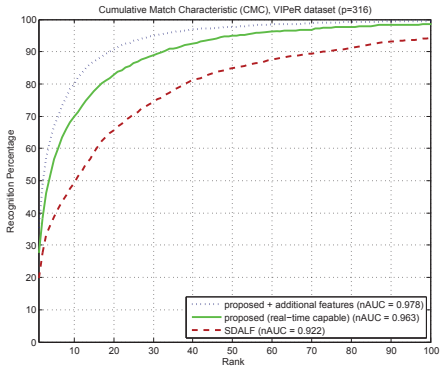
1 full wHSV+RGB+HSV+CIEL*a*b*+YUV+LBP-Hist. on 6 stripes (48,440 D.)
2 marginal RGB+HSV+YUV+LBP-Hist. on 6 stripes (2,580 Dimensions)
3 marginal RGB+HSV+YUV+LBP-Hist. on 75 patches (32,250 Dimensions)
4 full wHSV-Hist. + MSCR (ca. 2,142 Dimensions, $\sigma_{MSCR} = 16$)
5 wHSV + MSCR + LDFV (ca. 73,894 Dimensions), PCCA metric
6 marginal RGB+HSV+YUV+LBP-Hist. on 341 patches (146,630 Dimensions)
7 wHSV + MSCR + SDC (ca. 201,768 Dimensions, $\sigma_{MSCR} = 16$)
8 wHSV + MSCR on 6 body parts (ca. 312 Dimensions, $\sigma_{MSCR} = 16$)
9 wHSV + MSCR + BiCov (ca. 67,002 Dimensions, $\sigma_{MSCR} = 16$)
10 marginal RGB+YUV+HSV-Hist + Schmidt+Garbor filters on 6 stripes (2,784 D.)
11 wHSV + MSCR + RHSP / LBP (ca. 227 Dimensions, $\sigma_{MSCR} = 16$)

11,034 samples were semi-manually labeled[2]. To complicate re-identification, we eliminated similar appearances for each person automatically by clustering. Each person, for which 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. The dataset is very challenging and covers all difficulties described in Sect. II (see Fig. 7(d)).

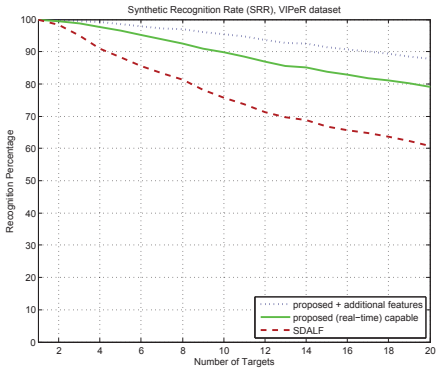
Fig. 7(e) and 7(f) show the CMC and SRR curve of our re-identification system in comparison to the SDALF approach, that extracts the same features. It is visible, that the modifications presented in Sect. III-B lead to a significant improvement of the re-identification rate. However, the SRR curve of our approach is considerably lower than on VIPeR. That means, appearance-based person re-identification on a robot (ROREAS) is far more difficult than re-identification of pedestrians in multiple static cameras with disjoint views (VIPeR). The SRR indicates, that recognizing the user within a group of five people will only succeed in 81% of all cases, using just a single observation. However, in our scenario, we can use multiple observations and contextual information to significantly increase the recognition rate. Therefore, the

[2]Labeling was assisted by automatic detection and tracking. Therefore, for each track ID the person ID had to be assigned. This amount of training data is sufficient to adequately learn distance metrics. This can also be confirmed for other applications, that most often provide datasets with this amount of training data. Additional experiments with more data did not show significant improvements. However, if no dataset is available, labeling this amount of images is quite labor intensive (approx. 12 hours).

(a) Samples of VIPeR dataset

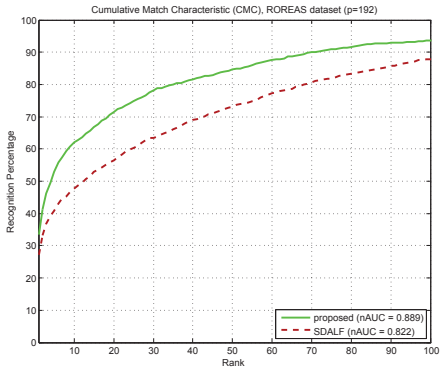


(b) CMC (VIPeR dataset)

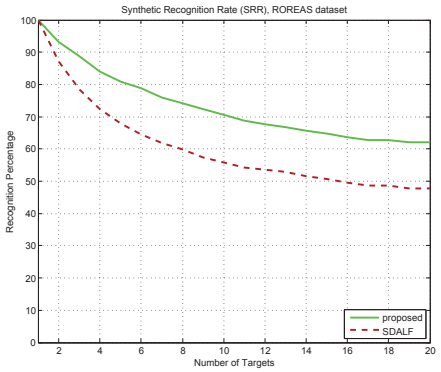


(c) SRR (VIPeR dataset)

(d) Employed robot, samples of dataset



(e) CMC (ROREAS dataset)



(f) SRR (ROREAS dataset)

Fig. 7. Re-identification performance on VIPeR (top) and ROREAS dataset (bottom). (a) Sample gallery and probe image pairs for VIPeR dataset [7]. (b, c) CMC and SRR curves for VIPeR dataset. (d) Challenging image samples of ROREAS dataset, which was recorded with a mobile robot driving through the corridor of a stroke rehab clinic. The dataset is characterized by high inner class variance and small inter class variance for groups of similar clothed people. (e, f) CMC and SRR curves for ROREAS dataset.

next subsection shall show the performance of the complete system.

C. Evaluation of Following and Guiding in Live Tests

To evaluate the benefit for the robot to use a re-identification module to resolve ambiguity in tracking, we performed live tests in the rehab clinic, where the robot shall be employed in the future. Over a period of six hours, the robot followed and guided three probands though the corridor of one ward of the clinic where later inpatients shall be coached during their self-training. Fig. 8 shows a map of the operational environment and the tree probands. Their appearances cover typical clothing: dark/black, light/gray, and colored clothes.

In each run, one of the probands was guided and followed respectively by the robot as shown in Fig. 8 for a distance of 400 m. The probands could freely choose their route and walking speed, but were instructed to behave like stroke patients (i.e. no running). The behavior of the robot was observed and manually corrected via a control tablet whenever the robot did not succeed. In these cases, the position of the user was marked on the control tablet, and the robot had to continue. We repeated guiding and following until a pure driving time of one hour was reached in each case. During this time, the robot guided the probands for an overall distance of 2 km and followed the probands for 2.4 km.

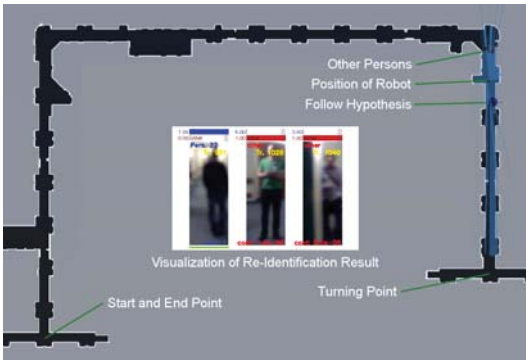


Fig. 8. Map of operational environment. Center: Exemplary visualization of re-identification where the three probands are around the robot.

To evaluate, which decision the robot would have made, if it did not use visual re-identification, a reference approach ran simultaneously. Whenever the track broke, it chose the new track by spatial distance to last observation. However, the robot ignored these decisions and behaved like the re-identification module suggested.

Tab. IV shows the results for follow and guide. As can be seen, the proposed person re-identification performed well and helped the robot to decrease the number of mismatches with other persons that would require manual correction. A drawback of visual user recognition is the confusion of

TABLE IV
RE-IDENTIFICATION PERFORMANCE IN LIVE TESTS IN A CLINIC

FOLLOW

run	pers.	hyp.	stop-t	stop-f	fp-mm.	pers.-mm.	ref
1	15	748	2	0	3 (2/1)	1	2
2	13	701	3	1	4 (4/0)	0	3
3	14	262	2	1	0	1	1
4	11	772	1	0	0	0	1
5	8	241	1	0	2 (1/1)	0	3
6	6	275	0	0	0	1	0
sum	67	2999	9	2	9 (7/2)	3	10

GUIDE

run	pers.	hyp.	stop-t	stop-f	fp-mm.	pers.-mm.	ref
1	8	386	1	1	1 (0/1)	0	1
2	13	465	0	2	2 (2/0)	0	1
3	10	847	0	0	0	0	0
4	5	247	0	1	0	0	0
5	12	154	0	0	0	0	1
sum	48	2099	1	4	3 (2/1)	0	3

LEGEND: **pers**: number of nearby persons while following the user for a distance of 400m along the floor of a rehab clinic. **hyp**: number of assigned tracking IDs for new person hypotheses (including valid false positive detections). **stop-t**: number of correct stops requested by the re-identification module due to lost user (correct behavior). **stop-f**: number of unnecessary stops requested by the re-identification module (uncritical). **fp-mm.**: mismatches of user with false positive person detections (lamps, and the like), in brackets: situation where robot could resolve situation due to user cooperation (tolerable) and where it could not (critical). **pers-mm.**: mismatches of user with other persons (critical). **ref**: reference method: mismatches of user with other persons (critical).

the user with false positive detections. Manual intervention was necessary three times, when the robot followed false positive detections and could not resolve the situation by itself. This happens if misaligned images showing walls or false detections that were assigned to the track of the user did appear after login in the enrollment phase. Then, the multi-modal template consists of observations of the user and false detections, that may match perfectly to later false detections. Since this happens quiet often, our person detection module should be improved further. Also, a fast visual tracking algorithm [9] can help to validate detections over time.

The robot did very well in stopping when the user was temporally not visible. The few additional stops are acceptable.

At rush-hour times, where the reference approach fails clearly, the visual re-identification performed very well. For example, in the situation shown in Fig. 1, the robot had to follow the proband on a zigzag course through seven people. The user was traced almost through all people, but then he was lost during an evasive maneuver. The robot immediately stopped as desired (highlighted in green in Tab. IV, run 4).

Real-time requirements were always met, and the re-identification module used only 3% of the CPU on average.

V. CONCLUSION

We implemented a person re-identification module that runs on a mobile robot to recognize its user in real-time, using only few amount of the robot's processing resources. It is robust to image blur, varying resolution and illumination, occlusions, and people with walking aids. State-of-the-art performance is confirmed on the standard VIPeR benchmark dataset, as well as on a scenario-specific dataset recorded at a stroke rehab clinic. Additionally, we tested the re-identification performance live in the addressed operation area during regular day-time routines. During the two hours of following and guiding probands on a track of 4.4 km, the robot came in close contact with 115 other people. Overall, the user was mismatched only three times. Even at rush-hour times, the robot was able to reliably follow and guide probands through the corridor of the clinic. The current performance is acceptable for the upcoming first tests with real patients, when an observer can correct the rare mistakes via a control tablet. For autonomous operation, however, the re-identification performance must be improved further.

Therefore, in our future work, we plan to fuse the proposed visual approach with a non-visual identification, based on a device carried by the patient that can be located by the robot via stereo ultrasound. Additionally, we plan to add more contextual information like predicted walking routes.

REFERENCES

[1] A. Andrade, A. Pereira, S. Walter, R. Almeida, R. Loureiro *et al.*, "Bridging the gap between robotic technology and health care," *Biomedical Signal Processing and Control*, no. 10, p. 6578, 2014.

[2] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *BMVC*, 2011.

[3] E. Einhorn, T. Langner, R. Stricker, C. Martin, and H.-M. Gross, "Mira - middleware for robotic applications," in *IROS*, 2012, pp. 2591–2598.

[4] M. Eisenbach, A. Kolarow, K. Schenk, K. Debes, and H.-M. Gross, "View invariant appearance-based person reidentification using fast online feature selection and score level fusion," in *AVSS*, 2012, pp. 184–190.

[5] 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 *IJCNN*, 2015.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[7] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS*, 2007.

[8] H.-M. Gross, K. Debes, E. Einhorn, S. Mueller, A. Scheidig, C. Weinrich, A. Bley, and C. Martin, "Mobile robotic rehabilitation assistant for walking and orientation training of stroke patients: A report on work in progress," in *SMC*, 2014, pp. 1880–1887.

[9] A. Kolarow, M. Brauckmann, M. Eisenbach, K. Schenk, E. Einhorn, K. Debes, and H.-M. Gross, "Vision-based hyper-real-time object tracker for human-robot interaction," in *IROS*, 2012.

[10] A. Kolarow, K. Schenk, M. Eisenbach, M. Brauckmann, H.-M. Gross *et al.*, "Apfel: The intelligent video analysis and surveillance system for assisting human operators." in *AVSS*, 2013, pp. 195–201.

[11] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *BMVC*, 2012.

[12] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV*, 2012, pp. 413–422.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, pp. 971–987, 2002.

[14] B. Prosser, W. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *BMVC*, 2010.

[15] M. Volkhardt, C. Weinrich, and H.-M. Gross, "People tracking on a mobile companion robot," in *SMC*, 2013, pp. 4354–4359.

[16] C. Weinrich, C. Vollmer, and H.-M. Gross, "Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images," in *IROS*, 2012, pp. 2147–2152.

[17] C. Weinrich, T. Wengefeld, C. Schroeter, and H.-M. Gross, "Generic distance-invariant features for detection of people with walking aid in 2d range data," in *RO-MAN*, 2014, pp. 767–773.

[18] F. Xiong, M. Gou, O. Camps *et al.*, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014, pp. 1–16.

[19] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.

[20] W. Zheng, S. Gong *et al.*, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.