A Probabilistic Exemplar Approach to Combine Laser and Vision for Person Tracking

Dirk Schulz
Department of Computer Science
University of Bonn
Bonn, Germany
Email: schulz@iai.uni-bonn.de

***Abstract*— This article presents an approach to person tracking that combines camera images and laser range data. The method uses probabilistic exemplar models, which represent typical appearances of persons in the sensor data by metric mixture distributions. Our approach learns such models from laser and from camera data and applies a Rao-Blackwellized particle filter in order to track a person's appearance in the data. The filter samples joint exemplar states and tracks the person's position conditioned on the exemplar states using a Kalman filter. We describe an implementation of the approach based on contours in images and laser point set features. Additionally, we describe how the models can be learned from training data using clustering and EM. Our experimental results show that the appearance of persons in camera image scan be tracked reliably using this approach and that it also allows to distinguish between persons during tracking.**

I. Introduction

The ability to keep track of the motions of people is of general importance for mobile robots operating in populated environments. Over the last decade, several mobile robots have been deployed in populated environments like office buildings [1], [2], [3], supermarkets [4], hospitals [5], and museums [6], [7]. The requirements on the quality of the motion tracking differs largely from task to task. For example, if one wants to adapt the robot's velocity to the walking speed of the people in its surrounding [8], or if one just wants to distinguish between static and dynamic parts of the environment [9], it is generally not important to keep track of individual persons. However, if the robot is intended to interact with a particular person over a longer period of time, like carrying loads for individual persons or guiding individuals, it becomes essential that the robot does not interchange its client with someone else.

For the first kind of application several tracking approaches have been developed over the last year. Most of these approaches rely on laser range sensors [10], [11], [12]. The main advantage of this sensor is the accuracy of its distance measurements. However, the sensor does not directly provide information that allows to distinguish between persons. Cameras provide this information, but vision-based tracking is very difficult from a mobile robot for several reasons. Small movements of the robot can lead to very large shifts in the image plane, lighting conditions can change, and, as the whole image content is non-stationary, it can be hard to distinguish the person being tracked from the background.

In this article we present a tracking approach that combines laser range data with camera images to overcome some of these problems. The approach employs two exemplar models of a walking person for this purpose. One model for the appearance of a walking person in laser range data and a second model for the appearance of a walking person in the robot's camera images. The general idea is that the laser beams which are reflected from a person provide information about the person's motion state, for example if the laser scanner measures several points on the surface of the person's legs. The laser data forms patterns which correlate with the appearance of the person in the image at the same point in time. By taking both kinds of features into account, a particle filter can be derived that requires only a small number of particles to track a person's position in the robot's surrounding and the position and shape of the person in the robot's camera images simultaneously. This is achieved by applying a Rao-Blackwellized particle filter that maintains a posterior over the person's position, its image exemplar state, and its laser exemplar state. The algorithm samples joint image and laser exemplar states and maintains a Kalman filter for each particle, whose updates are conditioned on the exemplar states, in order to track the person's position.

Additionally to the tracking algorithm, we describe an approach to learn the joint exemplar models from training data. This involves a clustering the training data into distinctive sets and an EM approach to learn the temporal transitions between the joint exemplar states. The remainder of this article is organized as follows. After discussing related work in the next section, we introduce the joint exemplar Rao-Blackwellized particle filter in Section III. In Section IV we describe the actual exemplar models used in our implementation in more detail and we explain how they are learned from training data. Before we conclude in Section VI, we give some experimental results obtained using the approach in Section V.

II. Related Work

Over the last years, several approaches for tracking moving people with mobile robots have been developed. Most of these

approaches use 2D laser scanners to observe and track people in the surrounding of the robot.

For example, Kluge et al. [13] describe an approach to estimate moving obstacles with an autonomous wheelchair. Their approach does not apply a motion-model to the objects so that they cannot reliably keep track of individual objects over time. Montemerlo et al. [14] addresses the problem of simultaneous localization and people tracking using range sensors. The authors use conditional particle filters to incorporate the robot's uncertainty about its own position into the tracking process, where each particle maintains Kalman filters for the objects being tracked. Fod et al. [15] present an approach to track multiple moving people within a workspace using statically mounted laser range-finders. They use Kalman filters to keep track of objects during temporary occlusions. Schulz et al. [12] propose a variant of Joint probabilistic data association filters [16] that replaces Kalman filters by particle filters to track multiple moving objects in laser range data.

These approaches have in common that they only keep track of the spatial motion of the objects. They do not try to distinguish between different appearances of the object within the laser data. To our knowledge, the only approach that distinguishes between different motion states in the laser data is by Taylor and Kleeman [17]. Their laser-based method tracks the repetitive motion pattern of a walking person's legs. For this purpose, the individual legs are tracked using a switching state Kalman filter.

As laser range scanners only provide proximity information, they can not be used to reliably identify or distinguish between persons during tracking. Several authors propose to combine a laser-based approach with vision in order to overcome this limitation. For example, in the context of learning motion patterns for individual persons, Bennewitz et al. [18] use color histograms to distinguish between persons, where the image region to consider for computing the histogram is selected based on the position of a person in the laser scan. Brooks and Williams [19] use skin colored blobs as image features indicating persons. Again, this approaches do not track the change of appearance of persons in the camera images. In contrast to this, our approach tracks the appearance of a person's legs in laser scans and the shape of the person in camera images during the tracking process. The appearance models for this purpose are learned from training data. Our method is largely inspired by the metric mixture approach introduced by Toyama and Blake [20]. However, there are some differences. First, we combine two sensor modalities and therefore two mixture models in one algorithm by tracking the joint exemplar states. Second, we do not learn exemplar transformations [21]. In purely vision-based exemplar approaches, these transformations describe the dynamic change of a person in the image. In our joint laser and vision-based approach, we learn the transition model for the exemplars only. Motion within the image is predicted based on the position prediction of the Kalman filter and a mapping of the person's location relative to the robot to its location within the image. This mapping is also learned during training.

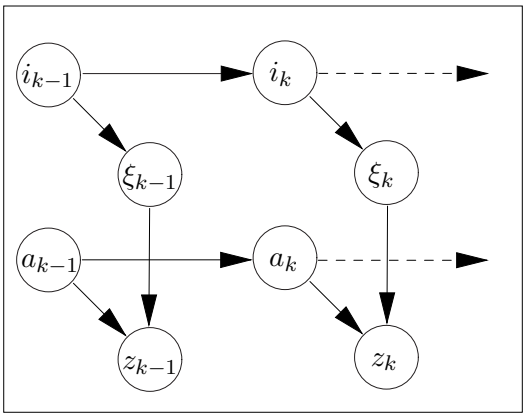

Fig. 1. The generic exemplar model according to Toyama and Blake [20]. i_k denotes the active mixture component at time k, while ξ_k denotes the mixture model. The a_k are the geometrical transformations, and z_k the observed images.

Other particle filter-based tracking approaches in the computer vision community aim at actually tracking the articulated human motions based on many degrees of freedom models of the human body, for example [22], [23]. However, most of these techniques are only intended for static cameras and have not yet been applied on mobile robots.

III. The Joint Exemplar Rao-Blackwellized Particle Filter

In this section we explain the inference part of our joint laser and vision tracking approach. Basically the approach builds on a Rao-Blackwellized particle filter, which maintains a joint probability distribution over the position of a person and its current appearance in laser scans and camera images. This particle filters samples appearances from two mixture distributions of possible appearances, and updates the position part of the distribution analytically using a Kalman filter for each sample. Our current implementation uses typical silhouettes of persons in camera images as prototypes in the vision-based model and typical 2D point sets representing laser measurements of a person's legs as prototypes for the laser-based model. Details on these particular models and how they are learned are given in Section IV. Here, we will briefly introduce the joint exemplar-based tracking approach using metric mixture distributions and we will explain how a person can be efficiently tracked based on these models using a Rao-Blackwellized particle filter. In the following mathematical derivations, time is indexed by subscripts, where the current time is denoted by k. The superscript l is used for laser-related random variables and the superscript c for camera-related random variables.

A. The Joint Exemplar Model

Following [21] an exemplar model consists of a set of "exemplars", $\Xi = \{\xi_1, \ldots, \xi_M\}$, which contain representatives of training data, and a distance function ρ, which measures the distance of any two points in exemplar space. It is assumed that an observation z_k at time k is drawn from a mixture

distribution, such that $z_k \approx T_\alpha \xi(k)$, where T_α is a geometric transformation and $\xi(k) \in \Xi$ is the exemplar at time k. The dynamics is usually modeled as a first order Markov chain $p(\xi_k, \alpha_k \mid \xi_{k-1}, \alpha_{k-1})$, where the transition probabilities as well as the mixture centers Ξ and their associated distribution parameters are learned from the training data. The graphical model for this generic exemplar technique is depicted in Figure 1.

The graphical model of our approach is depicted in Figure 2. Here, the variable x_k^l denotes the position and the velocity vector of the person in the robot's vicinity at time k, while x_k^c is the person's position within the camera image taken at time k. Laser scans and images are denoted by z_k^l and z_k^c and the index of the active exemplar states by $\mathcal{E}_k^l$ and $\mathcal{E}_k^c$, while ξ_k^l and ξ_k^c denote the exemplars themselves. For the sake of brevity we also write $\mathcal{E}_k$ for $(\mathcal{E}_k^l, \mathcal{E}_k^c)$, $\mathbf{z}_k$ for (z_k^l, z_k^c), and complete time sequences of random variables upto time k are abbreviated by the subscript $1:k$, e.g. $\mathbf{z}_{1:k}$ denotes the sequence of all laser and camera measurements upto time k.

Our model differs from the generic one in two respects: (1) We keep track of laser exemplars ξ_k^l and vision exemplar ξ_k^c simultaneously, and (2) we do not learn and keep track of the geometrical transformations α_k. Instead, we rely on a Kalman filter that estimates the position of a person in laser scans for this purpose. Additional random variables x_k^c are introduced to account for the uncertainty of mapping a spatial position to the position and scale of the person in the image. The main reason for not learning the geometrical transformations is that they are highly affected by the robot's own motion.

We follow the metric mixture approach proposed in [20] to evaluate the observation likelihoods. In this approach, it is assumed, that the observations are drawn from a metric mixture distribution, where a transformed exemplar $\tilde{\xi}$ serves as a center in a mixture component:

$$p(z \mid \tilde{\xi}) \propto \frac{1}{Z} \exp(-\lambda \rho(\tilde{\xi}, z)). \tag{1}$$

We use truncated quadratic chamfer distance for the exemplar distance ρ, which has two nice properties:

1) It can be computed fast both for images and for laser data. For this purpose, images are transformed to binary edge images and a distance transformation is applied. For laser scans, we average over the distances of the closest point of the scan to each point of the exemplar.
2) For quadratic chamfer distance, the metric distribution Equation 1 is approximately Gaussian, and the parameter λ and the partition function Z can be estimated from training data [20].

B. The Rao-Blackwellized Particle Filter

During tracking, it is our goal to determine the position of a person based on a sequence of laser range scans and camera images. Note, that the exemplar sets ξ_k^l,ξ_k^c, their associated mixture distributions, and the mapping uncertainty x_k^c are learned during training and remain fixed during tracking. We therefore treat them as background knowledge during the derivation of the filter.

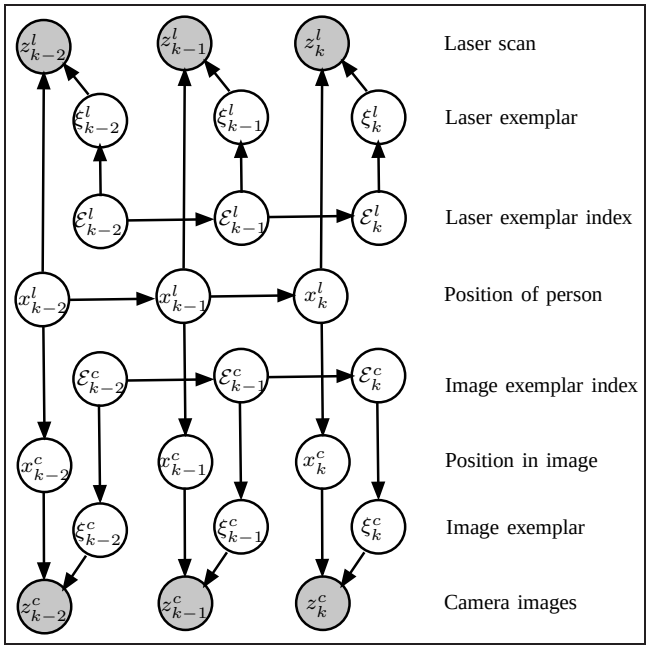

Fig. 2. Graphical model of the joint exemplar person tracking algorithm. x_k is the position of the person at time k. At each time step we receive a camera image z_k^c and a laser scan z_k^l. The shapes of the person's legs in the laser data and the silhouette of the person in the camera image are assumed to be drawn from the metric mixture distributions ξ_k^l and ξ_k^c respectively, where the hidden states $\mathcal{E}_k^l$ and $\mathcal{E}_k^c$ determine the active components. Finally, the random variable x_k^c accounts for the uncertainty of mapping spatial positions to image coordinates.

However, the exemplar states $\mathcal{E}_k$ are hidden, we therefore need to estimate the joint posterior over positions x_k^l and sequences of exemplar states $\mathcal{E}_{1:k}$. This joint posterior can be factorized by conditioning the positions x_k^l on the exemplar states $\mathcal{E}_{1:k}$:

$$p(x_k^l, \mathcal{E}_{1:k} \mid \mathbf{z}_{1:k}) = p(x_k^l \mid \mathcal{E}_{1:k}, \mathbf{z}_{1:k}) p(\mathcal{E}_{1:k} \mid z_{1:k}) \tag{2}$$

The key idea of Rao-Blackwellized particle filters (RBPF) is to compute Equation 2 by *sampling* exemplar states from $p(\mathcal{E}_{1:k} \mid z_{1:k})$ and then to update the positions x_k^l conditioned on the exemplar state of each sample. This way, the position estimates can be updated analytically using one Kalman filter for each sample.

More specifically our RBPF maintains a set of weighted samples, $S_k = \{\langle s_k^{(\iota)}, w_k^{(\iota)} \rangle \mid 1 \leq \iota \leq N\}$, where each sample $s_k^{(\iota)} = \langle \theta_k^{(\iota)}, \mathcal{E}_k^{(\iota)} \rangle$ of the sample set at time k consists of a Kalman filter $\theta_k^{(\iota)} = \langle \mu_k^{(\iota)}, \Sigma_k^{(\iota)} \rangle$ and the current exemplar states $\mathcal{E}_k^{(\iota)}$.

The generic RBPF algorithm generates a set S_k from the previous sample set S_{k-1} based on a new laser scan z_k^l and camera image z_k^c by first generating new exemplar states $\mathcal{E}_k$ distributed according to the posterior $p(\mathcal{E}_{1:k} \mid \mathbf{z}_{1:k})$. After that, the position part $\theta_k^{(\iota)}$ of each sample is maintained by applying Kalman filter updates on each sample. In our case, the position measurements to be integrated into the Kalman

filter are obtained by locally matching the exemplar prototype corresponding to $\mathcal{E}_k^{l(\iota)}$ of each sample against the current laser scan.

C. Sampling Exemplars

The efficiency of RBPFs strongly depends on the number of samples needed to represent the posterior $p(\mathcal{E}_{1:k} \mid z_{1:k})$. Due to the sequential nature of the estimation process, samples must be generated from the exemplar states of the previous time step. The posterior at time k can be written as

$$p(\mathcal{E}_{1:k}|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathcal{E}_{1:k}, \mathbf{z}_{1:k-1})p(\mathcal{E}_k|\mathcal{E}_{1:k-1}, \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} p(\mathcal{E}_{1:k-1}|\mathbf{z}_{1:k-1}) \tag{3}$$

$$= \frac{p(\mathbf{z}_k|\mathcal{E}_k, \theta_{k-1})p(\mathcal{E}_k|\mathcal{E}_{k-1}, \theta_{k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} p(\mathcal{E}_{1:k-1}|\mathbf{z}_{1:k-1}). \tag{4}$$

Here, Equation 3 follows from Bayes rule and the replacement of $p(\mathcal{E}_{1:k} \mid \mathbf{z}_{1:k-1})$ by $p(\mathcal{E}_k \mid \mathcal{E}_{1:k-1}, \mathbf{z}_{1:k-1}) p(\mathcal{E}_{1:k-1} \mid \mathbf{z}_{1:k-1})$. Equation 4 follows from Equation 3 by the fact, that the Kalman filter state θ_{k-1} and the latest exemplar states $\mathcal{E}_{k-1}$ are sufficient statistics for the previous observations $\mathbf{z}_{1:k-1}$ and exemplar state sequence $\mathcal{E}_{1:k-1}$.

In most cases it is impossible to sample directly from Equation 4. The approach most commonly used in particle filters is to evaluate Equation 4 from right to left in a three stage process [24]: First, draw samples $s_k^{(\iota)}$ from the previous sample set using the importance weights, then draw for each such sample a new sample from the predictive distribution $p(\mathcal{E}_k \mid \mathcal{E}_{k-1}^{(\iota)}, \theta_{k-1}^{(\iota)})$, and finally weight these samples proportional to the observation likelihood $p(\mathbf{z}_k \mid \mathcal{E}_k^{(\iota)}, \theta_{k-1}^{(\iota)})$. The last step, importance sampling, adjusts for the fact that samples are not drawn from the actual target distribution. This approach has the disadvantage, that the most recent observations $\mathbf{z}_k$ are not taken into account during sampling, which can lead to sample depletion, if the predicted distribution is a poor approximation of the true posterior [25].

Fortunately, in our case, we are able to take the latest observation into account. From Equation 4 follows that the optimal sampling distribution is

$$\frac{p(\mathbf{z}_k \mid \mathcal{E}_k, \theta_{k-1}) p(\mathcal{E}_k \mid \mathcal{E}_{k-1}, \theta_{k-1})}{p(\mathbf{z}_k \mid \mathbf{z}_{1:k-1})}. \tag{5}$$

and the importance weights are then computed according to

$$\sum_{\mathcal{E}_k} p(\mathbf{z}_k \mid \mathcal{E}_k, \theta_{k-1}) p(\mathcal{E}_k \mid \mathcal{E}_{k-1}, \theta_{k-1}), \tag{6}$$

where the summation goes over all possible joint exemplar states. If the number of joint exemplars is relatively small, we can compute this distributions analytically. Additionally, this sampling distribution allows for *look-ahead* weighting [26]. Since Equation 6 does not depend on the sample's current state, all new samples generated from the same sample $s_{k-1}^{(\iota)}$ obtain the same importance weights. For this reason, the importance weight can be incorporated already into the importance weight $w_{k-1}^{(\iota)}$ of the parent sample. This way, more samples are drawn from "fitter" parent samples in the next time step, which additionally mitigates the sample depletion problem.

D. Computing the Observation Likelihoods

It remains to be shown, how the actual observation likelihoods for the most recent camera image z_k^c and laser scan z_k^l can be determined. According to Equation 4 the observation likelihood is

$$p(\mathbf{z}_k \mid \mathcal{E}_k, \theta_{k-1}) = p(z_k^c, z_k^l \mid \mathcal{E}_k^c, \mathcal{E}_k^l, \theta_{k-1}) \tag{7}$$

$$= p(z_k^c \mid z_k^l, \mathcal{E}_k^c, \mathcal{E}_k^l, \theta_{k-1}) p(z_k^l \mid \mathcal{E}_k^l, \theta_{k-1}) \tag{8}$$

Here, Equation 8 follows from Equation 7 by applying the chain rule and the second factor of Equation 8 follows from the fact that z_k^l is conditionally independent from $\mathcal{E}_k^c$ according to our model,see Figure 2.

Now, the individual observation likelihood for z_k^l can in principle be computed by marginalizing over the laser position predicted by the Kalman filter at time k, $\theta_{k|k-1}$, i.e.

$$p(z_k^l \mid \mathcal{E}_k^l, \theta_{k-1}) = \int p(z_k^l \mid \mathcal{E}_k^l, \mathbf{x}_k^l = x) p(\mathbf{x}_k^l = x \mid \theta_{k|k-1})\, dx. \tag{9}$$

However, computing this integral is intractable in practice, since it would require to match the exemplar ξ_k^l associated to $\mathcal{E}_k^l$ at all locations. Instead, we use an approximative approach, which is also quite popular in scan-based mapping, and determine the best position $\hat{z}_k^l(\xi_k^l)$, for the exemplar ξ_k^l by matching the exemplar against the current scan starting from the predicted position. We assume a fixed covariance for the matching position, and approximate $p(z_k^l \mid \mathcal{E}_k^l, \theta_{k-1})$ as

$$p(z_k^l \mid \mathcal{E}_k^l, \theta_{k-1}) \approx p(z_k^l \mid \xi_k^l, \hat{z}_k^l(\xi_k^l)) p(\hat{z}_k^l(\xi_k^l) \mid \theta_{k|k-1}). \tag{10}$$

Note, that in the Kalman filter framework $p(\hat{z}_k^l(\xi_k^l) \mid \theta_{k|k-1})$ amounts to compute the likelihood of the innovation, while $p(z_k^l \mid \xi_k^l, \hat{z}_k(\xi_k^l))$ can be determined from the mixture distribution of the exemplar model.

The observation likelihood of the latest image z_k^c can be determined almost in the same way. The main difference is, that it is additionally conditioned on the laser scan z_k^l and exemplar state $\mathcal{E}_k^l$ at time k. For this reason, we have to marginalize over the updated Kalman filter state at time k, θ_k. Again, we approximate the likelihood by first computing the best matching position of the exemplar ξ_k^c in the image. This is carried out by first computing the expected pixel coordinates of the exemplar based on the estimated laser position, i.e. the Kalman filter mean μ_k. Then the best matching position in the image $\hat{z}_k^c(\xi_k^c)$ is computed by direct ascent. Finally, the likelihood is approximated as

$$p(z_k^c \mid \mathcal{E}_k^c, \theta_k) \approx p(z_k^c \mid \xi_k^c, \hat{z}_k^c(\xi_k^c)) \int p(\hat{z}_k^c(\xi_k^c) \mid x_k^c) p(x_k^c \mid \theta_k)\, dx_k^c. \tag{11}$$

1. **Inputs:** $S_{k-1} = \{\langle s_{k-1}^{(\iota)}, w_{k-1}^{(\iota)} \rangle \mid \iota = 1, \ldots, N\}$, laser scan z_k^l and image z_k^c
2. $S_k := \emptyset$ *// Initialize*
3. **for** $\iota := 1, \ldots, N$ **do** *// Compute Kalman prediction for each sample*
4. $\quad \theta_{k|k-1}^{(\iota)} = \langle \mu_{k|k-1}^{(\iota)}, \Sigma_{k|k-1}^{(\iota)} \rangle$
5. **for** $\iota := 1, \ldots, N$ **do** *// Update importance weights by matching possible joint exemplars to the next observation*
6. $\quad \widehat{w}_{k-1}^{(\iota)} \propto w_{k-1}^{(\iota)} \sum_{\mathcal{E}_k} p(\mathbf{z}_k \mid \mathcal{E}_k, \theta_{k-1}^{(\iota)}) p(\mathcal{E}_k \mid \mathcal{E}_{k-1}^{(\iota)} \theta_{k-1}^{(\iota)})$
7. **for** $\iota := 1, \ldots, N$ **do** *// Sample $s_{k-1}^{(\iota)}$ using updated weights and draw $s_k^{(\iota)}$ from corresponding set*
8. $\quad$ Sample $s_{k-1}^{(\iota)} = \langle \theta_{k-1}^{(\iota)}, \mathcal{E}_{k-1}^{(\iota)} \rangle$ from S_{k-1} with probability proportional to the updated importance weights $\widehat{w}_{k-1}^{(\iota)}$.
9. $\quad$ Draw new exemplars $\mathcal{E}_k^{(\iota)}$ proportional to $p(\mathbf{z}_k \mid \mathcal{E}_k^{(\iota)}, \theta_{k-1}^{(\iota)}) p(\mathcal{E}_k^{(\iota)} \mid \mathcal{E}_{k-1}^{(\iota)}, \theta_{k-1}^{(\iota)})$
10. $\quad$ Update the position estimates $\theta_k^{(\iota)}$ using Kalman filter updates with $\hat{z}_k^{l(\iota)}$, and $\theta_{k|k-1}^{(\iota)}$
11. $\quad s_k^{(\iota)} := \langle \theta_k^{(\iota)}, \mathcal{E}_k^{(\iota)} \rangle; \qquad S_k := S_k \cup \{\langle s_k^{(\iota)}, \frac{1}{N} \rangle\}$
12. **return** S_k

Table 1: The Joint Exemplar RBPF algorithm.

Here, the integral is required to marginalize over the uncertainty of mapping laser positions to pixel coordinates. We assume that the random variable x_k^c is constant and Gaussian. The integral can then be solved analytically. The actual mapping and its covariance matrix is learned during training.

E. The Joint Exemplar RBPF Algorithm

The algorithm is summarized in Table 1. First, the Kalman filters are predicted in step 4. Most of the work is then carried out in step 6. Here, sampling distributions for sampling exemplars are computed for each sample, according to Equation 5. This involves matching all the exemplars against the current laser scan and image. The normalizers of the sampling distributions are used as the lookahead weights of the previous samples. In step 8, samples are drawn from the previous sample set S_{k-1}. For each of these samples, a new pair of exemplar states $\mathcal{E}_k$ is drawn proportional to the corresponding sampling distribution computed in step 6. In step 10, the Kalman filter update is applied to each new sample making use of the best matching position $\hat{z}_k^l(\xi_k^{(\iota)})$ of the new exemplar in the current laser scan. Actually, this Kalman update has already been computed in step 6 prior to computing the observation likelihood of the image, and can be re-used here.

IV. Learning the Joint Exemplar Model

Before a person can be tracked with the joint exemplar approach, the model needs to be learned from training data. This involves learning the exemplar sets Ξ^l, Ξ^k, their associated distribution parameters, and the transition model $p(\mathcal{E}_k \mid \mathcal{E}_{k-1}, \theta_{k-1})$. Additionally, we learn the mapping of positions in laser scans to the pixel coordinates and its uncertainty $p(x_k^c \mid \theta_k)$.

A. Generating and aligning training data

The training data is obtained by recording a sequence of laser scans and images of a person walking in front of a stationary robot. In this situation we can obtain silhouettes of persons in the images using background subtraction. Similarly, we obtain laser features by considering only end points of beams, which do not hit static obstacles. Figure 4 presents characteristic examples of the silhouettes and laser features extracted this way.

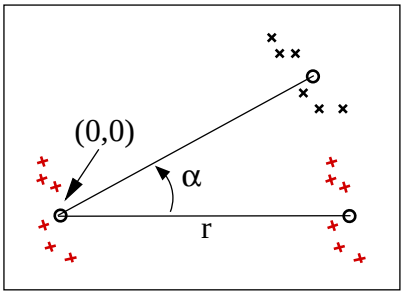

Fig. 3. The normalization of laser features: The laser scanner is in the (0,0) position. If we observe a laser feature in direction α at distance r, we first rotate the feature points by $-\alpha$ and then shift the result by $-r$.

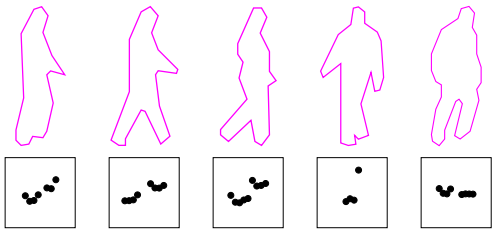

Fig. 4. Examples of silhouettes and laser features recorded during training. The contours are here scaled to the same height. The laser feature shown in the second row correspond to the respective silhouettes in the first row, but are normalized as shown in Figure 3

For computing the exemplar sets the pairwise similarity of the extracted training examples needs to be determined. For this purpose, the silhouettes and the laser features need to be aligned. To align two silhouettes we minimize the chamfer distance of the first silhouette based on the distance transformed image of the second silhouette using hill-climbing. The chamfer distance is the sum of the minimum squared distances of each edge pixels of the first silhouette to an edge

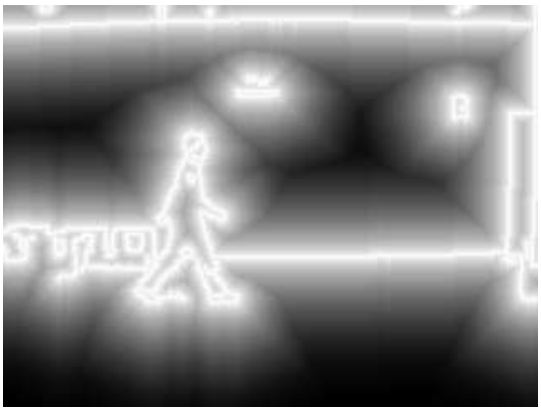

Fig. 5. A distance transformed edge image used for chamfer matching during contour tracking. To obtain this image, a canny edge detector is applied to the original image and a distance transform is applied.

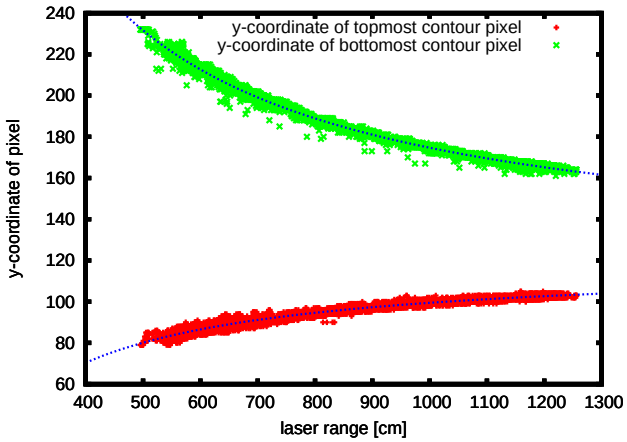

Fig. 6. y-coordinates of the topmost and bottommost point of the person in the training image as a function of the distance in the laser scan at the same point in time. The training data allows to estimate the projection function quite accurately.

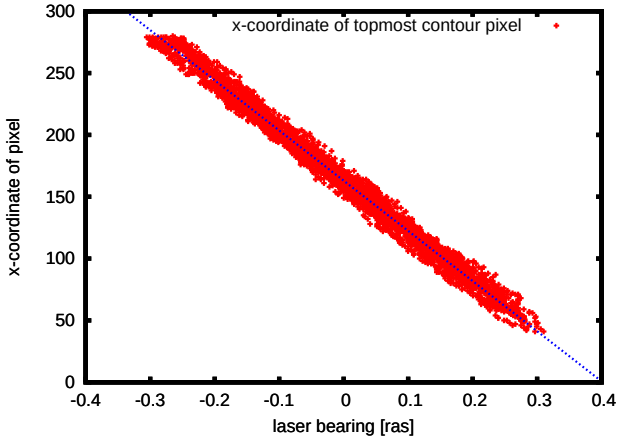

Fig. 7. x-coordinates of the topmost point of the person in the training image as a function of the direction in the laser scan at the same point in time. In the relevant range, the function is linear.

pixel of the second silhouette. The distance transform basically computes for each pixel of a binary image the Euclidean distance to the closest edge pixel, which allows to compute the chamfer distance fast. Figure 5 gives an example of a distance transformed image, where distance is encoded as gray-scale.

For aligning laser features, we make use of the fact that in a given stance the appearance of a person's legs in a laser scan only depend on the walking direction relative to the sensing direction. This is illustrated in Figure3. Here, the coordinate system is robot centered. We then account for the invariance by normalizing bearing and range to zero. This normalization process is illustrated in Figure 3. After normalization, the chamfer distance between two laser features is minimized based by computing the optimal displacement of one feature with respect to the other feature using an iterative least squares approach similar to ICP [27]. The only difference is that the orientation of the feature remains fixed. The result of the aligning process are two pairwise dissimilarity matrices, one for the silhouettes and one for the laser features.

B. Computing exemplar sets

Based on the dissimilarity matrices, exemplar sets can be computed using some pairwise clustering approach [28], [29]. We cluster the training examples into a pre-specified number of clusters. To determine the exemplar representing each cluster, we then follow the approach in [20] and choose the training example, whose maximum dissimilarity to any other example within the cluster is minimal. Finally, the parameters of the mixture components are estimated also using the approach proposed in [20], i.e. by assuming that the dissimilarities of each example in the cluster to its cluster center are drawn from a scaled χ^2-distribution.

C. Computing transition probabilities

Once the exemplar mixture distributions have been determined, the transition probabilities $p(\mathcal{E}_k \mid \mathcal{E}_{k-1})$ are learned. This is carried out using the EM algorithm, but keeping the mixture distribution fixed. Because the computation of the α and β values of the standard Baum-Welsh algorithm [30] becomes prohibitively expensive for larger state spaces, our current implementation adopts a technique known from factorial HMMs [31] and evaluates the E-step approximately based on Gibbs-sampling of exemplar state sequences $\mathcal{E}_{1:k}$. The maximization is then carried out using the transition frequencies counted during sampling. This has the additional advantage, that the transition probabilities can directly be learned in conditional form, i.e.

$$p(\mathcal{E}_k|\mathcal{E}_{k-1},\theta_{k-1}) = p(\mathcal{E}_k^c|\mathcal{E}_k^l,\mathcal{E}_{k-1}^c,\theta_{k-1})p(\mathcal{E}_k^l|\mathcal{E}_{k-1}^l,\theta_{k-1}),$$

which reduces the size of the model.

D. Computing the laser to image mapping

To compute the mapping of positions provided by the laser scanner to pixel coordinates in the image, we determine the topmost and bottommost image coordinates for the silhouettes in each training image. In Figure 6 the y-coordinates of these points in the image are plotted as a function of the distance to

Fig. 8. Outtake of the first experiment. The two persons are crossing their paths. The time gap between two images is one second. The matched contour based on the particle with the maximum a posterior weight are overlayed in pink and purple.The robot was moving at a speed of 60 cm/s during this experiment

the person measured by the laser scanner. We obtain a nonlinear relation, which results from the projection of the person to the image plane. Based on these training examples we compute least squares estimates of the projection functions. The relation between the x-coordinate in the image and the direction of the laser measurement is estimated independently, see Figure 7; in the relevant range for our tracking experiments, between 4 and 12 meters in front of the robot, this relation is almost linear. We then assume that the uncertainties in x and y directions are independent and follow Gaussian distributions.

V. Experimental Results

We evaluated our approach using a RWI B21 robot equipped with a SICK laser range scanner and a progressive scan VGA camera. The laser scanner is mounted at a height of approximately 40cm, such that it measures the distance to the legs of persons, while the camera is mounted at a height of 1.5m. The camera captures 30 frames per second with an image resolution of 320x240 pixel, while the laser scanner produces 37.5 scans per second and measures with an angular resolution of $0.5°$.

In order to test the performance of the approach on a moving mobile robot, we first learned exemplar models for two different persons. For this purpose, two sequences of 4000 camera images and laser scans were recorded for each person, where the persons had to perform straight line walks with different headings in front of the stationary robot. From the extracted sets of contours and laser features we learned exemplar models using 20 laser exemplars and 80 contour exemplars for both models.

To evaluate the performance of these models, the same two persons then walked in front of the moving robot for more then a minute, sometimes moving side by side, sometimes in opposite directions and also crossing their paths some time. The robot moved at a speed of 60 cm/s. The exemplar approach was then evaluated off-line based on the data recorded during this experiment.

In the first experiment, we tracked the two persons using independent exemplar trackers for both persons. Figure 8 shows example images from this experiment, where the pink and purple colored outlines indicate the contour exemplar of the sample with the maximum weight at the corresponding point in time. The two particle filters were able to track the persons reliably over the whole sequence of 2400 frames. Each particle filter used 100 particles during this experiment. The current C++ implementation achieves a frame rate of 17 frames per seconds for tracking a single person, when using this setting.

In a second experiment we tested, if the approach is capable of distinguishing between the two persons based on their exemplar models. For this purpose, we repeated the previous experiment, but this time using four independent particle filters, where each person is tracked with two filters using both learned exemplar models.

For each of the four particle filters, we recorded the sum of the sample weights after each time step, i.e. the sample normalizer, which corresponds to the likelihood $p(\mathbf{z}_k|\mathbf{z}_{1:k-1})$, which describes how well the filter explains the data observed so far. We then evaluated how well the approach is able to determine the correct assignment of filters to persons. Let us denote the four different likelihoods as $p_{aa}(\mathbf{z}_k|\mathbf{z}_{1:k-1})$,$p_{ab}(\mathbf{z}_k|\mathbf{z}_{1:k-1})$, $p_{ba}(\mathbf{z}_k|\mathbf{z}_{1:k-1})$,$p_{bb}(\mathbf{z}_k|\mathbf{z}_{1:k-1})$, where aa stands for the particle filter tracking person a using the exemplar model learned from data of person a and ab stands for the filter tracking person a using the model learned from data of person b. Based on these values, we compute after each image the posterior probability of the two joint assignments of filters to persons. These assignments are H_1, both persons are correctly identified, and H_0, the ids are erroneously swapped. We compute the posterior $p(H_1(k))$ starting with an initial prior $p(H_0(0)) = p(H_1(0)) = 0.5$,

$$\begin{aligned} p(H_0(k)) &= \alpha_k * p_{ab}(\mathbf{z}_k|\mathbf{z}_{1:k-1}) * p_{ba}(\mathbf{z}_k|\mathbf{z}_{1:k-1}) * p(H_0(k-1)) \\ p(H_1(k)) &= \alpha_k * p_{aa}(\mathbf{z}_k|\mathbf{z}_{1:k-1}) * p_{bb}(\mathbf{z}_k|\mathbf{z}_{1:k-1}) * p(H_1(k-1)), \end{aligned}$$

where the α_k are normalizers that ensures that the two probabilities sum up to one.

Figure 9 shows the result of this process obtained on 5 different parts of the data sequence. In all 5 cases our approach identified the correct hypothesis within 11 seconds, which indicates that the method is able to distinguish between known persons during tracking.

VI. Conclusion

In this article we introduced a joint exemplar Rao-Blackwellized particle filter for tracking a person in laser range data and camera images. The Rao-Blackwellized particle filter tracks the appearance of a person by sampling characteristic contours and laser features and it tracks the person's motion conditioned on the sampled appearance using a Kalman filter for each particle. The approach uses metric mixture models of the contours of a person in video images and of features of a person's legs in laser data to compute the observation

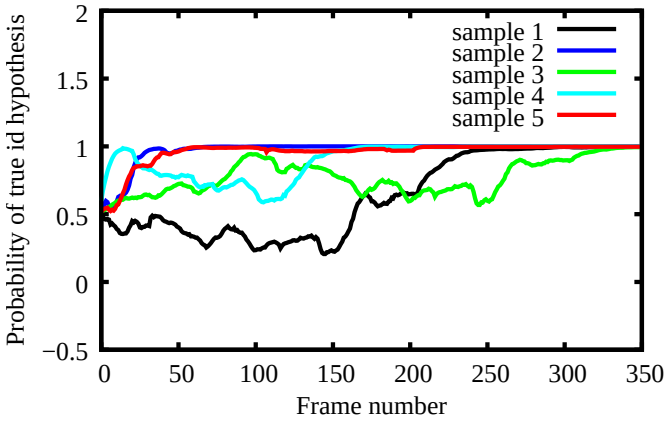

Fig. 9. Learning the correct id assignment during the tracking of two persons. Within 11 seconds the approach was able to figure out the correct id assignment to the persons being tracked in Figure 8 based in their different exemplar models; see text for details.

likelihood of different prototypical shapes in the data. These exemplar models can be learned from training sequences using clustering for obtaining the mixture models, and EM for obtaining the exemplar transition probabilities. Finally, we presented tracking results achieved with the approach, which show that it is able to reliably track two persons with a mobile robot that moves at a speed of 60 cm/s. The technique is also able to simultaneously figure out the identity of the persons during tracking based on their exemplar models.

However, there is still room for future work. Currently, the motion model of the person is based on a single linear Kalman filter only. In the future we want to extend the RBPF to take possible maneuvers into account by using a model switching approach, where parameters of the Kalman filter and the switching probabilities could also be learned from training data. We also want to additionally take color histograms into account and adapt the appearance models during tracking.

REFERENCES

[1] K. Arras and S. Vestli, "Hybrid, high-precision localization for the mail distributing mobile robot system MOPS," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 1998.

[2] I. Horswill, "Polly: A vision-based artificial agent," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 1993.

[3] R. Simmons, R. Goodwin, K. Haigh, S. Koenig, and J. O'Sullivan, "A layered architecture for office delivery robots," in *Proc. of the First International Conference on Autonomous Agents*, Marina del Rey, CA, 1997.

[4] H. Endres, W. Feiten, and G. Lawitzky, "Field test of a navigation system: Autonomous cleaning in supermarkets," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 1998.

[5] J. F. Engelberger, "Health-care robotics goes commercial: The 'helpmate' experience," *Robotica*, vol. 11, pp. 517–523, 1993.

[6] W. Burgard, A. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "Experiences with an interactive museum tour-guide robot," *Artificial Intelligence*, vol. 114, no. 1-2, 1999.

[7] S. Thrun, M. Bennewitz, W. Burgard, A. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, "MINERVA: A second generation mobile tour-guide robot," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 1999.

[8] S. Tadokoro, M. Hayashi, Y. Manabe, Y. Nakami, and T. Takamori, "On motion planning of mobile robots which coexist and cooperate with human," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1995.

[9] D. Hähnel, D. Schulz, and W. Burgard, "Map building with mobile robots in populated environments," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2002.

[10] A. Fod, A. Howard, and M. Mataric, "Laser-based people tracking," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2002.

[11] M. Bennewitz, W. Burgard, and S. Thrun, "Using EM to learn motion behaviors of persons with mobile ro bots," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2002.

[12] D. Schulz, W. Burgard, and D. Fox, "People tracking with a mobile robot using joint probabilistic data association filters," *International Journal of Robotics Research (IJRR)*, 2003.

[13] B. Kluge, C. Koehler, and E. Prassler, "Fast and robust tracking of multiple moving objects with a laser range finder," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2001.

[14] M. Montemerlo, S. Thun, and W. Whittaker, "Conditional particle filters for simultaneous mobile robot localization and people-tracking," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2002.

[15] A. Fod, A. Howard, and M. J. Matarić, "Laser-based people tracking," in *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2002.

[16] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*, ser. Mathematics in Science and Engineering. Academic Press, 1988.

[17] G. Taylor and L. Kleeman, "A multiple hypothesis walking person tracker with switched dynamic model," in *Proc. of the Australasian Conference on Robotics and Automation*, 2004.

[18] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun, "Learning motion patterns of people for compliant robot motion," vol. 24, no. 1, pp. 31–48, 2005.

[19] A. Brooks and S. Williams, "Tracking people with networks of heterogeneous sensors," in *Proc. of the Australian Conference on Robotics and Automation*, 2005.

[20] K. Toyama and A. Blake, "Probabilistic tracking with exemplars in a metric space," *International Journal of Computer Vision (IJCV)*, vol. 48, no. 1, pp. 9–19, 2002.

[21] B. Frey and N. Jojic, "Learning graphical models of images, videos and their spatial transformations," in *Proc. of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

[22] K. Choo and D. J. Fleet, "People tracking using hybrid monte-carlo filtering," in *Proc. of the International Conference on Computer Vision (ICCV)*, 2001, p. 321.

[23] L. Sigal, S. Bhatia, R. Stefan, B. M. J., and M. Isard, "Tracking loose-limbed people," in *Proc. of Computer Vision and Pattern Recognition*, vol. I, 2004, pp. 421–428.

[24] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo in Practice*. New York: Springer-Verlag, 2001.

[25] M. Pitt and N. Shephard, "Filtering via simulation: auxiliary particle filters," *Journal of the American Statistical Association*, vol. 94, no. 446, 1999.

[26] R. Morales-Menéndez, N. de Freitas, and D. Poole, "Real-time monitoring of complex industrial processes with particle filters," in *Advances in Neural Information Processing Systems 15*, 2002.

[27] F. Lu and E. Milios, "Robot pose estimation in unknown environments by matching 2d range scans," in *IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, 1994.

[28] V. Roth, J. laub, M. Kawanabe, and J. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," *IEEE Transactions on Patern Analysis and Machine Intelligence (PAMI)*, vol. 25, no. 12, pp. 1540–1550, December 2003.

[29] T. Hofmann and J. Buhmann, "Pairwise data clustering by deterministic annealing," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 19, no. 1, pp. 1–14, January 1997.

[30] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*. IEEE, 1989.

[31] Z. Ghahramani, , and M. I. Jordan, "Factorial hidden Markov models," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., vol. 8. MIT Press, 1995, pp. 472–478.