Auton Robot (2009) 26: 141–151
DOI 10.1007/s10514-009-9112-4

Classifying dynamic objects

An unsupervised learning approach

Matthias Luber · Kai O. Arras · Christian Plagemann · Wolfram Burgard

Received: 7 November 2008 / Accepted: 6 March 2009 / Published online: 27 March 2009

Abstract For robots operating in real-world environments, the ability to deal with dynamic entities such as humans, animals, vehicles, or other robots is of fundamental importance. The variability of dynamic objects, however, is large in general, which makes it hard to manually design suitable models for their appearance and dynamics. In this paper, we present an unsupervised learning approach to this model-building problem. We describe an exemplar-based model for representing the time-varying appearance of objects in planar laser scans as well as a clustering procedure that builds a set of object classes from given observation sequences. Extensive experiments in real environments demonstrate that our system is able to autonomously learn useful models for, e.g., pedestrians, skaters, or cyclists without being provided with external class information.

Keywords Classification · Unsupervised learning · Detection and tracking

1 Introduction

The problem of tracking dynamic objects and modeling their time-varying appearance has been studied extensively in robotics, engineering, computer vision, and other areas. On one hand, the problem is hard as the appearance of objects is ambiguous, partly occluded, may vary quickly over time, and is perceived via a high-dimensional measurement space. On the other hand, the problem is highly relevant in practice—especially in future applications for mobile robots and intelligent cars. Consider, for example, a service robot deployed in a populated environment such as a pedestrian precinct. Tasks like collision-free navigation or interaction require the ability to recognize, distinguish, and track moving objects including reliable estimates of object classes, e.g., "adult", "infant", "car", "dog", etc.

In this paper, we consider the problem of detecting, tracking, and classifying moving objects in sequences of planar range scans acquired by a laser sensor. We present an exemplar-based model for representing the time-varying appearance of moving objects as well as a clustering procedure that builds a set of object classes from given observation sequences in conjunction with a Bayes filtering scheme for classification. The proposed system, which has been implemented and tested on a real robot, does not require labeled object trajectories, but rather uses an unsupervised clustering scheme to automatically build appropriate class assignments. By pre-processing the sensor stream using state-of-the-art feature detection and tracking algorithms, we obtain a system that is able to learn and re-use object models on-the-fly and without human intervention. The resulting set of object models can then be used to (1) recognize previously seen object classes and (2) improve data segmentation and association in ambiguous multi-target tracking situations. We furthermore believe that the object models may be used in various applications to associate semantics with recognized objects depending on their classes.

M. Luber (✉) · K.O. Arras · C. Plagemann · W. Burgard
Department for Computer Science, Albert-Ludwigs-University Freiburg, 79110 Freiburg, Germany
e-mail: luber@informatik.uni-freiburg.de

K.O. Arras
e-mail: arras@informatik.uni-freiburg.de

C. Plagemann
e-mail: plagem@informatik.uni-freiburg.de

W. Burgard
e-mail: burgard@informatik.uni-freiburg.de

Fig. 1 Six examples of relevant object classes considered in this paper. Our proposed system learns probabilistic models of their appearance in planar range scans and the corresponding dynamics. The classes are denominated Pedestrian (PED), Buggy (BUG), Skater (SKA), Suitcase (SUI), Cyclist (CYC), and Kangaroo-shoes (KAN)

2 Related work

Exemplar-based models are frequently applied in computer vision systems for dealing with the high dimensionality of visual input. Toyama and Blake (2002), for instance, used probabilistic exemplar models for representing and tracking human motion. Their approach is similar to ours in that they also learn probabilistic transition models. As the major differences, the range-bearing observations used in this work are substantially more sparse than visual input and we also address the problem of learning different object classes in an unsupervised way. Plagemann et al. (2005) used exemplars to represent the visual appearance of 3D objects in the context of an object localization framework. Kruger et al. (2006) learned exemplar models to realize a face recognition system for video streams. Exemplar-based approaches have also been used in other areas such as action recognition (Drumwright et al. 2004) or word sense disambiguation (Ng and Lee 1996). Wren et al. (1997) introduce a people modeling and tracking system for color images. It uses a multi-class model of shape and color and has an explicit background model to perform image segmentation.

There exists a large body of work on laser-based object and people tracking in the robotics literature (Schulz et al. 2001; Fod et al. 2001, 2002; Montemerlo and Thrun 2002; Arras et al. 2007). People tracking typically requires carefully engineered or learned features for track identification and data association and often a-priori information about motion models. This has been shown to be the case also for geometrically simpler and rigid object such as vehicles in traffic scenarios (MacLachlan and Mertz 2006). Cui et al. (2006) describe a system for tracking single persons within a larger set of people, given the relevant motion models are known.

The work most closely related to ours has recently been presented by Schulz (2006), who combined vision- and laser-based exemplar models to realize a people tracking system. In contrast to his work, our main contribution is the unsupervised learning of multiple object classes that can be used for tracking as well as for classifying dynamic objects. Ilg et al. (2003) also follow a prototype-based approach. In contrast to this work, they explicitly align time series using Dynamic Time Warping to perform a clustering into prototypes.

Periodicity and self-similarity have been studied by Cutler and Davis (2000), who developed a classification system based on the autocorrelation of appearances, which is able to distinguish, for example, walking humans from dogs.

A central component of our approach detailed in the following section is an unsupervised clustering algorithm to produce a suitable set of exemplars. Most approaches to cluster analysis (Hartigan 1975) assume that all data is available from the beginning and that the number of clusters is given. Recent work in this area also deals with sequential data and incremental model updates (Tasoulis et al. 2006; Chis and Grosan 2006). Ghahramani (2004) gives an easily accessible overview of the state-of-the-art in unsupervised learning.

As an alternative to the exemplar-based approach, researchers have applied generic dimensionality reduction techniques to deal with high-dimensional and/or dynamic appearance distributions. PCA and ICA have, for example, been used to recognize people from iris images (Wang and Han 2005) or their faces (Fortuna and Capson 2004). Recent advances in this area include latent variable models, such as Gaussian process latent variable models (GPLVMs) (Lawrence 2005).

The approach of Wang et al. (2006), termed Gaussian process dynamical models (GPDMs), builds on the idea that the high-dimensional data which is observed over time actually lies on a low-dimensional manifold. They build on GPLVMs to learn and represent the low-dimensional embedding in a nonparametric way. The feasibility of this approach has been shown for the different problem of body pose tracking from visual input.

Jenkins and Matarić (2004) extended Isomap (Tenenbaum et al. 2000), which is another popular method for nonlinear dimensionality reduction, by a spatio-temporal component which allows to model high-dimensional data that changes over time. One of their example instantiations of the model shows that it develops into a HMM-like structure for clustered data.

3 Modeling object appearance and dynamics using exemplars

Exemplar models are representations for both, appearance and appearance dynamics. They are a choice consistent with the motivation for an unsupervised learning approach avoiding manual feature selection, parameterized physical models (e.g., human gait models), and hand-tuned classifier creation.

This section describes how the exemplar-based models of dynamic objects are learned. Based on a segmentation and tracking system presented in Sect. 6, we assume to have a discrete track for each dynamic object in the current scene. Over time, these tracks describe trajectories that we analyze regarding appearance and dynamics of the corresponding objects.

3.1 Problem description

The problem we address in this work can be formalized as follows. Let $T = \langle Z_1, \ldots, Z_m \rangle$ be a *track*, i.e., a time-indexed observation sequence of appearances Z_t, $t = 1, \ldots, m$, of an object belonging to an *object class* C. Then we face the following two problems:

1. *Unsupervised learning*: Given a set of observed tracks $\mathcal{T} = \{T_1, T_2, \ldots\}$, learn classes $\{C_1, \ldots, C_n\}$ of objects in an unsupervised manner. This amounts to setting an appropriate number n of classes and to learn for each class C_j a probabilistic model $p(T \mid C_j)$ that characterizes the time-varying appearance of tracks T associated with that class.
2. *Classification*: Given a newly observed track T and a set of known object classes $\mathcal{C} = \{C_1, \ldots, C_n\}$, estimate the class probabilities $p(C_j \mid T)$ for all classes.

Note that "unsupervised" in this context does not mean that *all* model parameters are learned from scratch, but rather that the important class information (e.g. "pedestrian", "cyclist") is not supplied to the system. The underlying segmentation, tracking, and feature extraction subsystems are designed to capture a wide variety of possible object appearances and the unsupervised learning task is to build a compact representation of object appearance that generalizes across instances.

3.2 The exemplar model

Exemplar models (Toyama and Blake 2002) aim at approximating the typically high-dimensional and dynamic appearance distribution of objects using a sparse set $\mathcal{E} = \{E_1, \ldots, E_r\}$ of significant observations, termed *exemplars* E_i. Similarities between concrete observations and exemplars as well as between two exemplars are specified by a distance function $\rho(E_i, E_j)$ in exemplar space. Furthermore, each exemplar is given a prior probability $\pi_i = p(E_i)$, which reflects the prior probability of a new observation being associated with this exemplar. Changes in appearance over time are dealt with by introducing transition probabilities $p(E_i \mid E_j)$ between exemplars w.r.t. a predefined iteration frequency. Formally, this renders the exemplar model a first-order Markov chain, specified by the four elements $\mathcal{M} = (\mathcal{E}, B, \pi, \rho)$, which are the exemplar set $\mathcal{E}$, the transition probability matrix B with elements $b_{i,j} = p(E_i \mid E_j)$, the priors π, and the distance function ρ. All these components can be learned from data, which is one of the central topics of this paper.

3.3 Exemplars for range-bearing observations

In a laser-based object tracking scenario, the raw laser measurements associated with each track constitute the appearance $Z = \{(\alpha_i, r_i)\}_{i=1}^{l}$ of the objects, where α_i is the bearing, r_i is the range measurement, and l is the number of laser end points in the respective laser segment.

To cluster the laser segments into exemplars, the individual laser segments need to be normalized with respect to rotation and translation. This is achieved using the state information estimated by the underlying tracker. Here, the state of a track $\mathbf{x} = (x, y, v_x, v_y)^T$ is composed of the target position (x, y) and velocities (v_x, v_y). The velocity vector can then be used to calculate the heading of the object. Translational invariance is achieved by shifting the center of gravity of the segment to $(0, 0)$, rotational invariance is gained from zeroing the orientation in the same way. After normalization, all segments have a fixed position and orientation.

Rather than using the raw laser end points of the normalized segments as observations (see Schulz 2006), we calculate the so called *likelihood field* (Thrun 2001) on a regular grid for each of them. In this model, the likelihood of a range measurement is a function of the Euclidean distance d_{euc} of the respective endpoint of the beam to the closest obstacle in the environment. The likelihood of each cell (x, y) is then calculated using a Gaussian distribution $\mathcal{N}(d_{\text{euc}}(x, y); 0, \sigma^2)$ with zero mean and standard deviation σ which reflects the sensor noise. In the past, likelihood fields have been used successfully for tasks like localization or scan matching. The main advantage of this approach is that the distance function for observations can be defined independently of the number of laser end points in the segment and that likelihood estimation for new observations can be performed efficiently. We will henceforth denote the grid representation of an appearance Z_i as G_i. Figure 2 shows an example of a track, a laser segment, the normalized segment, and the corresponding grid for a walking pedestrian.

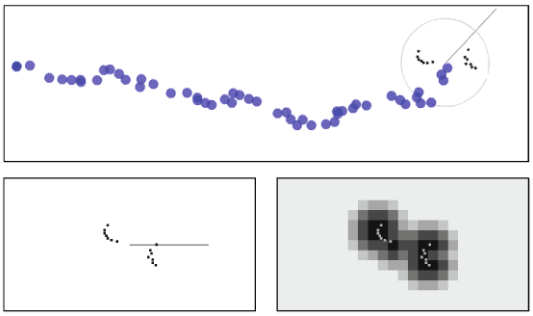


Fig. 2 Pre-processing steps illustrated with a pedestrian observed via a laser range finder. The *top* figure shows the trajectory of the subject moving from *left* to *right*. The segmentation and tracking system yields estimates of target location (shown as a trace of *large dots*), orientation (shown by a *line*), and velocity. The *bottom left* figure shows the raw range readings (*small dots*) that are normalized such that the estimated motion direction is zeroed. The resulting grid-based representation G generated from the set of normalized laser end points is depicted in the *bottom right* figure

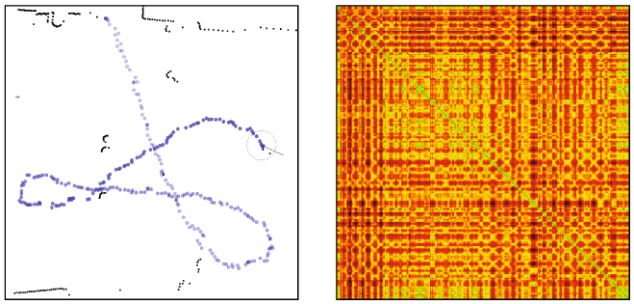


Fig. 3 Trajectory (*left*) and self-similarity matrix (*right*) of a pedestrian walking in a large hallway. The track consists of 387 observations. The walking cycle of the pedestrian causes a periodic structure of the self-similarity matrix. The relative stability of this structure demonstrates how the exemplar representation is able to uncover salient appearance properties with invariance to the subject's heading and to self-occlusion by the legs

3.4 Validation of the exemplar approach

The exemplar representation has a strong impact on both the creation of the exemplar set from a sequence of appearances and the unsupervised creation of new object classes. This motivates a careful analysis of the choices made. To illustrate the practicability of the exemplar model for our purpose, we analyzed the self-similarity of exemplars for tracks of objects from relevant object classes. We define the similarity $S(G_{t_1}, G_{t_2})$ of two observations obtained at times t_1 and t_2 as the absolute correlation

$$S(G_{t_1}, G_{t_2}) := \sum_{(x,y)\in\mathcal{B}} |G_{t_1}(x,y) - G_{t_2}(x,y)|, \tag{1}$$

where $\mathcal{B}$ is the bounding box of the grid-based representations of the observations Z_{t_1} and Z_{t_2}.

Figure 3 visualizes the self-similarity matrix for 387 observations of a pedestrian. Both axes of this matrix (Fig. 3, right) correspond to the time with t_1 along the horizontal and t_2 along the vertical axis. The gray values that encode self-similarity range from bright to dark. Whereas light gray stands for maximal correlation, black represents minimal self-similarity. The diagonal is maximal by definition as the distance of an observation to itself is zero.

We recognize the periodic structure of the matrix, which is caused by the strong self-similarity of the appearance of the pedestrian over a walking cycle. This is not self-evident as the appearance of the walking person in laser data changes with the heading of the person relative to the sensor. Poor normalization (e.g., because of inaccurate heading estimates of the underlying tracker) or a poor exemplar representation (e.g., which is too sensitive to measurement noise) would have removed the periodicity in the data. This illustrates that the normalization and the grid-based representation of appearance has sufficiently good invariance properties, so that a small amount of salient appearance patterns, i.e. *exemplars*, and the transitions between them are well suited for our goal to learn and classify dynamic objects.

3.5 Learning the exemplar model

In this section, we will describe how exemplar models are learned from observation sequences. This involves the distance function ρ, the exemplar set $\mathcal{E}$, the prior probabilities π_i, and the transition probabilities $p(E_i \mid E_j)$.

3.5.1 Distance function for exemplar learning

We assess the similarity of two observations Z_i and Z_j based on a distance function applied to the corresponding grid-based representations G_i and G_j. Interpreting the grids as histograms we employ the Euclidean distance for this purpose:

$$\rho(G_i, G_j) = \sqrt{\sum_{(x,y)} (G_i(x,y) - G_j(x,y))^2}. \tag{2}$$

The function is used for both, the clustering and the Gaussian observation model in (4) described hereafter.

3.5.2 Exemplar set

Exemplars are representations that generalize object appearance. To this aim, similar appearances are associated and merged into clusters. In our current system, we apply k-means clustering (Hartigan 1975) to partition the full data set into r clusters $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_r$ (see Fig. 4).

Strong outliers in the training set—which cannot be merged with other observations—are retained by the clustering process as additional, non-representative exemplars.

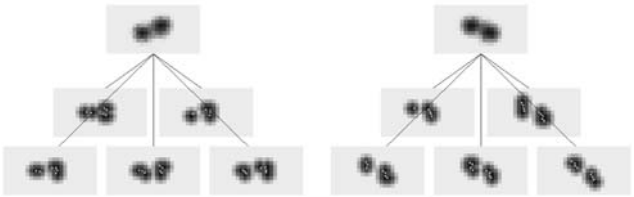


Fig. 4 Example clusters of a pedestrian. The diagram shows the centroids of two clusters (exemplars E) each created from a set of five observations G

Such observations may occur for several reasons, e.g., when a tracked object performs atypical movements, when the underlying segmentation method fails to produce a proper foreground segment, or due to sensor noise. To achieve robustness with respect to such outliers, we accept an exemplar only if it was created from a minimum number of observations. This assures that the resulting exemplars characterize only states of the appearance dynamics that occur often and are representative.

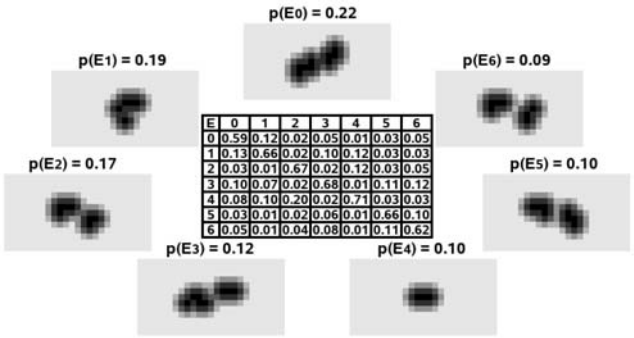
3.5.3 Transition probabilities

Once the clustered exemplar set has been generated from the observation sequence, the transition probabilities between exemplars can be learned. As defined in Sect. 3.2, we model the dynamics of the appearance of an object using hidden Markov models (HMM). The transition probabilities are obtained by pair-wise counting. A transition between two exemplars E_i and E_j is counted each time an observation that has minimal distance to E_i is followed by an observation with minimal distance to E_j. As there is a non-zero probability that some transitions are never observed although they exist, the transition probabilities are initialized with a small value to moderately smooth the resulting model.

Accordingly, the exemplar priors π_i are determined by counting the number of contributing observations G in a cluster. See Fig. 5 for the learned exemplar model of a pedestrian.

4 Classification

Having learned the exemplar set and the transition probabilities as described in the previous section, both can be used to classify tracks of different objects in a Bayesian filtering framework. More formally, given the grid representations $\langle G_1, \ldots, G_m \rangle$ of the observations of a track T and a set of learned classes $\mathcal{C} = \{C_1, \ldots, C_n\}$, we want to estimate the class probabilities $p_t(C_k \mid T)_{k=1}^n$ for every time step t. The estimates for the last time step m then reflect the consistency of the entire track with the different exemplar models. These quantities can thus be used to make classification decisions.



E	0	1	2	3	4	5	6
0	0.59	0.12	0.02	0.05	0.01	0.03	0.05
1	0.13	0.66	0.02	0.10	0.12	0.03	0.03
2	0.03	0.01	0.67	0.02	0.12	0.03	0.05
3	0.10	0.07	0.02	0.68	0.01	0.11	0.12
4	0.08	0.10	0.20	0.02	0.71	0.03	0.03
5	0.03	0.01	0.02	0.06	0.01	0.66	0.10
6	0.05	0.01	0.04	0.08	0.01	0.11	0.62

Fig. 5 Laser-based exemplar model of a pedestrian. The transition matrix is shown in the *center* with the exemplars sorted counterclockwise according to their prior probability

4.1 Estimating class probabilities over time

Each exemplar model $\mathcal{M}^i$ represents the distribution of track appearances for its corresponding object class C_i. Thus, a combination of *all* known exemplar models $\mathcal{M}^{\text{comb}} = \{\mathcal{M}^1, \ldots, \mathcal{M}^n\}$ covers the entire space of possible appearances—or, more precisely, of all appearances that the robot has seen so far. We construct the exemplar set $\mathcal{E}^{\text{comb}}$ of $\mathcal{M}^{\text{comb}}$ by simply building the union set of the individual exemplar sets $\mathcal{E}^k$ of all models $\mathcal{M}^k$. The transition probability matrix B^{comb} as well as the exemplar priors π^{comb} can be obtained from the B^k matrices and the π^k in a straightforward way since we assume that the corresponding exemplar sets do not intersect. This assumption means that objects do not change their class during the time of observation, that is to say, for example, that no skater takes off his shoes and becomes a pedestrian. Therefore all cross-model transition probabilities in B^{comb} are set to zero.

Given this combined exemplar model, a belief function Bel_t for the class probabilities $p_t(C_k \mid T)_{k=1}^n$ can be updated recursively over time using the well-known Bayes filtering scheme. For better readability, we introduce the notation E_i^k to refer to the ith exemplar of model $\mathcal{M}^k$. According to the Bayes filter, the belief about object classes is initialized as

$$Bel_0(E_i^k) = p(\mathcal{M}^k) \cdot \pi_i^k, \tag{3}$$

where π_i^k denotes the prior probability of E_i^k and $p(\mathcal{M}^k)$ stands for the model prior, which we assumed to be uniform (or can be estimated from a training set). Starting with G_1, we now perform the following recursive update of the belief function for every G_t:

$$Bel_t(E_i^k) = \eta_t \cdot p(G_t \mid E_i^k) \cdot \sum_l \sum_j p(E_i^k \mid E_j^l) \times Bel_{t-1}(E_j^l). \tag{4}$$

In this equation, η_t is a normalizing factor ensuring that $Bel_t(E_i^k)$ sums up to one over all i and k, and $p(G_t \mid E_i^k)$ is

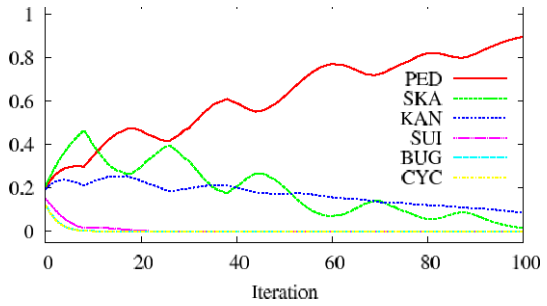


Fig. 6 The graphs show the evolution of the probabilities of different classes over time during an experiment in which a pedestrian is being observed. The x-axis refers to the time t. The classes that compete most are pedestrian (*solid line*) and skater (*dashed line*). The periodicity in the graphs corresponds to the walking or skating cycle respectively. Both cycles have very similar appearance in the laser data. Integrated over time, however, the pedestrian class obtains maximum posterior probability, which corresponds to the ground truth

the Gaussian observation likelihood using the distance function in (2).

The estimates of exemplar probabilities $Bel_t(E_i^k)$ at time t can be summed up to yield the individual class probabilities

$$p_t(\mathcal{M}^k \mid T) = \sum_i Bel_t(E_i^k) \ . \tag{5}$$

At time $t = m$, that is, when the entire observation sequence has been processed, the $p_m(\mathcal{M}^k \mid T)$ constitute the resulting estimates of the class probabilities of our model. In particular, we define

$$\mathcal{M}^{\text{best}}(T) := \text{argmax}_k \ p_m(\mathcal{M}^k \mid T) \tag{6}$$

as the most likely class assignment for track T. To visualize the filtering process described above, we give an example run for a pedestrian track T in Fig. 6 and plot the class probabilities for five alternative object classes over time.

5 Unsupervised learning of object classes

In this section, we explain how to learn a set of object classes from scratch in an unsupervised manner. Objects of a previously unknown type will always be assigned to some class by the Bayes filter. The class with the highest resulting probability estimate provides the current best, yet suboptimal description of the object at the time. A better fit would always be achieved by creating a new, specifically trained model for this particular object instance. Thus, we are faced with the classic model selection problem, that is, choosing between a more compact vs. a more precise model for explaining the observed data. As a selection criterion, we employ the *Bayes factor* (Kass and Raftery 1995) which considers the amount of evidence in favor of a model relative to an alternative one.

More formally, given a set of known classes $\mathcal{C} = \{C_1, \ldots, C_n\}$ and their respective models $\{\mathcal{M}^1, \ldots, \mathcal{M}^n\}$, let T be the track of an object to be classified. We determine the current best matching model $\mathcal{M}^{\text{best}}(T)$ according to (6) and learn a new, fitted model $\mathcal{M}^{\text{new}}(T)$ as described in Sect. 3.5. To decide whether T should be added to $\mathcal{M}^{\text{best}}(T)$ or rather to $\mathcal{M}^{\text{new}}(T)$ by adding a new object class C^{new} to the existing set of classes, we calculate the model probabilities $p(\mathcal{M}^{\text{best}}(T) \mid T)$ and $p(\mathcal{M}^{\text{new}}(T) \mid T)$ using the Bayes filter. The ratio of these probabilities yields the factor

$$K = \frac{p(\mathcal{M}^{\text{new}}(T) \mid T)}{p(\mathcal{M}^{\text{best}}(T) \mid T)} \ , \tag{7}$$

that quantifies how much better the new model describes this object instance relative to the current best matching model. While large values for a threshold on K favor more compact models (fewer classes and lower data-fit), lower values lead to more precise models (more classes, in the extreme case overfitting the data). As alternative model selection criteria, one could use the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC), for example. However, during our experimental evaluation, the Bayes factor yielded accurate results and, thus, we leave the comparison to alternative choices to future work.

We now describe how to identify a threshold on K, so that the system achieves a human-like class granularity, that is, a balance between model precision and compactness which is similar to how humans classify dynamic objects. To this aim, we collected a training set consisting of instances of the classes *pedestrian*, *skater*, *cyclist*, *buggy*, and *kangaroo*. We first compared the current best models and the fitted models of objects of the same class and calculated the factors K according to (7). Then we carried out the same comparison with objects of different classes with randomly selected tracks. Table 1 gives the relative number of pairs for which different values of K—ranging from 1 to 20—were exceeded. It can be seen that, e.g., for $K \geq 4$, all pedestrians are merged to the same class (PED/PED), but also that there is a poor separation (40%) between pedestrians and skaters (PED/SKA). Given this set of tested thresholds K, the best trade-off between precision and recall is achieved between $K \geq 2$ and $K \geq 4$. We therefore chose $K \geq 3$.

Interestingly, this threshold on K coincides with the interpretation of "substantial evidence against the alternative model" of Kass and Raftery (1995). Note that fitting the threshold K to a labeled data-set does not render our approach a supervised one, since no specific class labels—which is the crucial information in this task—are supplied to the system. This step can rather be compared to learning regularization parameters in alternative models to balance data-fit against model complexity.

Table 1 Percentages of incorrectly separated (*top five rows*) and correctly separated (*bottom five rows*) track pairs. A Bayes factor is sought that trades off separation of tracks from different classes and association of tracks from the same class

	$K \geq 1$	$K \geq 2$	$K \geq 4$	$K \geq 8$	$K \geq 20$
PED/PED	41%	2%	0%	0%	0%
SKA/SKA	58%	7%	0%	0%	0%
CYC/CYC	79%	32%	14%	10%	8%
BUG/BUG	78%	47%	21%	9%	1%
KAN/KAN	60%	40%	21%	11%	3%
PED/KAN	46%	3%	0%	0%	0%
PED/SKA	100%	83%	40%	10%	0%
CYC/BUG	100%	100%	100%	99%	50%
BUG/KAN	100%	100%	100%	100%	82%
CYC/KAN	100%	100%	100%	98%	92%

6 Segmentation and tracking

The segmentation and tracking system takes the raw laser scans as input and produces the tracks with associated laser segments for the exemplar generation step. To this end, we employ a Kalman filter-based multi-target tracker with a constant velocity motion model. We use this model since it makes mild assumptions about the motion of targets of unknown type. Practical experiments with a constant acceleration motion model have been made without sensible changes in performance.

The observation step in the filter amounts to the problem of partitioning the laser range image into segments that consist in measurements on the same dynamic objects and to estimate their center. This is done by subtracting successive laser scans to extract beams that belong to dynamic objects. If the beam-wise difference is above the sensor noise level, the measurement is marked and grouped into a segment with other moving points in a pre-defined radius.

We compared four different techniques to calculate the segment center: mean, median, average of extrema, and the center of a circle fitted through the segments points (for the latter the closed-form solutions from Arras et al. 2007 were taken). The last approach leads to very accurate results when tracking pedestrians, skaters, and people on kangaroo shoes but fails to produce good estimates with person pushing a buggy and cyclists. The mean turned out to be the smoothest estimator of the segment center.

Data association is realized with a modified nearest neighbor filter. It was adapted so as to associate multiple observations to a single track. This is necessary to correctly associate the two legs of pedestrians, skaters, and kangaroo shoes that appear as nearby blobs in the laser range image. Although more advanced data association strategies, motion models, or segmentation techniques have been described in the related literature, we found the system effective for our purposes.

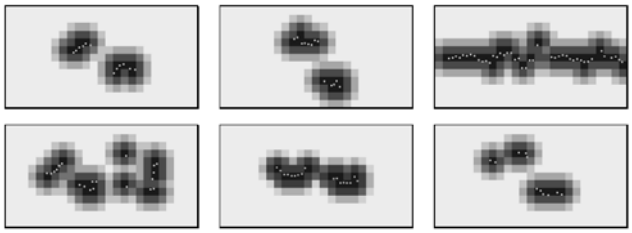


Fig. 7 *Top left* to *bottom right*: Typical exemplars of the classes *pedestrian*, *skater*, *cyclist*, *buggy*, *suitcase*, and *kangaroo*. The direction of motion is from *left* to *right*. Pedestrians and skaters have very similar appearance but differ in their dynamics. Pedestrians and subjects on kangaroo-shoes have similar dynamics but different appearances (mainly due to metal springs attached at the backside of the shoes). We use both information to classify these objects

7 Experiments

We experimentally evaluated our approach with six object classes: *pedestrian* (PED), *skater* (SKA), *cyclist* (CYC), *person pushing a buggy* (BUG), *person pulling a suitcase* (SUI), and *people on kangaroo-shoes* (KAN) (see Fig. 1). We recorded a total of 500 tracks. The sensor employed was a SICK LMS291 laser range finder mounted at a height of 15 cm above ground. The tracks include walking and running pedestrians, skaters with small, wide, or no pace (just rolling), cyclists at slow and medium speeds, people pushing a buggy, pedestrians pulling a suitcase, and subjects on kangaroo shoes that walk slowly and fast. Note that pedestrians, skaters, and partly also kangaroo shoes have very similar appearance in the laser data but differ in their dynamics. See Fig. 7 for typical exemplars of each class. The implementation of our system runs in real-time on a 2 GHz single-core CPU. The cycle time for single tracks is around 43 Hz when sensor data are immediately available. Most time is spent in the k-means clustering algorithm (about 65%).

7.1 Supervised learning experiments

In the first group of experiments, we test the classification performance in the supervised case. Each training set was composed of a single, typical track for each class including their labels *PED*, *SKA*, *CYC*, *BUG*, *SUI*, or *KAN*. The exemplar models were then learned from these single tracks. Based on the resulting prototype models, we classified the remaining 494 tracks. We repeated this experiment ten times, the averaged results are shown in Table 2.

Pedestrians are classified correctly in 96.2% of the cases whereas 3.3% are incorrectly associated to the skater class. A manual analysis of these 3.3% revealed that the misclassification occurred typically with running pedestrians whose

Table 2 Classification rates in percent in the supervised experiment. Whereas the rows correspond to the ground truth, the columns contain the obtained classification results

Classes	PED	SKA	CYC	BUG	SUI	KAN
Pedestrian	96.2	3.3	0	0	0	0.5
Skater	2.4	97.5	0	0	0	0.1
Cyclist	0	1.6	98.4	0	0	0
Buggy	0	0	0	97.2	0	2.8
Suitcase	6.4	0	0	0	85.4	8.0
Kangaroo	14.7	0	0	0	0	85.3

appearance *and* dynamics resemble those of skaters. A percentage of 0.5% were classified to be a person on kangaroo-shoes. All these tracks belonged to running pedestrians, too. We obtain a rate of 97.5% for skaters with one track (0.1%) falsely classified as kangaroo-shoes and 2.4% classified as pedestrians. The latter group was found to skate slower than usual with a small pace, thereby resembling pedestrians. Cyclists are classified correctly in 98.4% of the cases. None of them was falsely recognized as pedestrians, buggies, suitcases, or person on kangaroo-shoes. But it appeared that the bicycle wheels produced measurements that resemble skaters taking big steps. This lead to a rate of 1.6% of cyclists falsely classified as skaters. A percentage of 97.2% of the buggy tracks were classified correctly. Only 2.8% were found to be a subject on kangaroo-shoes. In this particular case, the track contained measurements in which the front of the buggy was partially outside the field of view of the sensor with two legs of the person still visible. The pedestrians pulling a suitcase were correctly classified in 85.4%. Unfortunately, 6.4% were classified as pedestrians and 8% were considered to walk on kangaroo shoes. Typically, the people in these tracks walked with a lower pace, so that both legs and the suitcase appeared as the legs of a pedestrian or the kangaroo shoes. Subjects on kangaroo shoes were correctly recognized at a rate of 85.3% with 14.7% of the tracks falsely classified as pedestrians. The manual analysis revealed that the latter group consisted mainly of kangaroo shoe novices taking small steps and thus appearing like pedestrians.

In conclusion, we find that, given the limited information provided by the laser data and the high level of self-occlusion naturally occuring in this setting, the results indicate that our exemplar models are expressive enough to discriminate between relevant object classes accurately.

7.2 Unsupervised learning experiments

In the second experiment the classes were learned in an unsupervised manner. The entire set of 500 tracks from all six classes was presented to the system in random order.

Table 3 Unsupervised learning results. Whereas the rows contain the learned classes, the columns show the number of classified objects. The last column shows the manually added labels and the last row contains the total number of tracks of each class

Classes	PED	SKA	CYC	BUG	SUI	KAN	
Class 1 (209)	187	5	0	0	3	17	"PED"
Class 2 (114)	7	107	0	0	0	0	"SKA"
Class 3 (41)	0	0	41	0	0	0	"CYC"
Class 4 (23)	0	0	23	0	0	0	"CYC"
Class 5 (26)	0	0	1	25	0	0	"BUG"
Class 6 (23)	0	0	0	23	0	0	"BUG"
Class 7 (38)	0	0	0	0	38	0	"SUI"
Class 8 (23)	0	0	0	0	23	0	"SUI"
Total (500)	194	112	65	48	64	17	

Each track was either assigned to an existing class or was taken as basis for a new class according to the learning procedure described above. As can be seen in Table 3, eight classes have been generated for our data set: one class for *pedestrians* (PED), one for *skaters* (SKA), two for *cyclists* (CYC), two for *buggies* (BUG), two for *suitcases* (SUI), and none for *kangaroo shoes* (KAN).

Class one (labeled PED) contains 187 pedestrian tracks (out of 194), 5 skater tracks, 4 suitcase tracks, and 17 kangaroo tracks resulting in a true positive rate of 89.5%. Class two (labeled SKA) holds 107 skater tracks (out of 112) and 7 pedestrian tracks yielding a true positive rate of 93.9%. Given the resemblance of pedestrians and skaters, the total number of tracks and the extent of intra-class variety, this is an encouraging result that shows the ability of the system to discriminate objects that vary predominantly in their dynamics. Classes three and four (labeled CYC) contain 41 and 23 cyclist tracks respectively. No misclassifications occurred. The classes five and six (labeled BUG), hold 25 and 23 buggy tracks with a bicycle track as the single false negative in class five. The last two classes, seven and eight (labeled SUI), consists of 38 and 23 tracks of pedestrians pulling a suitcase. Again no misclassifications occurred. The representation of cyclists, buggies, and suitcases by two classes is due to the larger variability in their appearance and more complex dynamics. The discrimination from the other three classes is exact—no pedestrians, skaters, or subjects on kangaroo shoes were classified to be a cyclist or a buggy.

The system did not produce a specific class for subjects on kangaroo shoes as all instances of the latter class were included in the pedestrian class. The best known model for all 17 kangaroo tracks was always class one which has previously been created from a pedestrian track. This results in a false negative rate of 8.1% from the view point of the pedestrian class. This result confirms the outcome of the supervised experiment where the highest misclassification rate

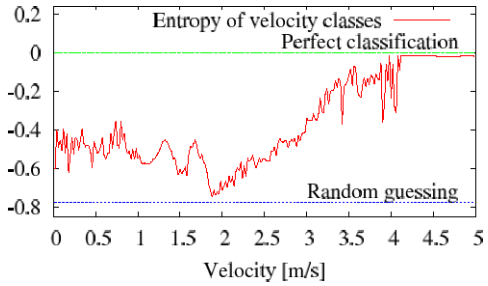


Fig. 8 Analysis of the track velocities as alternative features for classification. While high velocity is a strong indicator for a certain class (CYC), there is a higher confusion in the low and medium range

(14.7%) was found to be between pedestrians and subjects on kangaroo shoes (see Table 2).

7.3 Analysis of track velocities

The data set of test trajectories that was used in our experiments contains a high level of intra-class variation, like for example skaters moving significantly slower than average pedestrians or even pedestrians running at double their typical velocity. To visualize this diversity and to show that simple velocity-based classification would yield unsatisfactory results, we calculated a velocity histogram for all six classes. For every velocity bin, we calculated the entropy $H(v_i) = \sum_{j=1}^{6}(p(c_j|v_i) \cdot \log p(c_j|v_i))$ and visualized the result in Fig. 8. Note that the uniform distribution over six classes, which corresponds to random guessing, has an entropy of $6 \cdot (1/6 \cdot \log(1/6)) \approx -0.778$, which is shown by a straight, dashed line. As can be seen from the diagram, high and low velocities are strong indicators for certain classes while there is a high level of confusion in the medium range.

7.4 Classification with a mobile robot

To demonstrate the practicability of the approach for a mobile sensor, an additional supervised and unsupervised experiment was carried out with a moving platform. A total of 18 tracks has been collected: 3 pedestrian tracks, 5 skater tracks, 4 cyclist tracks, and 6 suitcase tracks (kangaroo shoes and buggies were unavailable for this experiment). The robot moved with a maximal velocity of 0.75 m/s and an average velocity of 0.35 m/s. A typical robot trajectory is depicted in Fig. 9.

For the supervised experiment, the trained models from one of the supervised experiments in Sect. 7.1 have been reused to classify the tracks collected from the moving platform. All objects were classified correctly. Table 4 contains the classification probabilities (t being the track length) averaged over all tracks in the corresponding class. The last two columns contain the probabilities for the classes BUG and KAN, all being close to zero. The lowest classification

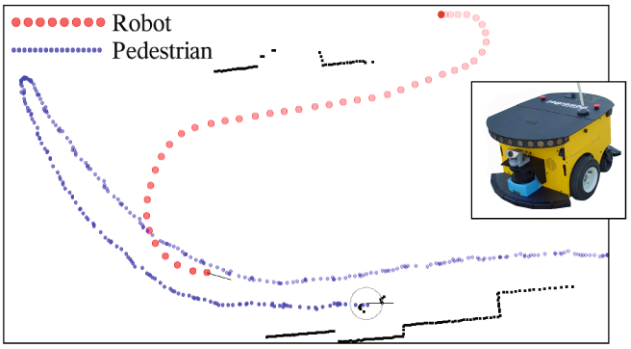


Fig. 9 Trajectory of the robot (an ActivMedia PowerBot) and a pedestrian over a sequence of 450 observations

Table 4 Averaged classification probabilities for the supervised experiment with the moving platform. All objects have been classified correctly

Classes	PED	SKA	CYC	SUI	BUG	KAN
Pedestrian	0.86	0	0	0	0	0.14
Skater	0.08	0.83	0	0	0	0.09
Cyclist	0.01	0	0.85	0.08	0.06	0.01
Suitcase	0.01	0	0	0.94	0.03	0.02

probability in this experiment was a pedestrian track which still had the probability 0.73 of being a pedestrian.

In the unsupervised experiment, the tracks have been presented to the system in random order without prior class information. The clustering result compared to human classification was exact: four classes were created autonomously—each containing the tracks of exactly one object category.

8 Conclusions and outlook

We have presented an unsupervised learning approach to the problem of tracking and classifying dynamic objects. In our framework, the appearance of objects in planar range scans is represented by a probabilistic exemplar model in conjunction with a hidden Markov model for dealing with the dynamically changing appearance over time. Extensive real-world experiments including 500 recorded trajectories show that (a) the model is expressive enough to yield high classification rates in the supervised case and that (b) the unsupervised learning algorithm produces meaningful object classes consistent with the true underlying class assignments. Additionally, our system does not require any manual class labeling and runs in real-time.

In future research, we plan to strengthen the interconnection between the tracking process and the classification module, i.e., to improve segmentation and data association given the estimated posterior over future object appearances.

Acknowledgements This work has partly been supported by the EC under contract number FP6-IST-045388, the German Federal Ministry of Education and Research (BMBF) within the research projects DESIRE under grant number 01IME01F and the German Research Foundation (DFG) under contract number SFB/TR-8.

References

Arras, K. O., Martínez Mozos, O., & Burgard, W. (2007). Using boosted features for the detection of people in 2d range data. In *Proc. of the int. conf. on robotics & automation*.

Chis, M., & Grosan, C. (2006) Evolutionary hierarchical time series clustering. In *6th int. conf. on intelligent systems design and applications (ISDA)* (pp. 451–455). Washington, DC, USA.

Cui, J., Zha, H., Zhao, H., & Shibasaki, R. (2006). Robust tracking of multiple people in crowds using laser range scanners. In *18th int. conf. on pattern recognition (ICPR)*. Washington, DC, USA.

Cutler, R., & Davis, L. (2000). Robust real-time periodic motion detection, analysis, and applications. *PAMI*, *22*(8), 781–796.

Drumwright, E., Jenkins, O. C., & Mataric, M. J. (2004). Exemplar-based primitives for humanoid movement classification and control. In *Proc. of the int. conf. on robotics & automation (ICRA)*.

Fod, A., Howard, A., & Mataríc, M. (2001). Fast and robust tracking of multiple moving objects with a laser range finder. In *Proc. of the int. conf. on robotics & automation (ICRA)*.

Fod, A., Howard, A., & Mataríc, M. (2002). Laser-based people tracking. In *Proc. of the int. conf. on robotics & automation (ICRA)*.

Fortuna, J., & Capson, D. (2004). Ica filters for lighting invariant face recognition. In *17th int. conf. on pattern recognition (ICPR)* (pp. 334–337). Washington, DC, USA.

Ghahramani, Z. (2004). *Unsupervised learning*. New York: Springer.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Ilg, W., Bakir, G., Mezger, J., & Giese, M. (2003). On the representation, learning and transfer of spatio-temporal movement characteristics. In *Humanoids*, July 2003, electronical version.

Jenkins, O. C., & Matarić, M. J. (2004). A spatio-temporal extension to Isomap nonlinear dimension reduction. In *The int. conf. on machine learning (ICML)* (pp. 441–448). Banff, Alberta, Canada, July 2004.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 733–795.

Kruger, V., Zhou, S., & Chellappa, R. (2006). Integrating video information over time. example: Face recognition from video. In *Cognitive vision systems* (pp. 127–144). New York: Springer.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*, 1783–1816.

MacLachlan, R., & Mertz, C. (2006). Tracking of moving objects from a moving vehicle using a scanning laser rangefinder. In *Intelligent transportation systems 2006* (pp. 301–306). New York: IEEE.

Montemerlo, M., & Thrun, S. (2002). Conditional particle filters for simultaneous mobile robot localization and people tracking. In *Proc. of the int. conf. on robotics & automation (ICRA)*.

Ng, H. T., & Lee, H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of the 34th annual meeting on association for computational linguistics* (pp. 40–47).

Plagemann, C., Müller, T., & Burgard, W. (2005). Vision-based 3d object localization using probabilistic models of appearance. In *Pattern recognition, 27th DAGM symposium*, Vienna, Austria (Vol. 3663, pp. 184–191). New York: Springer.

Schulz, D. (2006). A probabilistic exemplar approach to combine laser and vision for person tracking. In *Robotics: science and systems*. Cambridge: The MIT Press.

Schulz, D., Burgard, W., Fox, D., & Cremers, A. (2001). Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In *Proc. of the int. conf. on robotics & automation (ICRA)*.

Tasoulis, D. K., Adams, N. M., & Hand, D. J. (2006). Unsupervised clustering in streaming data. In *6th int. conf. on data mining—workshops (ICDMW)* (pp. 638–642). Washington, DC.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Thrun, S. (2001). An probabilistic online mapping algorithm for teams of mobile robots. *International Journal of Robotics Research*, *20*(5), 335–363.

Toyama, K., & Blake, A. (2002). Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, *48*(1), 9–19.

Wang, Y., & Han, J.-Q. (2005). Iris recognition using independent component analysis. *Machine Learning and Cybernetics*, *7*, 4487–4492.

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2006). Gaussian process dynamical models. In *Advances in neural information processing systems* (pp. 1441–1448). Cambridge: The MIT Press. Proc. NIPS'05.

Wren, C., Azarbayejani, A., Darrell, T., & Pentl, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 780–785.

Matthias Luber received the Diploma degree in computer science, with specialization in artificial intelligence and mobile robotics from the University of Freiburg, Freiburg, Germany, in 2007. He is currently working towards the Ph.D. degree in robotics in the Social Robotics Laboratory, University of Freiburg. His research interests include people tracking, learning, motion prediction, and probabilistic multi-hypothesis tracking.

Kai O. Arras received his Masters in electrical engineering from ETH Zurich in 1996 and his Dr. es science degree from EPFL Lausanne in 2003. After a post-doctoral stay at the Center for Autonomous Systems at KTH Stockholm, he founded a robot company, Nurobot Automation and Artefacts GmbH, Zurich. In 2006 he joined the Autonomous Intelligent Systems Group at the University of Freiburg as a post-doc, and in 2007, he joined Evolution Robotics, Pasadena, USA, as a Senior Research Scientist. Since July 2008 he holds a Junior Research Group Leader position at the University of Freiburg and is head of the Social Robotics Laboratory in the Computer Science Department.

Christian Plagemann is a postdoctoral researcher at the Artificial Intelligence Lab of Stanford University, USA. He obtained his Ph.D. in computer science from the University of Freiburg, Germany, and his M.Sc. in computer science from the University of Karlsruhe, Germany. His research interests include machine learning, robotics and autonomous systems technology with a special focus on nonparametric Bayesian methods and regression problems.

Wolfram Burgard is a professor for computer science at the University of Freiburg where he heads of the Laboratory for Autonomous Intelligent Systems. He received his Ph.D. degree in Computer Science from the University of Bonn in 1991. His areas of interest lie in artificial intelligence and mobile robots. In the past, Wolfram Burgard and his group developed several innovative probabilistic techniques for robot navigation and control. They cover different aspects including localization, map-building, path-planning, and exploration. Wolfram Burgard is a member if IEEE and AAAI and Fellow of the European Coordinating Committee for Artificial Intelligence (ECCAI).