2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Multi-User Identification and Efficient User Approaching by Fusing Robot and Ambient Sensors

Ninghang Hu*, Richard Bormann*, Thomas Zwölfer, and Ben Kröse

***Abstract*— We describe a novel framework that combines an overhead camera and a robot RGB-D sensor for real-time people finding. Finding people is one of the most fundamental tasks in robot home care scenarios and it consists of many components, *e.g.* people detection, people tracking, face recognition, robot navigation. Researchers have extensively worked on these components, but as isolated tasks. Surprisingly, little attention has been paid on bridging these components as an entire system. In this paper, we integrate the separated modules seamlessly, and evaluate the entire system in a robot-care scenario. The results show largely improved efficiency when the robot system is aided by the localization system of the overhead cameras.**

I. Introduction

Globally, aging of populations is becoming a potential problem. The growing group of elderly people requires efficient and accurate care-giving at an affordable level, and robots may offer a solution in future. In recent years, researchers have been extensively working on the tasks of people detection [1], people tracking [2], face recognition [3], robot navigation, and robot controls, however, mainly as isolated tasks instead of combining these systems for real life applications. In this paper, we study these tasks jointly, and we propose a unified system that integrates these components in a home care scenario, as seen in Fig. 1. The system is very efficient and suitable for real-time applications. Moreover, the single components are complementary to help improving the robustness of the entire system.

Two fundamental tasks in robot home care scenarios are people localization and people identification. They are also the elemental components for more advanced tasks, *e.g.* activity recognition [4], [5]. Commonly used sensors for these tasks include overhead cameras and RGB-D sensors on mobile robots. The overhead cameras are usually fixed at the ceiling, covering most of the areas in the room. The cameras only need to be calibrated once so that the coordinates of the detected person can be transformed easily from the image space to the ground-plane of the room. As the camera is mounted at the ceiling, people in the video are less likely to be occluded by each other. The overhead camera commonly has a wide field of view. Thereby one camera is often sufficient for detecting and tracking people in the whole room. Despite these benefits, it is very difficult for the overhead camera to tell people's identity. Faces can hardly be seen at many locations. The most prominent parts of people are the clothes, but they may be changed over time. Consequently, the overhead camera may be enough to find the person, but it is not sufficient for people identification.

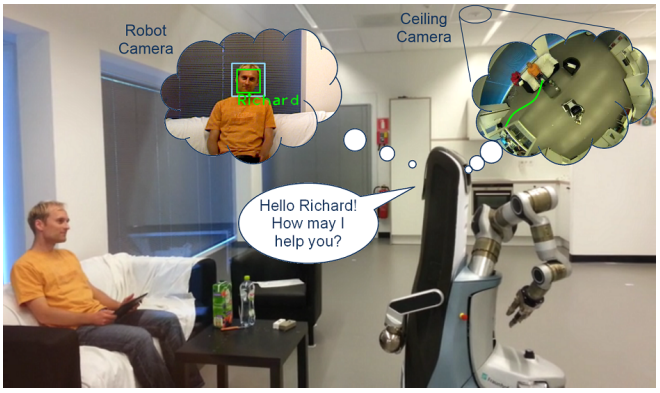


Fig. 1. Fusion of robot and environment cameras for direct user approach.

The sensors on the robot, *e.g.* the Microsoft Kinect, provide a complementary view to the overhead camera. The on-board cameras are commonly mounted at a level that keeps the human face in sight. The RGB-D sensor provides both the color image from a color camera and the depth image from a range camera. By fusion of the depth image and color image, the face can be recognized robustly [3]. However, the RGB-D sensor is limited in both the range and the view angle. When people are too close, the face is outside the field of view; when they are far away, the accuracy and resolution of face data drops quickly. An advantage of the combination with ceiling cameras for tracking is that the robot does not need to keep monitoring the persons all the time. Hence, the robot may carry out other tasks, rather than allocating its resources to the task of tracking each person.

In this paper, we propose a system that combines the robot RGB-D sensor and the overhead cameras for real-world applications. The architecture of the proposed system is shown in Fig. 2. The system consists of three modules: a) people detection and tracking, b) people identification, and c) a joint tracker that combines both kinds of information. The first module finds multiple people that are present in the room using two overhead cameras. The second module identifies people using a Kinect sensor that is mounted on the robot. The third module collects information from the first two modules and associates tracks with human identities.

The research has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 287624

*The two authors contribute equally to the paper.

Ninghang Hu and Ben Kröse are with Informatics Institute, Faculty of Science, University of Amsterdam, 1098 XH Amsterdam, The Netherlands. n.hu,b.j.a.krose@uva.nl

Richard Bormann and Thomas Zwölfer are with the Institute for Manufacturing Engineering and Automation, Fraunhofer IPA, 70569 Stuttgart, Germany. richard.bormann@ipa.fraunhofer.de, thomas.zwoelfer@ipa.fraunhofer.de

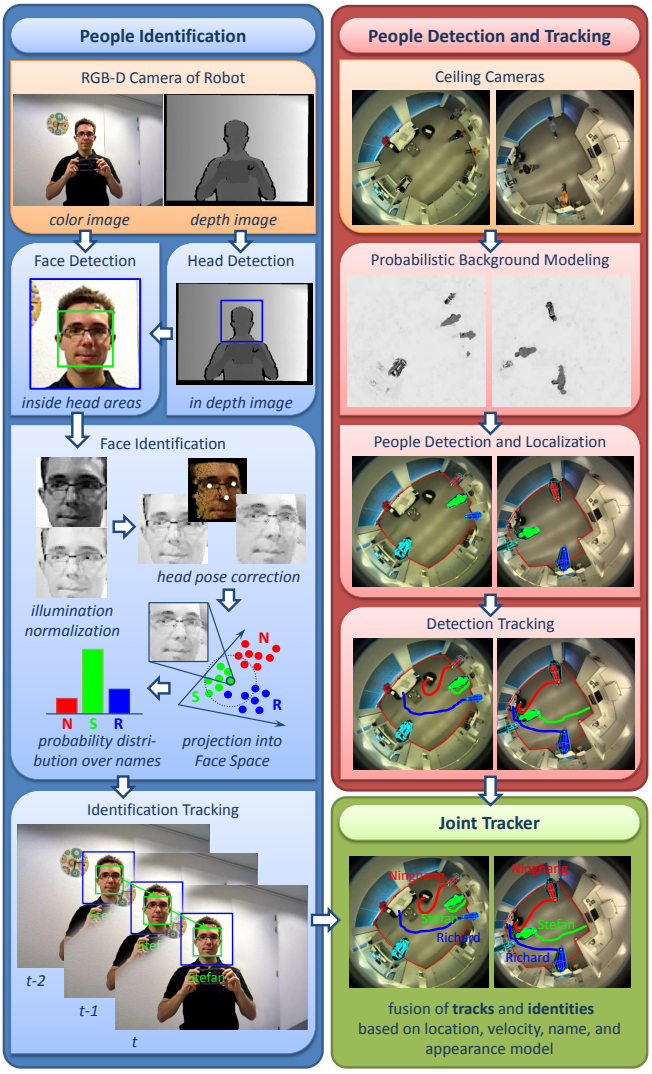


Fig. 2. An overview of the combined tracking and recognition system.

The contributions of the paper are:

1) We propose a system that integrates the overhead cameras with the RGB-D sensor for human localization and identification tasks. All components of our system follow a probabilistic approach so that the system is more robust to noise. The experiments show that the system is very efficient and can be applied in real-time.
2) Instead of evaluating the components as isolated tasks, we evaluate the effectiveness of the whole system together with robot path planning and navigation in a real life scenario, *i.e.* finding one or more persons.
3) We show a novel way of associating the faces with human tracks. We represent the face locations as weighted particles and the faces are associated with the tracks by evaluating the weighted particles in a set of Kalman Filters.

The remainder of the paper is structured as follows. We review the related work in Section II. Afterwards, we introduce three main components of the system, *i.e.* people localization and tracking in Section III, people identification in Section IV, and the joint tracker in Section V. We show the results of our experiments in Section VII.

II. Related Work

In the past few years, there has been considerable work on people detection and people tracking using computer vision techniques. Most of the work only adopts a single type of sensor, *e.g.* color cameras [6], [7], [8], depth cameras [9], [10], or laser range finders [11]. More recently, researchers worked on combining different sensors to make the tracking system more reliable. Cui et al. [12], Kristou et al. [13] and Kobilarov et al. [14] utilize a mobile robot platform to detect people with a laser range finder and video cameras. Luber et al. [15] fuse the data from a depth camera and a color camera. Nakazawa et al. [16] combine multiple cameras for multi-people tracking. For people detection, we employ a similar approach as in [17] and [18]. Instead of fusing the on-board laser range finder and the ambient camera, we combine the Kinect sensor with ambient cameras. We use the Kinect sensor for identifying people and ambient cameras for tracking. The results from the two components are fused in a probabilistic way by a joint tracker.

Face recognition is usually split up into the detection of face regions in an image and the actual identification of the detected face image patches. The former task has been approached on color images using the Viola-Jones classifier [19]. Later, many extensions have been introduced [6], [20]. Face detection is tackled on point clouds using local curvature features [10]. An RGB-D fusion system for combined head detection in depth images and face detection in color images [21] is used for face detection in this paper.

There exists a large variety of methods for face identification that might be divided into projection methods [7], [22], [23], [24], local pattern-based methods [25], [26], generative models [27], [28], and sparse representations [29]. The latter represent the space of known faces with a set of carefully chosen gallery images whereas generative methods construct an illumination and pose model for each individual from training data recorded under specialized lighting conditions. Both kinds of methods suffer from long training and/or recognition times. Given the robustness and real-time demands of robotics applications, projection methods like Eigenfaces [22] or Fisherfaces [7], or local pattern-based methods are preferable. The first construct a handy representation of identities by projecting the high dimensional face image matrix into a low-dimensional subspace, which commonly reduces intra-class variance and amplifies inter-class differences. Local pattern-based methods compute dense local binary patterns [25] or local ternary patterns [26] which become accumulated in histograms over spatially constrained areas. The robustness of these methods can be improved by applying illumination normalization techniques like histogram equalization, logarithmic transform [30], gamma correction [31], discarding the low-frequency Discrete Cosine Transform coefficients [32], Difference of Gaussians filtering, or contrast equalization [26] to the face image in advance. Compensation measures for varying head

pose such as multiple orientation modeling on the training data [27], [33] or face plane estimation [10], [34], [35] also increase the robustness of the identification system. The face recognition module used for this work bases on our earlier work on robust real-time face recognition systems [3].

Most of the previous work focuses on solving only one type of the tasks and little work has been done on combining modules and evaluating the system as a whole. In this paper, we work on bridging the gaps between the different components and build an integrated tracking and recognition system that is suitable for real home care scenarios.

III. People Localization and Tracking

In this section, we introduce the sub-system for people localization and tracking. The pipeline of the sub-system is shown in Fig. 2 (red panel).

A. People Localization

The ground-plane area is discretized into small regions for localizing people. Our goal is to find in which regions the people are located. We define k as a random variable to indicate the index of a region, and $\boldsymbol{X} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_C}\}$ as a set of images that are observed from the overhead cameras. We formulate the people detection problem in a Bayesian fusion framework [36].

$$P(k|\boldsymbol{X}) \propto P(k) \prod_{c=1}^{C} P(\boldsymbol{X_c}|k) \tag{1}$$

where $P(k)$ is a prior distribution of seeing a person located at region k, and $P(\boldsymbol{X_c}|k)$ is the likelihood of seeing the image $\boldsymbol{X_c}$ given there is a person at location k. Here we assume images captured from different cameras are conditionally independent given the location of the person. In this paper, we make a simple assumption that $P(k)$ follows uniform distribution, although it can also be learned from the training data.

The likelihood term $P(\boldsymbol{X_c}|k)$ in (1) computes the joint probability of all pixels from a single camera given that a person is located at k.

$$P(\boldsymbol{X_c}|k) = \sum_{x_c \in \Phi(k)} P_f(x_c) + \sum_{x_c \notin \Phi(k)} P_b(x_c) \tag{2}$$

where x_c is a pixel value in $\boldsymbol{X}_c$. $P_b(x_c)$ computes the background probability of the pixel x_c, and $P_f(x_c)$ computes the foreground probability. As we do not know any prior knowledge about how a person looks like, the foreground probability is a constant for all pixels. The background probability is estimated by the Adaptive Gaussian Mixture Model [37], see Fig. 2. $\Phi(k)$ is a binary mask where the ones indicate the foreground pixels and zeros the background pixels. The mask is generated by evaluating a 3D human-shape template at location k and then projecting the 3D template onto the image plane. Pixels within the area of the projection are considered as the foreground, and otherwise the pixels are labeled as the background. We sum up the foreground and background probabilities over all pixels in the image, and then we are able to evaluate the product of the likelihood terms with respect to all cameras as in (1).

We evaluate the posterior probability using (1) for all the regions on the ground-plane. People are detected at regions with local maximal posterior probabilities. The algorithm is implemented using logarithm likelihood probabilities to avoid numerical problems.

Next, we introduce the algorithm that generates tracks out of the frame-based detections.

B. Tracking

We use an online tracker that updates tracks every frame. The tracker associates the detected people with the existing tracks by evaluating the matching scores between the detection and the model of the tracks. After that, the new detection is appended to the corresponding track and tracks get updated separately using a Kalman Filter [38].

We compute the matching score using two cues, *i.e.* the appearance of a human and the location. Assume there are M detections and N tracks generated at time t. The matching score between the m^{th} detection and n^{th} track is computed as

$$S_t(m,n) = \Psi_1(h(o_m^t), h_n^{t-1}) \Psi_2(o_m^t, s_n^{t-1}, v_n^{t-1}) \tag{3}$$

where o_m^t is the location of the detected person, s_n^{t-1} is the previous tracked location, and v_n^{t-1} is the velocity of the track. Both s_n^{t-1} and v_n^{t-1} can be estimated by the Kalman Filter. Ψ_1 measures the similarity between the appearance template h_n^{t-1} and the histogram extracted at the detected location $h(o_m^t)$. Here we use the Bhattacharyya distance [39] for comparing two color histograms

$$\Psi_1(h(o_m^t), h_n^{t-1}) = \sqrt{1 - \sum_u \sqrt{h_u(o_m^t) h_{n,u}^{t-1}}} \tag{4}$$

Ψ_2 measures the consistency between the observed location o_m^t and the new location predicted by the Kalman Filter.

$$\Psi_2(o_m^t, s_n^{t-1}, v_n^{t-1}) = \mathcal{N}(v_n^{t-1} + s_n^{t-1} - o_m^t, \sigma^2) \tag{5}$$

We can evaluate the matching score using (3) for all track-detection pairs. Finding the optimal assignment from the table of matching scores can be efficiently solved using the Hungarian Algorithm [40].

IV. People Identification

The system for recognizing people with the robot's RGB-D camera consists of a head and face detection module, a face recognition module, and a face identification tracker, see Fig. 2 (blue panel).

A. Face Detection

The detection module searches for face images using the Kinect sensor. The detection has two stages: first, a Viola-Jones detector [19] is applied in the depth image for finding the objects that look like a human head. These candidate regions are then verified in the color image using another Viola-Jones classifier that is trained on color images. The whole detection procedure is detailed and evaluated in our previous work [3], [21].

B. Face Identification

Subsequently, those image patches which contain a face are processed by the recognition module. The recognition starts with pre-processing the face image with a gamma transform ($\gamma = 0.2$) and a downscaling of the first 5 coefficients of the Discrete Cosine Transform on the gray scale image by a factor of 50. This realizes an illumination normalization which renders the recognition algorithm more robust against lighting conditions that are different from the training data. Then, the algorithm tries to detect facial features like the eyes or the nose with a Viola-Jones classifier and generates a virtual frontal perspective on the recorded face if those face features can be identified successfully. This measure diminishes negative effects on recognition stemming from a badly aligned face image. Eventually, the recognition is conducted by projecting the pre-processed face image into a lower-dimensional space which minimizes the variance of training face images of the same person and maximizes the inter-class variance according to Fisher's Linear Discriminant [7]. The identity estimate is found as the nearest neighbor from training data in this space or as a probability distribution constructed from the labels of the neighborhood. Again, a thorough explanation and evaluation of those methods is given in [3].

C. Identity Tracking

The detector and recognizer work based on single image frames. However, due to noise or misalignment, only using the frame-based recognition may be problematic. Therefore, the recognized identities are filtered by a tracking module for stable identity estimation.

The tracker firstly matches the recognition results in consecutive frames. The matching score is computed between the previous detection i and current detection j:

$$C(i,j) = \|X_i - X_j\|_{L_2} + \alpha \|P_i - P_j\|_{\chi^2} + \beta \|H_i - H_j\|_{\chi^2} \quad (6)$$

where X denotes the 3d face coordinates in space, P is the probability distribution over all labels, and H is the histogram of local binary patterns of the head region. α and β are optional weighting factors for the single metrics. The first term measures the distance to the last detection, the second term computes the similarity in label predictions, and the third one establishes visual similarity of the tracked image regions. The global minimum cost assignment between previous and current recognition is found with the Hungarian method [40]. New detections in the current set are added afterwards and initialized with their estimated label probability distribution. To smooth sporadic false recognitions, the estimated probability distributions are filtered temporally with a Hidden Markov Model (HMM). The final identity assignment is estimated for each frame by considering the label probabilities for each detection as inverse costs and computing the globally optimal assignment with the Hungarian Method.

V. Joint Tracker

So far we have introduced two systems. One system detects, localizes, and tracks all persons in the room. The system, however, only gets the track ID and does not know who the person is. The other system identifies the person using the robot sensor, but the robot has to keep the person in sight all the time. The joint tracker solves the problem of the two separate systems by combining both of the sensors. The joint tracker assigns the tracks as unknown persons when they have not been recognized by the robot. Once people are identified by the robot, names of people are immediately associated with the tracks.

To increase the robustness of the system, we fuse data from the two sensors in a probabilistic way. In robotic scenarios, both the locations of the tracked persons and the locations of the robot can be very noisy. Our robot is localized using the SLAM approach [41], and the robot location is represented as a set of weighted particles. As the Kinect sensor is registered in the robot's transformation tree, we can always transform the face locations that are detected by the Kinect sensor into the particle representation in world coordinates. Let the set of the particles be $L = \{(w_1, \boldsymbol{l}_1), (w_2, \boldsymbol{l}_2)...(w_N, \boldsymbol{l}_N)\}$, where $\boldsymbol{l}$ is the location of a particle in world coordinates and w is the weight of the particle.

The posterior distribution of the human location returned by the Kalman Filter is a Gaussian distribution, with its mean indicating the most likely position of people and variance indicating the uncertainty. The set of Kalman Filters from different tracks proposes multiple Gaussian density distributions at each time step. Our goal is to associate the detected faces with those Gaussian PDFs. We compute the score of associating the i^{th} track with the j^{th} face as

$$Q(i,j) = \sum_{n=1}^{N} \frac{w_n}{C\sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left((\boldsymbol{l}_n^j - \boldsymbol{\mu}_i)^T {\boldsymbol{\Sigma}_i}^{-1} (\boldsymbol{l}_n^j - \boldsymbol{\mu}_i)\right) \quad (7)$$

where $\boldsymbol{\mu}_i$ is the mean location and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the i^{th} track. C is a constant that does not affect the assignments. The best assignments between the tracks and the detected faces are calculated in the same way as in people tracking. Once the track has been associated with the face, the name of the person is attached to the track until a new face is detected for that track. In such a way, the correct name can be recovered if a person is wrongly recognized at the beginning.

VI. Strategies for User Recognition and Approaching

There are two fundamental tasks in Human-Robot Interaction (HRI). The first task is that the robot needs to identify unknown users that are present in the room. The other task is that the robot is asked to approach a specific person, *e.g.* for completing a delivery. This section presents different algorithms for tackling these two tasks. For each of the tasks, we compare two algorithms a) using the robot only, and b) the robot assisted by an overhead camera.

1) Uninformed User Identification: Algorithm 1 describes the identification of present people without help from an external system. The robot starts searching for users by successively moving to random locations and turning around by 360° each time. In a real situation, users can be supposed

to move around so that a systematic search should not bear any advantages over a random strategy. Face detection and recognition is continuously running and every found user is stored internally and announced via speech. Since there is no way to verify that all present users have been found, the terminating condition may be application-driven, *e.g.* finding a certain set of people, searching for 5 minutes, etc. Laser scanner-based leg detection is not used as support because of false alarms at house interior, limited view through occlusions and difficulties with sitting persons.

Algorithm 1 Uninformed recognition of present users.

```
function IDENTIFYUSERSUNINFORMED
  recognitions ← ∅
  while TerminatingCondition = False do
    goal ← computeAccessibleRandomPosition()
    recognition ← moveToAndRotate360(goal)
    recognitions ← {recognitions, recognition}
  return recognitions
```

2) Informed User Identification: The combined system allows approaching each human detected by the external tracking system directly and recognizing the face (see Algorithm 2). In contrast to Algorithm 1, this algorithm is aware of having labeled all present users.

Algorithm 2 Informed recognition of present users.

```
function IDENTIFYUSERSINFORMED
  detections ← getTrackedHumans()
  recognitions ← ∅
  for all detections do
    goal ← computePositionOnPerimeter(detection)
    recognition ← moveTo(goal)
    recognitions ← {recognitions, recognition}
  return recognitions
```

3) Uninformed User Approach: Algorithm 3 displays the method to approach a specific user using the sensors of the robot only. After checking the last known user position, a random search strategy is employed to navigate the robot through the environment searching for the desired person.

Algorithm 3 Uninformed search for a specific user.

```
procedure APPROACHUSERUNINFORMED(targetName)
  targetLocation ← ∅
  goal ← lastKnownUserLocation
  while robotLocation ≠ targetLocation do
    recognitions ← moveToAndRotate360(goal)
    targetLoc. ← checkForUser(recognitions, targetName)
    if targetLocation ≠ ∅ then moveTo(targetLocation)
    goal ← computeAccessibleRandomPosition()
```

4) Informed User Approach: Algorithm 4 utilizes the additional sensory information from the external cameras to approach the specific person directly given that he or she has been recognized previously and tracked in the meantime.

Algorithm 4 Informed search for a specific user.

```
procedure APPROACHUSERINFORMED(targetName)
  detections ← getTrackedHumans()
  targetLocation ← checkForUser(detections, targetName)
  if targetLocation = ∅ then
    APPROACHUSERUNINFORMED(targetName)
  else
    moveTo(targetLocation)
```

VII. Experiment and Results

In our experiments, the robot is asked to complete two different tasks, both of which occur frequently in home care scenarios with robot assistance. The first task is to let the robot identify all the people that are present in the room. The second one is to let the robot find a specific person. We evaluate the efficiency of the integrated system by measuring the average time that the robot needs to complete the tasks.

A. Experiment Setup

The experiments are setup in a domestic environment, see Fig. 3 and 4. There are two GV-FE420 cameras mounted on the ceiling, one above the sofa area and the other one in the kitchen. Both cameras are calibrated to the ground-plane with the highest resolution (2048×1944 pixels). After calibration, we can easily project the image coordinates onto the ground-plane. For people detection and tracking, we use only $1/4$ of the full resolution for efficient processing and we find that is sufficient to give stable tracks. The cameras have a very wide field of view (over 180 degrees), and provide a good overview of the entire room. We adopt a PC with an Intel Core i7-3770K (3.50GHz) processor for people detection, tracking and also fusing the incoming face recognition data provided by the robot. The processing rate of this system is around 9 Hz.

The mobile service robot is a Care-O-bot® 3 which features an omni-directional mobile base with a flexible torso, a 7 DOF manipulator and a movable sensor head that contains a pair of stereo cameras and a Kinect RGB-D camera. The control script, navigation and person recognition software are run on two build-in PCs. The people detection and recognition module processes the RGB-D data from the Kinect camera at a resolution of 640×480 pixels. The module uses an Intel Core i7-E610 (2.53GHz) PC and delivers recognition results at 6 Hz whereat head and face detection generate 90% of the computational effort which can be reduced by using images of lower resolution.

For each experiment, the robot was commanded by a control script whose functionality corresponds to the search strategies proposed in Section VI. Both experiments were conducted with 5 different subjects. To demonstrate the advantages of the combined recognition and tracking system, we carried out the experiments in two flavors, *i.e.* once with the robot sensor only and once combining the robot sensor with the ceiling-mounted cameras. The environment setup of the two experiments was always the same for comparison.

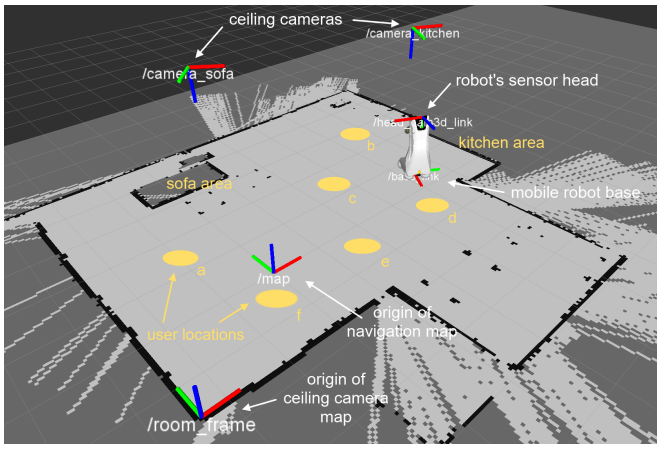


Fig. 3. Floor plan of the experimentation site with user locations a-f highlighted and the coordinate systems of the ceiling and robot cameras, the navigation and camera maps and the robot base drawn into.

Besides the sofa area, we further defined six locations in the room, see the labels a-f in Fig. 3. These points indicate the test locations where people were standing during our experiments.

B. Identification of All Present Users

The first experiment measures the performance of system on identifying and tracking all present users in the room. This is a prerequisite for tasks like activity recognition of multiple users or fast reacting to an individual call from one of the human users. The setup is that two or three subjects are distributed in the environment among positions a-f. We ask the users to look at the robot when it approaches so that the robot can recognize their faces. In a real scenario, the robot can either attract a user's attention by speech or move around the person until the face becomes visible. These options have been excluded from the experiments to avoid subjective factors in the results.

As a baseline, the robot is asked to find all users with its own sensors only. Algorithm 1 is applied to navigate the robot randomly until all people have been identified. This scenario is evaluated under 10 combinations of user locations, see Table I. We report both the time for finding all users and the number of wrongly identified persons. The baseline results vary largely in the time for finding all persons. It can be as fast as within one minute but it may also take more than 5 minutes. On average, all subjects are identified within 2 minutes and 10 seconds. All of the 26 person labels within the 10 sessions have been assigned correctly, and only 3 times people were wrongly identified initially but corrected after a couple of seconds. This yields an initial recognition rate of almost 90%. The two main drawbacks of the baseline approach are: a) the algorithm needs to be terminated manually as the robot never knows whether all present subjects have been found and b) the robot has to keep all users constantly in sight in order to keep track of them, which is not desirable for real-world applications.

The ambient camera system can detect and track humans. The information is shared with the robot via wireless connection so that the robot stays informed about the locations of all present persons along with their identities attached to the tracks, even when people are not visible to the robot. We use Algorithm 2 in our second experiment. After all persons have been visited and identified, this algorithm terminates automatically and announces that all users have been found. The times that are needed to identify all present users are listed in column "combined system" of Table I. The results show that the combined system is significantly better than the random search. The combined system requires only 1 minute and 5 seconds on average to complete the task. In contrast, the baseline approach takes twice the time than the combined system. The standard deviation of the combined system is around 19 seconds, which is much more stable than the time of 1 minute and 26 seconds in random search. In our experiments, the performance of the face recognition system is very stable as the guided approach can directly navigate the robot into an advantageous distance to the subjects so that the faces are always captured with high quality. The ceiling-camera system fails at detecting people at position f for 4 times because people at that location are heavily occluded by the robot from both of the ambient cameras. Apart from these errors, the system performs outstanding, yielding an accuracy of correctly tracked and identified people of 85%.

TABLE I

RESULTS FOR FINDING ALL PRESENT USERS.

user	robot only		combined system		
locations	time	rec. errors	time	rec. errors	track. errors
b, c, d	5:48	0	1:10	0	0
b, c, e	2:18	0	1:05	0	0
b, c, f	1:14	1	1:27	0	0
c, d, e	1:28	0	0:40	0	0
c, d, f	2:56	0	1:20	0	1
d, e, f	0:58	0	1:35	0	1
b, e	1:30	2	0:53	0	0
b, f	2:00	0	1:04	0	1
c, f	1:00	0	0:40	0	1
b, d	2:30	0	0:51	0	0
average (3p)	2:27	1/18	1:13	0/18	2/18
average (2p)	1:45	2/8	0:52	0/8	2/8
average (all)	2:10	3/26	1:05	0/26	4/26
stddev (all)	1:26		0:19		

C. Approaching a Specific User

The second experiment evaluates the performance of our system on approaching a specific user, which is widely used in delivery tasks. In our scenario, the user is sitting on the couch, calling the robot, and ordering something. The robot then goes to the kitchen and in the meantime the user stands up and moves to another place. The task of the robot is to find the user for delivery. The experiment is conducted at three levels of difficulty: 1) with a cooperative user facing the robot all the time, 2) with a busy user facing a fixed direction, and 3) with two users facing the robot constantly.

The results of the first session are shown in the upper part of Table II. Column "robot only" corresponds to the baseline case when only the robot sensors are used, see Algorithm 3. We report the times that are spent by the robot on approaching the user among positions a-f as well as the

TABLE II

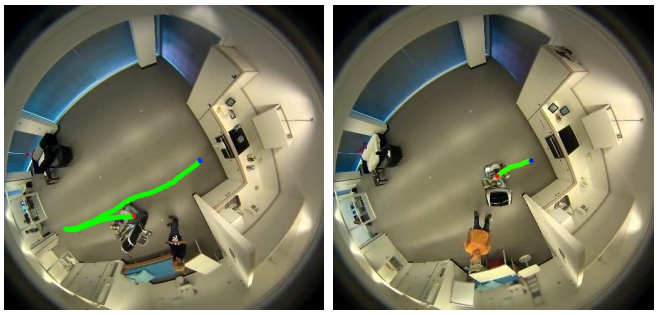
RESULTS FOR SEARCHING A SPECIFIC USER.

user	robot only		combined system		
location(s)	time	rec. errors	time	rec. errors	track. errors
with one user facing the robot constantly					
sofa	0:28	0	0:28	0	0
a	1:28	0	0:34	0	0
b	0:40	0	0:18	0	0
c	0:33	0	0:21	0	0
d	0:28	0	0:18	0	0
e	0:35	0	0:29	0	0
f	1:18	0	0:31	0	0
average	0:47	0/7	0:26	0/7	0/7
with one user facing a fixed direction					
a (+x)	1:12	0	0:33	0	0
a (-y)	7:37	0	0:43	0	0
b (-x)	0:26	0	0:21	0	0
b (-y)	9:59[1]	0	0:17	0	0
c (-x)	0:29	0	0:20	0	1
c (-y)	2:27	0	0:20	0	0
d (-x)	2:39	0	0:16	0	0
d (+y)	2:56	0	0:20	0	0
e (+x)	6:42	0	0:28	0	0
e (+y)	0:25	0	0:28	0	0
f (+x)	1:39	0	0:31	0	0
f (+y)	0:55	0	0:31	0	0
average	2:30	0/12	0:26	0/12	1/12
with two users facing the robot constantly					
b, a	0:38	0	0:15	0	0
b, c	2:17	0	0:20	0	0
b, d	1:12	0	0:21	0	0
c, d	0:22	0	0:17	0	0
d, a	0:43	0	0:16	0	0
d, f	0:37	0	0:14	0	0
e, a	1:03	0	0:26	0	0
e, b	1:21	1	0:22	0	0
f, b	1:43	0	0:52	0	0
f, e	1:10	0	0:27	1	0
average	1:07	1/20	0:23	1/20	0/20
average (all)	1:34	1/39	0:25	1/39	1/39
stddev (all)	1:44		0:09		

[1] aborted after 10 minutes of search, excluded from cumulative statistics

sofa area after leaving the kitchen area. All delivery times are quite fast with an average of 47 s. The delivery tasks are accomplished successfully as all of the users are correctly recognized. These results are compared with the performance of the "combined system" that keeps tracking the user after being identified. The times of delivery using Algorithm 4 are a little more than half of Algorithm 3, with only 26 s on average and deliveries are all successful.

The second session of this experiment requires the users to face a fixed direction. This makes the approaching a harder problem as the robot may not obtain a good perspective of the face. In Table II, signs + and - refer to the user's orientation, *e.g.* "a (+x)" means that the user stands at location a, facing the positive direction of the x axis. The results show that the combined system significantly outperforms the system that only uses the robot sensors. The average search time for the robot-only system is 2 minutes 30 seconds which is 5 times longer than the 26 seconds when using the combined system. The worst case of the robot-only system occurs at "b (-y)", which has to be manually terminated after 10 minutes of unsuccessful search. For the combined system, the results of the second session are similar to those of the first session.



(a) (b)

Fig. 4. Robot trajectories of finding a specific person when (a) using random search and (b) assisted with the external tracking system.

This is because the user's gaze direction is irrelevant when the position is already known from the overhead cameras. A speed up could possibly be obtained for the uninformed search by utilizing full body recognition and moving around found persons trying to recognize the face. For both systems, no recognition error is observed, but the tracking system loses the user once after delivery in case "c (-x)".

The third session is a more realistic situation where multiple people are present and the robot is required to approach the correct person. The lower part of Table II presents the results of 10 experiments with two users. The user who has ordered in the sofa area goes to the position that is underlined in Table II, the other user is located at the second position. Both users are told to face the robot constantly. Accordingly, the average time of approaching is 1:07 for the robot-only system, which is similar to the experiments with only one user. A recognition error occurs once when the order is delivered to a wrong person. Using the combined sensor system, the robot can be guided to the correct person in approximately the same time as in the tests before. Apart from one recognition error during delivery, the tasks are successfully completed.

In conclusion, the delivery tasks can be accomplished with the accuracy of 96.5% using both strategies of approaching. With the combined system, however, the delivery is three times faster than the unguided search, and the results have a smaller deviation. Hence, using the combined system is more efficient and more stable than just using the robot sensors for our tasks.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we use ambient cameras for detecting and tracking people, and we use a robot-mounted RGB-D sensor for identifying people. The information from the both sensors is combined in a joint tracking system for efficient and accurate people finding. The results show that by leveraging these two complementary types of sensors, the robot can be guided to approach people directly instead of searching through the room. This significantly reduces the approaching time and provides a more intuitive and comfortable way for the users to interact with robots.

Based on the people tracking and identification system proposed in this paper, we are currently working on activity recognition and prediction for multiple users. The working system allows the robot for deliberately offering assistance on critical tasks and may serve for long-term medical check-ups by detecting deviations from the user's daily activity schemes. We furthermore plan to enhance the approaching algorithms with proxemics [42], *i.e.* having the robot to approach the user in a human-acceptable way.

REFERENCES

[1] G. Gate, A. Breheret, and F. Nashashibi, "Centralized fusion for fast people detection in dense environment," in *Proc. IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 76–81.

[2] M. Montemerlo, S. Thrun, and W. Whittaker, "Conditional particle filters for simultaneous mobile robot localization and people-tracking," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 1. IEEE, 2002, pp. 695–701.

[3] R. Bormann, T. Zwölfer, J. Fischer, J. Hampp, and M. Hägele, "Person recognition for service robotics applications," in *accepted for publication at the 13th International IEEE-RAS International Conference on Humanoid Robots*, 2013.

[4] N. Hu, G. Englebienne, and B. Kröse, "Posture recognition with a top-view camera," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 2152–2157.

[5] N. Hu, G. Englebienne, Z. Lou, and B. Kröse, "Learning Latent Structure for Activity Recognition," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[6] J. Li, T. Wang, and Y. Zhang, "Face Detection using SURF Cascade," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2183–2190.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[8] N. Hu, H. Bouma, and M. Worring, "Tracking individuals in surveillance video of a high-density crowd," in *Proc. of SPIE Vol*, vol. 8399, 2012, pp. 839 909–1.

[9] Y. Zhu and K. Fujimura, "A bayesian framework for human body pose tracking from depth image sequences," *Sensors*, vol. 10, no. 5, pp. 5280–5293, 2010.

[10] R. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. Moeslund, and G. Tranchet, "An RGB-D Database Using Microsoft's Kinect for Windows for Face Detection," in *Proc. Intern. Conference on Signal Image Technology and Internet Based Systems*, 2012, pp. 42 –46.

[11] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard, "Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities," in *Proc. IEEE Internat. Conference on Robotics and Automation*. IEEE, 2008, pp. 1710–1715.

[12] J. Cui, H. Zha, H. Zhao, and R. Shibasaki, "Multi-modal tracking of people using laser scanners and video camera," *Image and vision Computing*, vol. 26, no. 2, pp. 240–252, 2008.

[13] M. Kristou, A. Ohya, and S. Yuta, "Target person identification and following based on omnidirectional camera and LRF data fusion," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2011.

[14] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, "People tracking and following with mobile robot using an omnidirectional camera and a laser," in *Proc. IEEE International Conference on Robotics and Automation*. IEEE, 2006, pp. 557–562.

[15] M. Luber, L. Spinello, and K. O. Arras, "People tracking in RGB-D Data with on-line boosted target models," in *Proc. International Conference on Intelligent RObots and Systems*, 2011, pp. 3844–3849.

[16] N. Atsushi, K. Hirokazu, H. Shinsaku, and I. Seiji, "Tracking multiple people using distributed vision systems," in *Proc. IEEE Internat. Conf. on Robotics and Automation*, vol. 3. IEEE, 2002, pp. 2974–2981.

[17] N. Hu, G. Englebienne, and B. J. Kröse, "Bayesian fusion of ceiling mounted camera and laser range finder on a mobile robot for people detection and localization," in *Human Behavior Understanding*. Springer, 2012, pp. 41–51.

[18] A. A. Mekonnen, F. Lerasle, A. Herbulot, *et al.*, "External cameras and a mobile robot for enhanced multi-person tracking," in *Proc. Internat. Conference on Computer Vision Theory and Applications*, 2013.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[20] V. Jain and E. Learned-Miller, "Online domain adaptation of a pre-trained cascade of classifiers," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 577–584.

[21] J. Fischer, D. Seitz, and A. Verl, "Face detection using 3-d time-of-flight and colour cameras," in *Proc. ISR/ROBOTIK*, 2010, pp. 112–116.

[22] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[23] I. Naseem, R. Togneri, and M. Bennamoun, "Linear Regression for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.

[24] C.-Y. Zhang and Q.-Q. Ruan, "Face Recognition Using L-Fisherfaces," *Journal of Information Science and Engineering*, vol. 26, no. 4, pp. 1525–1537, 2010.

[25] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[26] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[27] A. Georghiades, P. Belhumeur, and D. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[28] K. Lee, J. Ho, and D. Kriegman, "Acquiring Linear Subspaces for Face Recognition under Variable Lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[29] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.

[30] H. Liu, W. Gao, J. Miao, and J. Li, "A Novel Method to Compensate Variety of Illumination in Face Detection," in *Joint Conference on Information Sciences*, 2002, pp. 692–695.

[31] T. Goel, V. Nehra, and V. P. Vishwakarma, "Comparative Analysis of various Illumination Normalization Techniques for Face Recognition," *International Journal of Computer Applications*, vol. 28, no. 9, 2011.

[32] W. L. Chen, M. J. Er, and S. Q. Wu, "Illumination Compensation and Normalization for Robust Face Recognition Using Discrete Cosine Transform in Logarithm Domain," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 36, no. 2, pp. 458–466, 2006.

[33] A. Pentland, B. Moghaddam, and T. Starner, "View-based and Modular Eigenspaces for Face Recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 84–91.

[34] S. Gurbuz, E. Oztop, and N. Inoue, "Model free head pose estimation using stereovision," *Pattern Recognition*, vol. 45, no. 1, pp. 33–42, 2012.

[35] D. Beymer, "Face Recognition under Varying Pose," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 756–761.

[36] G. Englebienne and B. J. Kröse, "Fast bayesian people detection," in *Proc. Benelux AI Conference, BNAIC*, 2010.

[37] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. International Conference on Pattern Recognition*, vol. 2. IEEE, 2004, pp. 28–31.

[38] G. Welch and G. Bishop, "An introduction to the kalman filter," 1995.

[39] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.

[40] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[41] G. Grisetti, C. Stachniss, and W. Burgard, "Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters," *IEEE Transactions on Robotics*, pp. 34–46, 2007.

[42] E. Hall, *The hidden dimension*. Doubleday, 1966.