in: Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics (IEEE-SMC 2013), Manchester, GB, pp. 4384-4390, IEEE Computer Society CPS 2013

Appearance-Based 3D Upper-Body Pose Estimation and Person Re-Identification on Mobile Robots

Christoph Weinrich, Michael Volkhardt, Horst-Michael Gross
Neuroinformatics and Cognitive Robotics Lab
Ilmenau University of Technology
Ilmenau, Germany
christoph.weinrich@tu-ilmenau.de

***Abstract*—In the field of human-robot interaction (HRI), detection, tracking and re-identification of humans in a robot's surroundings are crucial tasks, e. g. for socially compliant robot navigation. Besides the 3D position detection, the estimation of a person's upper-body orientation based on monocular camera images is a challenging problem on a mobile platform. To obtain real-time position tracking as well as upper-body orientation estimations, the proposed system comprises discriminative detectors whose hypotheses are tracked by a Kalman filter-based multi-hypotheses tracker. For appearance-based person recognition, a generative approach, based on a 3D shape model, is used to refine these tracked hypotheses. This model evaluates edges and color-based discrimination from the background. Furthermore, for each person the texture of his or her upper-body is learned and used for person re-identification. When computational resources are limited, the update rate of the model-based optimization reduces itself automatically. Thereby the estimation accuracy decreases, but the system keeps tracking the persons around the robot in real-time. The person's 3D pose is tracked up to a distance of 5.0 meters with an average Euclidean error of 18 cm. The achieved motion independent average upper-body orientation error is 22°. Furthermore, the upper-body texture is learned on-line which allowed a stable person re-identification in our experiments.**

***Index Terms*—person tracking, upper-body pose estimation, person re-identification, appearance model**

I. INTRODUCTION

The detection and tracking of human position and upper-body orientation is an important requirement to improve human-robot interaction (HRI), e.g. for the realization of socially compliant navigation behaviors or polite contact initiation.

The spatial relation between a person and a robot is part of nonverbal communication. A person's upper-body orientation towards the robot permits to estimate the human's notice of the robot or even the human's interest in an interaction - a recognition task which is highly relevant for many robotic scenarios in public or private environments, as for example supermarkets [1], public buildings [2] or own apartments [3], [4]. Thus, it is a basis for the decision whether to approach or better to avoid a human. Likewise, the robot's navigation behavior in relation to a person's pose has socio-emotional importance. Accordingly, the robot's navigation behavior should be adapted to the person's pose. For example, the navigation behavior should conform to the proxemics [5], whereby the personal space could be modeled by a set of elliptic regions relatively to the human pose and upper-body orientation [6].

Our mobile robot SCITOS G5 used in this study and in former projects [1], [2] is equipped with an omni-directional camera system, 1.5 m above the ground, and two laser range scanners, which cover 360° of a horizontal plane, 0.4 m above the ground. Both modalities are used to detect and track persons around the robot. All detected and tracked person hypotheses about torso position and orientation (with uncertainties) are represented by 6D Gaussian distributions. Thereby, we obtain a flexible modular system, where different detectors can be added or replaced. For this work, we use a laser-based leg detector [7] and a visual upper-body detector, which additionally provides rough estimates of the upper-body orientation [8], based on HOG features [9]. These detectors complement each other well, because the laser-based hypotheses have low uncertainties regarding the distance to the robot, and the vision-based hypotheses have low uncertainties regarding the height of the detected persons. Furthermore, the laser-based detector has a high update rate, and the vision-based detector has low false positive rate. In order to combine the positive advantages of both detectors, a tracker processes the asynchronous hypotheses of both detectors. Figure 1 shows an overview of the whole system.

The main focus of this work is the refinement of the accuracy of the tracked hypotheses, particularly of their orientation estimation. For this purpose, the parameters of a 3D upper-body model are optimized by Particle Swarm Optimization (PSO) [10], to match the appearance of the currently observed image. Furthermore, the 3D model is used to learn the texture of each tracked hypothesis for person re-identification. This permits to recognize a person who left and re-entered the robot's detection area. In this work, we considered important, that the system keeps tracking people, even when the available computing capacity reduces, e.g. when the robot has to process further tasks. In such situations, we accept increasing uncertainties of the resulting hypotheses.

The next section reviews state-of-the-art work, which is related to our approach. Thereafter, Sec. III describes how the detectors and the tracker are used to obtain tracking hypotheses

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

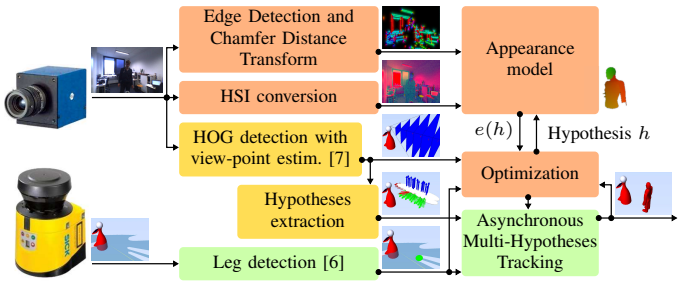


Fig. 1. System overview: The asynchronous tracker and the leg detector (green) have low computing requirements and work with high update rate to continuously track multiple persons. The HOG detector (yellow) works with lower frequency, but therefore, it has a lower false positive rate than the leg detector. Since the tracker only outputs hypotheses, that have been supported by both detectors at least once, the HOG detector significantly reduces false positive hypotheses. Furthermore, it provides rough upper-body orientation estimations. The remaining modules (orange) are used for appearance-based refinement of the pose hypotheses to learn person-specific color models for person re-identification.

for the people in the robot's surrounding. The 3D appearance model-based optimization of these hypotheses is specified in Sec. IV. Section V shows the performance of our approach in the context of experimental studies in multi-person tracking.

II. RELATED WORK

Human 3D pose estimation is challenging, because of the complexity of human articulation and appearance. Therefore, many approaches depend on multiple camera views [11], [12] and are thus not applicable on mobile robots. Approaches that are based on active depth sensors [13], [14], like Kinect™, are less applicable in the aforementioned scenarios, as they are limited due to the required data bandwidth, computational and power resources, interferences through external IR light sources (sunlight), and their limited field of view.

Approaches that perform on monocular camera images are advantageous, because they can be used with wide-angle cameras or even omni-directional cameras. These cameras are relatively inexpensive, have high information content, and are nowadays standard equipment on many mobile robots.

Approaches for pose estimation are often divided into bottom-up (discriminative, conditional, recognition-based) and top-down (generative) approaches. In our approach, the HOG detection with view-point estimation as well as the laser-based leg detection belong to the discriminative approaches. However, the optimization based on the appearance model forms a generative analysis-by-synthesis approach.

Commonly, top-down approaches synthesize a model to render a human pose hypothesis into the image plane and compare it with the observation based on image features. The 3D models are mostly modeled by primitives, like cones [15]. Like in [11], our appearance model uses a mesh based shape model. Many approaches use silhouette features for comparison [16]. Since background segmentation is relatively difficult when the camera is moving, we prefer to use edge features like in [17]. Other approaches use HOG descriptors [18] or SIFT features [19]. Additionally to edge features, we evaluate the HOG multi-resolution pyramid, because we already use HOG features for discriminative person detection.

As reported in [19], generative approaches are claimed to be computationally demanding in comparison to discriminative approaches due to the high computational costs for projecting a huge set of 3D pose candidates into the image and to compare each projection with the person image using low-level image features, like edges and silhouettes. For this reason, some approaches comprise a combination of a top-down approach with a discriminative bottom-up approach [20]. Thereby, the parameter space of the generative model is significantly reduced. Furthermore, monocular, articulated 3D pose estimation is challenging, because of the ambiguity of human 3D poses in monocular images. This is particularly evident, if only the silhouette is considered. For this reason, in [20] view-point specific HOG detectors are used within a first stage to solve the ambiguity regarding the orientation. This approach is very similar to our HOG detection with rough orientation estimation [8], whereas in [8] the orientation classification is faster due to the application of a decision tree. In [20] two further stages are applied for 2D body part estimation and to recover 3D poses. However, all of these stages are processed sequentially, and this purely visual approach is not real-time applicable. Our proposed approach makes use of the laser range scanner, and it performs pose tracking in parallel with appearance-based pose optimization. Thereby, it achieves tracking of person's position in real-time (10Hz) and estimation of joint angles at lower update rate (approx. 1Hz).

III. DETECTORS AND TRACKER

A. Detectors

Leg detector: The approach of [7] is used to detect pairs of legs in the 2D laser scan $\boldsymbol{S}(t)$ at time t. To create a 3D torso hypothesis from a leg pair's position, it is assumed that the torso is located above the legs. Since the leg position provides no evidence of the vertical torso position, the distribution over human torso heights is used for the leg-based torso hypotheses $\boldsymbol{H}_{LEG}(t)$. The leg-based hypotheses have infinite uncertainties regarding the upper-body orientation. In turn, the leg-based observation model has relatively low uncertainty regarding the distance to the laser scanner.

HOG upper-body detector: The sliding window approach in [8] on a multi-resolution pyramid of the input image $\boldsymbol{I}$ is used to determine the probability $p(\boldsymbol{I}_w|c)$ that image window w could be observed, when the respective image detail $\boldsymbol{I}_w$ shows a person with upper-body orientation class c. The output is a multi-resolution pyramid of probabilities with $c = 8$ layers per resolution level, because each upper-body orentation class covers 45° in our approach. Using a calibrated camera, a given pose hypothesis in 3D space can be mapped onto a position in the multi-resolution pyramid, yielding the corresponding probabilities or vice versa. The hypotheses extraction module (Fig. 1) transforms each pyramid position w, with a probability $p(\boldsymbol{I}_w|c)$ above a certain threshold into 3D space. Note, that the observation model of these hypotheses $\boldsymbol{H}_{HOG}(t)$ has relatively large uncertainty regarding the distance to the camera.

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

B. Tracker

The asynchronous multi-hypotheses tracker provides filtered hypotheses $\boldsymbol{H}(t)$ with 10Hz update rate based on the detections $\boldsymbol{H}_{LEG}(t)$, $\boldsymbol{H}_{HOG}(t)$ and the appearance hypotheses $\boldsymbol{H}_{APP}(t)$, which are described below. For each update, all detections that have been made since the last update are processed based on their timestamp t. However, some detectors, such as the HOG detector, have larger processing time than 100ms. This means that the detections with timestamp t are only available at time $t+\Delta t$ and thus after the corresponding tracker update. To handle these out-of-sequence detection hypotheses, the detections are predicted to the current timestamp. This is basically done by increasing the uncertainties of these hypotheses. A track ID i is assigned to all hypotheses $\boldsymbol{h}_i(t) \in \boldsymbol{H}(t)$, while they are constantly tracked over time. Whenever a person leaves the robot's detection area and re-enters it, a new track ID is assigned to the person's hypotheses. Note, that position and orientation are filtered independently in this work. The motion direction of moving hypotheses is not used to support the orientation estimation.

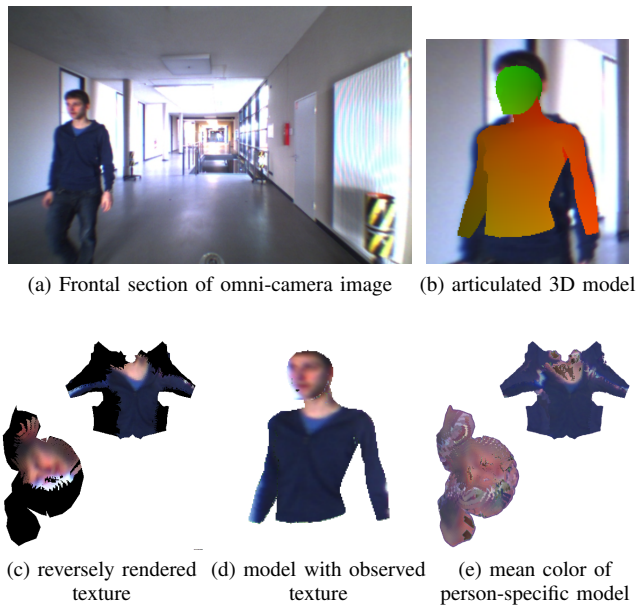
IV. OPTIMIZATION WITH 3D APPEARANCE MODEL

The previously described tracker provides a 6D hypothesis with uncertainties $\boldsymbol{h}_i(t)$ about the torso position and orientation of each currently tracked person i. Each pose distribution is the starting point for the appearance-based optimization of a 3D model. To model the diversity of the human appearance, this model has 14 degrees of freedom (DOFs). Most of these model parameters $\boldsymbol{\theta}$ are not tracked, because of low update rates and an uncertain motion model that would not justify the computing effort.

The appearance model represents a matching function $f(\boldsymbol{I}, \boldsymbol{\theta})$, which aims to correlate with the likelihood $p(\boldsymbol{I}|\boldsymbol{\theta})$, that the current image $\boldsymbol{I}$ might be observed given the pose parameter vector $\boldsymbol{\theta}$. The model parameters and the matching function $f(\boldsymbol{I}, \boldsymbol{\theta})$ are specified below. Thereafter, the description of the model parameter optimization is given. The optimized parameters $\boldsymbol{\theta}$ are used to provide appearance-based hypotheses $\boldsymbol{H}_{APP}(t)$ and for learning a color model of each tracked person's texture for re-identification if the person was lost from view.

A. Appearance Model

The 3D model of an average upper-body without hands (Fig. 2b) was generated with MakeHuman™[21]. We have fixed the model's degrees of freedom (DOFs) for the hands, gender, age, muscle mass, weight, breast size, proportion etc., because a more complex 3D model would require more parameters to be estimated, which would lead to higher computational costs during the optimization process. To model the torso position and upper-body orientation (4DOF), the head pan and tilt (2DOF), articulation of both upper arms (6DOF) and the bend of the elbows (2DOF), the model already has 14 DOFs. Furthermore, an additional DOF is used to model the color model for different people. To calculate $f(\boldsymbol{I}, \boldsymbol{\theta})$, several



(a) Frontal section of omni-camera image (b) articulated 3D model

(c) reversely rendered texture (d) model with observed texture (e) mean color of person-specific model

Fig. 2. The observed image (a) is reversely rendered into the texture (c) of the articulated 3D model (b). Thereby, the forward rendered 3D model with this texture (d) would appear like the observed image. The reversely rendered textures (c), which only show partial areas of a person's upper-body, are used to adapt a complete person-specific color model (e), which is used for person re-identification.

features of the image $\boldsymbol{I}$ are evaluated on a graphics processing unit (GPU). The GPU is mainly used for efficient match value calculation by special shader programs. It does not need to be very powerful to calculate the following match values:

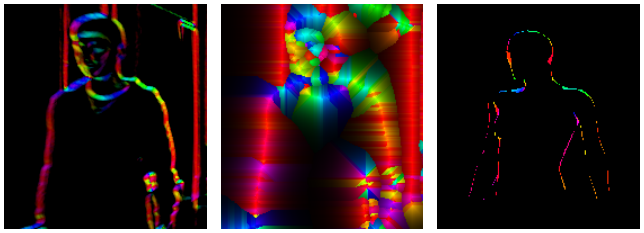
Edge Model: In relation to the great variance of texture and color of people's clothes, a comparatively invariant feature can be found in the image gradients. The success of robust detection approaches, like HOG [9], proves the relevance of these features. The edge model compares the expected edge gradient orientations $\boldsymbol{\theta}^O$ of the 3D model pixel by pixel to the gradient orientations $\boldsymbol{I}^O$ in the image, whereas a Gaussian is used to model the pixelwise match value based on the gradient orientation difference. The respective magnitudes of the model gradients $\boldsymbol{\theta}^M$ and the image gradients $\boldsymbol{I}^M$ are used as weights for calculation of the weighted mean $f_{Edg}(\boldsymbol{I}, \boldsymbol{\theta})$ of all pixel's match values.

The expected gradients (Fig. 3c) are modeled by special vertex and pixel shader programs on the GPU. The vertex normals of the model are projected onto the image plane and this is interpreted as expected edge gradient orientation $\boldsymbol{\theta}^O$. The magnitudes $\boldsymbol{\theta}^M$ of the expected edges result from the dot product of the model's normals and the viewing direction to the model's surface. This is similar to "Cel Shading".

For each observed image, the edge detection module (Fig. 1) calculates a gradients orientation image $\boldsymbol{I}^O$ and a magnitude image $\boldsymbol{I}^M$ like in [9]. Thereby, simple 2x2 Robert's Cross kernels are used for horizontal and vertical edge detection. To reduce noise and emphasize the relevant edges, a nonlinear filter is applied to the magnitude images $\boldsymbol{I}^M$, suppressing low values and emphasizing the higher ones. Additionally,

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

in order to smooth the magnitude image and therewith the matching function $f_{Edg}(\boldsymbol{I}, \boldsymbol{\theta})$, the gradient magnitudes are spatially spread out similar to the Chamfer distance transform [22], taking edge orientation from the highest magnitude in the surrounding pixels. This algorithm allows to propagate the edge information to arbitrary distance at constant time. Fig. 3b shows the resulting gradient image, which is used for matching with the expected gradient image (Fig 3c).



(a) raw gradients image (b) propagated gradients image (c) modeled gradients (expected)

Fig. 3. The edge gradients of the observed image (a) are propagated based on chamfer distance transformation (b). This increases robustness regarding deviations of the modeled gradients (c). In all images the gradient orientation is coded by color and the magnitude by intensity.

Foreground/Background Divergence (FBD) model: The previously described edge model is affected by strongly structured clothes or background, because of the contours of the human cannot be distinguished from texture of the human or the background. Therefore, the FBD model $f_{FBD}(\boldsymbol{I}, \boldsymbol{\theta})$ values the observed image $\boldsymbol{I}$'s foreground/background segmentation by the model parameters $\boldsymbol{\theta}$, based on color. A 2D color histogram (hue and saturation) of the whole visible upper-body surface $p_F(H, S)$ and a second histogram $p_B(H, S)$ of the margin around the visible upper-body pixels are compared using the Bhattacharyya distance [23]. The FBD match value $f_{FBD}(\boldsymbol{I}, \boldsymbol{\theta}) = 1.0 - BD(p_F(H, S), p_B(H, S))$ is high, when the model parameters $\boldsymbol{\theta}$ lead to different color distributions of foreground and background (as determined by the model). Due to the histogram calculation on image areas, small changes of the model parameters $\boldsymbol{\theta}$ lead to small changes of the histograms. Thereby, the matching function $f_{FBD}(\boldsymbol{I}, \boldsymbol{\theta})$ is inherently smooth. The foreground histograms could also be used for person re-identification, but this has not been investigated yet.

Color Model: In contrast to the previously described models, the color model $\boldsymbol{c}_m$ is person-specific. Accordingly, a universal color model $\boldsymbol{c}_0$ for hypotheses optimization of unknown persons and multiple color models $\boldsymbol{c}_{m>0}$ for optimization and re-identification of already tracked persons are utilized. Before the use of the different color models is explained in more detail, the match value function $f_{Col}(\boldsymbol{I}, \boldsymbol{\theta})$ for any color model $\boldsymbol{c}_m$ is specified. For given model parameters $\boldsymbol{\theta}$, "reverse rendering" is applied to project the observed HSI-image on the model's texture. In other words, the texture is calculated which would have caused the observed image (Fig. 2).

The color model operates in HSI color space. For each texture pixel, a Gaussian distribution on the HSI color is specified. The mean color of such a color model is shown in Fig. 2e. Its parameters are learned on-line using maximum a posteriori (MAP) estimation. If a tracked person moves in front of the robot and sequentially shows the entire surface of its upper-body to the camera, a complete color model is learned. The currently observed texture is matched with the Gaussian texture model, by determination of the average likelihood over all visible texture pixels, that the observed texture color belongs to the model. Initially, a universal color model $\boldsymbol{c}_0$ is used for hypothesis optimization. Then, the optimized model parameters $\boldsymbol{\theta}$ are used to adapt the universal model and store it as person-specific model. A mapping of track id i to person id m is used to apply the same person-specific color model m for optimization and adaption while a hypothesis is tracked. Whenever a new track id i occurs, the generic color model $\boldsymbol{c}_0$ is used during optimization. Thereafter, the optimized parameters $\boldsymbol{\theta}$ are used to check if a person-specific color model $\boldsymbol{c}_{m>0}$ matches better than the generic model $\boldsymbol{c}_0$. In that case, this is considered as re-identification, the color model is adapted, and the mapping of the new track id to the person-specific color model is added. If the observation reaches greatest likelihood for the generic model, the generic model is updated and stored as new person-specific model. Furthermore, the mapping of the current track id to the new person id is added.

B. Discriminative Models

The HOG-based detector and the leg detector are used to generate discrete hypotheses for the tracker. However, the outputs of these detectors are also used to improve the appearance-based optimization. In the following, we describe, how the detector outputs are used to calculate match values for a given parameter configuration $\boldsymbol{\theta}$.

HOG model: As described in Sec. III-A, the tracker processes discrete hypotheses, which are extracted from the HOG filter pyramid by a threshold operation. However, for evaluation of the model parameters $\boldsymbol{\theta}$ the resulting HOG match value $f_{HOG}(\boldsymbol{I}, \boldsymbol{\theta})$ is calculated by transformation of the model parameters $\boldsymbol{\theta}$ into the HOG pyramid and interpolation of the probability values of the adjacent orientations, pyramid levels, horizontal and vertical positions.

Leg model: For consideration of the leg detections (Sec. III-A) during optimization, the torso position is extracted from the parameters $\boldsymbol{\theta}$. Then the leg detector-based torso hypotheses with uncertainties are used to calculate the likelihood for this position $f_{Leg}(\boldsymbol{S}, \boldsymbol{\theta})$.

C. Match value calculation

The previously described partial match values are combined to the overall match function $f(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta})$ by the gamma operator, known from fuzzy logic. It is a compromise between product and weighted mean:

$$f(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta}) = \gamma \left(\sqrt[5]{\prod_{M \in \{\text{Edg,FBD,Col,HOG,Leg}\}} \omega_M f_M(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta})} \right) + (1-\gamma) \left(\frac{1}{5} \sum_{M \in \{\text{Edg,FBD,Col,HOG,Leg}\}} \omega_M f_M(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta}) \right) \quad (1)$$

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

The applied gamma γ and the weighting factor ω_M for each of the 5 models are specified in the experiments section (V).

D. Optimization

Each tracked person hypothesis is optimized by PSO [10]. In our case, the particle swarm consists of 20 particles. Each of them represents a 14-dimensional parameter configuration $\boldsymbol{\theta}$. The particle swarm is initialized according to the Gaussian distribution of the hypothesis. The parameters for joint orientations, that are not tracked (head pose, etc.), are initialized according to a predefined Gaussian distribution. The particle's velocity vectors are initialized based on predefined probabilities as well. Then, the PSO is performed for maximal 20 iterations. Thereby, the matching function $f(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta})$ over the parameter-space $\boldsymbol{\Theta}$ is used as optimization criteria.

V. EXPERIMENTS

Before the accuracy of the upper-body orientation is evaluated, the matching function of the appearance model is visualized.

A. Matching Functions

Ideally, the matching values $f(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta})$ increase continuously and spaciously while the model parameters $\boldsymbol{\theta}$ converge with the actual 3D pose parameters of a person in the robot's surroundings. This means the models need to be tolerant to deviations of the parameters from the actual person's pose to support the optimization process. On the other hand, the models need to be specific enough, so that the matching function has a well distinctive maximum for the correct parameters.

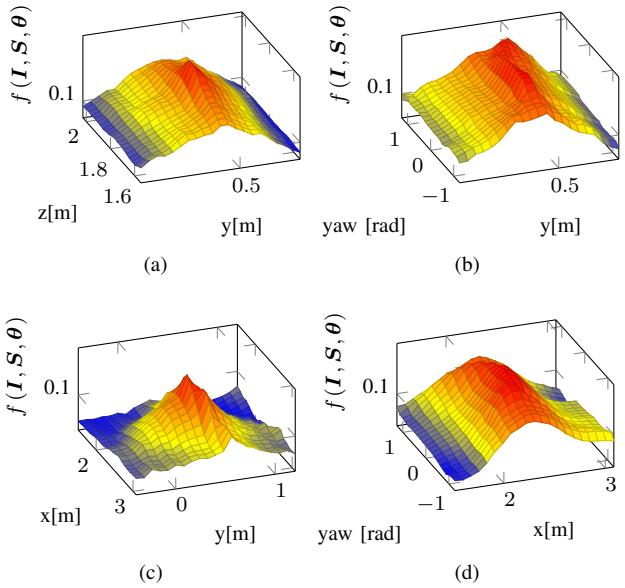


Fig. 4. Matching function $f(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{\theta})$ with generic color model $\boldsymbol{c}_0$ over two of the 14 parameter of the 3D model. The correct parameters are located in the center of the respective plot.

The performance of the used models regarding these criteria is illustrated by Fig. 4. The applied function parameters (Equ. 1) are $\gamma = 0.1$, $\omega_{Edg} = 0.1$, $\omega_{FBD} = 1.0$, $\omega_{Col} = 1.0$, $\omega_{HOG} = 0.1$ and $\omega_{Leg} = 0.1$. The correct pose parameters are located in the center. Fig. 4a and 4c show, that the matching function has good gradients regarding the torso position. Fig. 4d and 4b show, that the matching function is less sensitive to the upper-body orientation. But the correct parameter configuration is still distinguishable. The influence of the person specific color models is shown in Fig. 5. In contrast to Fig. 4, the correct pose has a more pronounced maximum, which enables the person re-identification.

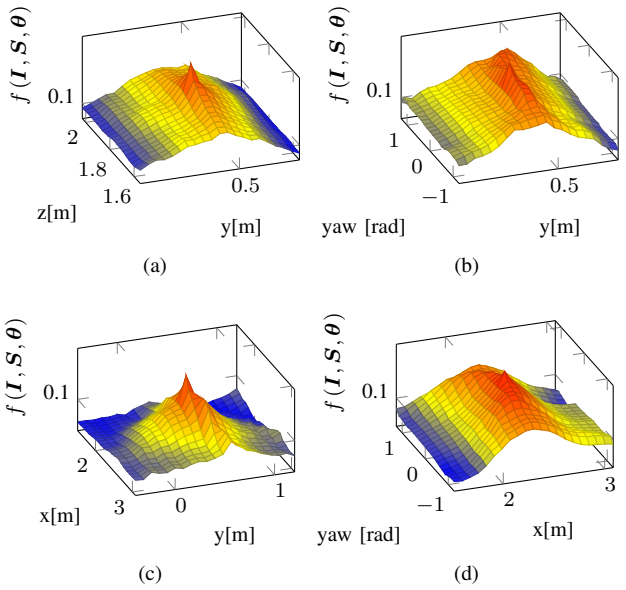


Fig. 5. Matching function $f_{Col}(\boldsymbol{I}, \boldsymbol{\theta})$ of learned color model over two of the 14 parameter of the 3D model. The correct parameters are located in the center of the respective plot.

B. Upper-Body Pose and Person Re-Identification

To evaluate the proposed tracking system, we performed an experiment, where 3 test persons walked repeatedly through an evaluation area in front of the robot (Fig. 6). An external multi-laser tracking system [24] was used to track the persons' 2D ground truth positions. Each person's height was measured manually once. Because the probands were only allowed to walk in the direction of their upper-body orientation, the ground truth upper-body orientation could easily be calculated from the motion direction.

Before the evaluation has been performed, a universal color model was learned, based on the observations of five different people. Furthermore, two person-specific color models were learned, for other people than the three probands. This is to test, whether the probands are actually detected as previously unknown persons and new person-specific models are created. A false recognition as an already tracked person would be counted as mismatch. The three test persons entered the detection area two times. The first time they had to be perceived as previously unknown, and the second time they had to be recognized.

To measure the performance of our tracking system, we use the Multiple Object Tracking (MOT) performance metric

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

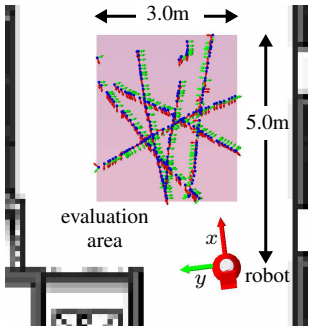


Fig. 6. Bird's-eye view on the experimental setup. The axis arrows show the position and orientation of the tracked hypotheses within the evaluation area.

[25]. Table II shows the MOT Precision (MOTP), which is the average position error, and MOT Accuracy (MOTA), which shows the accuracy and consistency of the tracker:

$$MOTA = 1 - \frac{\sum_k Miss_k + FP_k + MME_k}{\sum_k G_k} \qquad (2)$$

Thereby $Miss_k$ is the number of missed ground truth hypotheses at evaluation step k and FP_k is the number of false positive detections. For validation of the correspondence between ground truth poses and tracked hypotheses, the Euclidean distance with a threshold of 0.6m is used. The mismatch error MME_k specifies how many person id mismatches are made, and G_k denotes the number of all labels for evaluation time k.

TABLE I
COMPUTING TIME

	CPU cycles	processing time [ms]
Tracking	$1.5 \cdot 10^6 \pm 1.6 \cdot 10^6$	0.6 ± 0.6
Leg Detection	$1.7 \cdot 10^6 \pm 1.8 \cdot 10^6$	0.6 ± 0.6
HOG detection with view-point estimations	$5.1 \cdot 10^8 \pm 1.1 \cdot 10^8$	184 ± 39
Appearance-based optimization of tracked hypotheses	$2.9 \cdot 10^9 \pm 3.4 \cdot 10^8$	1068 ± 121

TABLE II
TRACKING QUALITY

	without optimization	with optimization
# evaluations	355	355
$\overline{MME}$	$1.72 \cdot 10^{-2}$	0.0
MOTP	0.187	0.173
MOTA	0.97	0.99
$\overline{\text{X error}}$ [cm]	10.5 ± 0.71	10.2 ± 0.68
$\overline{\text{Y error}}$ [cm]	11.5 ± 0.64	11.3 ± 0.63
$\overline{\text{Z error}}$ [cm]	6.35 ± 0.02	3.02 ± 0.03
$\overline{\text{orientation error}}$ [°]	24.64 ± 7.68	22.34 ± 7.57

The tracking was performed on an Intel® Core™ i7 CPU with 2.8 GHz and an Nvidia® GeForce GTX 470 GPU. The computation time for each of the modules is shown in table I. Table II shows, that the proposed tracking system is generally suitable to track the position of walking persons. Regarding the upper-body orientation estimation, note that the motion direction is not used to support orientation estimation during the experiment. Thereby, it is guaranteed that the orientation estimation is independent from the movement, and therefore we do not need to perform a motion-dependent evaluation. The absolute mean error without the proposed optimization is 24.64°. This is relatively large, because each orientation class spans 45°, and additionally classification is not perfect. However, it is shown as well, that the refinement by the appearance-based 3D model does not significantly improve the pose estimation. This probably has two reasons. Firstly, a false classification of the upper-body orientation class supports a false orientation during refinement. Secondly, the 3D model has not the same proportions as the used probands. We plan to investigate the exact reason in the near future. However, for the color model the precision of the appearance-based pose and joint angle estimation is much more important, than the pose estimation accuracy. Because a precise pose estimation enables to map each real-world point of the upper-body surface to an according point of the texture, even when the accuracy of the pose estimation is low. Accordingly, significant improvements are made for person re-identification. This is reflected by the reduction of the mean mismatch error $\overline{MME}$ and the improved multi-object tracking accuracy (MOTA).

VI. CONCLUSIONS

This work presents an approach for discriminative person detection and tracking in real-time. Simultaneously, the tracked hypotheses are used as initial pose of an articulated 3D upper-body model, whose appearance is matched to the monocular image. The optimized model allows to learn a texture model for each tracked person using "reverse rendering". These person-specific textures are used for person re-identification. This allows to recognize, whether people were already tracked, and to distinguish already tracked people.

ACKNOWLEDGMENT

This work has received funding from the Ph.D. Graduate School on Image Processing and Image Interpretation at Ilmenau University of Technology.

REFERENCES

[1] Gross et al., "TOOMAS: Interactive shopping guide robots in everyday use - final implementation and experiences from long-term field trials," in *Proc. Int. Conf. on Intel. Robots and Systems*, 2009, pp. 2005–2012.
[2] Stricker et al., "Interactive mobile robots guiding visitors in a university building," in *Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2012, pp. 695–700.
[3] Gross et al., "Further progress towards a home robot companion for people with mild cognitive impairment," in *Proc. of Int. Conf. on Systems, Man, and Cybernetics (SMC)*, 2012, pp. 637–644.
[4] C. Schröter, S. Müller, M. Volkhardt, E. Einhorn, C. Huijnen, H. van den Heuvel, A. van Berlo, A. Bley, and H.-M. Gross, "Realization and user evaluation of a companion robot for people with mild cognitive impairments," in *Proc. of Int. Conf. on Robotics and Automation (ICRA)*, 2013, pp. 1145–1151.
[5] E. T. Hall, "A system for the notation of proxemic behavior," *American Anthropologist*, vol. 65, no. 5, pp. 1003–1026, 1963.

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

[6] E. Pacchierotti et al., "Human-robot embodied interaction in hallway settings: a pilot user study," in *Int. Workshop on Robot and Human Interactive Communication (RO-MAN)*, 2005, pp. 164 – 171.

[7] K. Arras, O. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Proc. of Int. Conf. on Robotics and Automation (ICRA)*, 2007, pp. 3402 –3407.

[8] Weinrich et al., "Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images," in *Proc. Int. Conf. on Intel. Robots and Systems*, 2012, pp. 2147–2152.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[10] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. of IEEE International Conference on Neural Networks (ICNN)*, vol. 4, 1995, pp. 1942–1948.

[11] M. Hofmann and D. Gavrila, "Multi-view 3d human pose estimation in complex environment," *Int. Journal of Computer Vision (IJCV)*, vol. 96, no. 1, pp. 103–124, 2012.

[12] A. Elhayek et al., "Spatio-temporal motion tracking with unsynchronized cameras," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1870 –1877.

[13] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.

[14] D. Droeschel and S. Behnke, "3d body pose estimation using an adaptive person model for articulated icp," in *Proc. 4th Int. Conf. on Intelligent Robotics and Applications (ICIRA)*, 2011, pp. 157–167.

[15] Sigal et al., "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision (IJCV)*, vol. 98, pp. 15–48, 2012.

[16] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 1, pp. 44–58, 2006.

[17] M. Dimitrijevic, V. Lepetit, and P. Fua, "Human body pose detection using bayesian spatio-temporal templates," *Computer Vision and Image Understanding (CVIU)*, vol. 104, no. 2-3, pp. 127–139, 2006.

[18] P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2241–2248.

[19] J. Brauer, W. Hübner, and M. Arens, "Generative 2d and 3d human pose estimation with vote distributions," in *International Symposium on Advances in Visual Computing (ISVC)*, 2012, pp. 470–481.

[20] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 623–630.

[21] M. Bastioni et al., "Ideas and methods for modeling 3d human figures: the principal algorithms used by makehuman and their implementation in a new approach to parametric modeling," in *Proc of the 1st Bangalore Annual Compute Conf. (COMPUTE)*, 2008, pp. 1–6.

[22] D. G. Bailey, "An efficient euclidean distance transform," in *Proc. Int. Workshop on Combinatorial Image Anal. (IWCIA)*, 2004, pp. 394–408.

[23] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhya: The Indian Journal of Statistics (1933-1960)*, vol. 7, no. 4, pp. 401–406, 1946.

[24] K. Schenk, M. Eisenbach, A. Kolarow, and H.-M. Gross, "Comparison of laser-based person tracking at feet and upper-body height," in *Proc. of German Conf. on Advances in Artif. Intel. (KI)*, 2011, pp. 277–288.

[25] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing (JIVP)*, vol. 2008, pp. 1–10, 2008.

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