in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische Universitt Ilmenau, Germany, Vol. I, pp. 451-456, Universittsverlag 2007

S. Müller / A. Scheidig / A. Ober / H.-M. Gross

Making Mobile Robots Smarter by Probabilistic User Modeling and Tracking

Abstract

This paper considers the problem of tracking and modeling users of a mobile service robot. In order to adapt the behavior of an interacting robot to the user's preferences, the system has to know about the people in its surrounding and to model their properties. To that purpose, a consistent probabilistic model is introduced here, which is realizing the tracking process and storing the information.

1 Introduction

The weak acceptance of a robot actively offering services is a hard problem. People are not willing to interact with a stupid computer. One way to increase the rate of interactions is to make the robot smarter by selecting a behaviour more appropriate to the specific person. One essential step on that way, is to model the state of a user properly. In particular, the robot has to know where people are in its surrounding and what are their objectives. The task of people tracking typically is done using different sensors which are integrated in a probabilistic model, e.g. a particle or kalman filter. In former work [3] we developed a probabilistic multi-cue people tracker, which successfully runs on our shopping robot SCITOS. Besides the presence of possible users, further information like their gender, age and if they are in a hurry or willing to interact are necessary for an adequate reaction of the robot. Therefore, the simple model of people's positions has been improved by modeling different people and their properties. Because many of these properties can not be observed continuously, it is necessary to remember and recognize people, which typically is done by analyzing face images [5] or by a color model of a person. In our model both methods are used as cues for identification, which is done implicitly in the model. Based on a knowledge base like that, the robot can infere various facts about its situation for an intelligent action selection.

On a mobile robot there are different sensory systems gathering information about the robots environment (see fig. 1). A probabilistic model is the central place where all the

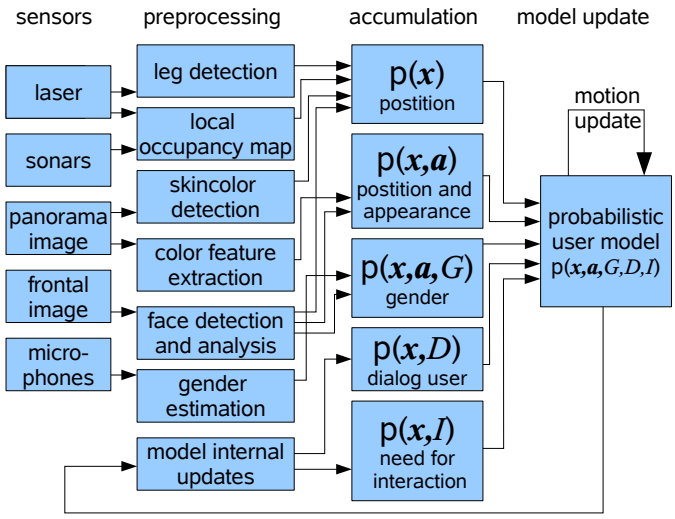

Figure 1: Layers of our systems architechture: (left) sensors providing raw data, (middle) preprocessing modules extracting information, (right) short time aggregation of information for model update

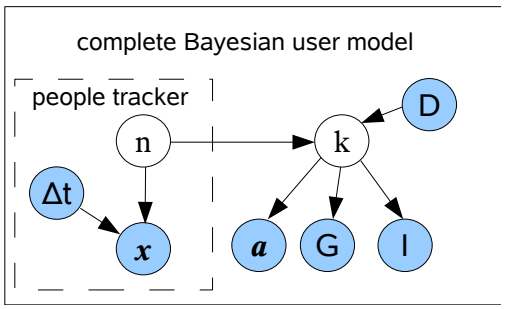

Figure 2: Bayesian network of the probabilistic model: $\mathbf{x}$ position of hypothesis n, Δt time difference for represening movement trajectories, $\mathbf{a}$ appearance vector (face and color) of person k, D is the person in dialog with the robot, I need to interact

data are merged together. Our robot HOROS, a Pioneer platform based interacting robot, is equipped with some range sensors like a laser and sonars, two cameras yielding a panorama image and frontal images, and a pair of microphones for sound analysis mounted at the head.

In the following section the architechture of our system is described in more detail, before the probabilistic user model is explained. At the end of the paper some ongoing experiments are shown which are made for learning a classification of people's movement trajectories.

2 System Architecture

The environment recognition system of HOROS is based on a couple of sensors providing raw data as ranges, visual inputs, and a stereo audio signal. Based on these, different preprocessing modules are extracting information on people hypotheses in the surrounding of the robot. A brief overview on these modules is given below. After single observations of possible people's positions $\mathbf{x}$, person's color profile and facial features $\mathbf{a}$ or their gender G are extracted, they are aggregated in the accumulation layer. Similar to the human brain here the different modalities are combined with their sensory context. Thus, observations on the appearance of a person are combined with a position, bearing only position estimation

from color detection are combined with distance measurements in the local occupancy map. Further, an extracted color pattern or an appearance vector from face analysis is assigned to a position and gender estimation from face analysis as well as from sound analysis are associated to an appearance $\mathbf{a}$ and a position. As a last source of information there are model internal observations. E.g. the need of a person to interact with the robot can be estimated from the movement trajectory (see below) and the person standing in front of the robot is supposed to be that user who is in dialog with the robot. In addition to the improved quality of observations by combining them, the accumulation layer compensates the asynchronously occuring observations using a time slicing mechanism. In each time slice all the observations are accumulated before one update at the model is done.

Knowing all the aggregated observations, which are represented as a joint probability, the probabilistic model is updated. In fig. 2 the Bayesian network is shown. Here, besides the shaded circles, representing observed variables, the non shaded circles contain discrete hidden variables for the component n in a mixture of Gaussians for the position $\mathbf{x}$ and the id k of the known people. The user properties are represented as conditional probabilities $p(\mathbf{a}|k)$ for the appearance, which is modelled as a Gaussian, $p(G|k)$ for the Bernoulli distributed gender, and $p(I|k)$ for the also Bernoulli distributed property of need for interaction with the robot. In order to decide who is the person, which is currently in dialog with the robot, $p(k|D)$ is representing a discrete distribution over the person ids k. There is one k standing for a non person or nobody, such that it is possible to express that nobody is in dialog, as well as that a position hypothesis n belongs to a non human object.

On each time step, the model is updated in three steps. In order to realize a Bayesian state estimation, similar to a Bayes-filter, first we have to perform a motion update. Here the positions are propagated according to the history which is represented by Δt and a motion model of the robot. Further the property of being dialog user $p(k|D)$ and the need for interaction $p(I|k)$ are diffused in order to allow a change over time. Second step is to infere the hidden and the unobserved variables given the current observations from accumulation layer. From that step we get a new prior distribution for the third step, which is a MAP estimation of the distribution parameters given the new observation and the prior model from last time step. Following this update regime, the model always contains the complete knowledge about people in the surrounding of the robot, but also the knowledge on absent people inspected before. With the current model the robot can easily infere e.g. the position of the dialog user by evaluation of $p(\mathbf{x}|D = 1)$ or the gender of a person trying to interact

by infering $p(G|I = 1, D = 0)$. Du to the simple structure of the Bayesian network, it is easy to introduce further variables for user properties. All it takes is a source of information producing observations on this variable. Due to the ability of our face analysis system to estimate the age of a person, we would easily be able to add a variable A and a conditional probability $p(A|k)$ to represent the age of a user.

3 Recognition subsystems

As mentioned above, there are different preprocessing modules generating observations on the various state variables. A first module is consuming the laser scans, while classifying the segments of the scan as a pair of legs. Using a simple collection of criteria explained in [1], segments with a limited variance in distance and a defined size are classified as possible legs. If there are two legs within a distance of less than $0.5m$ a Gaussian hypothesis for a position $\mathbf{x}$ observation is generated and sent to the accumulation layer. The laser scan and the noisy sonar measurements furthermore are integrated into an occupancy gridmap representing the local environment of the robot. This map is used by the accumulation layer to query the distance of objects observed in a known direction, as the skin color detections in the panorama image. Using a multi-instances particle filter [6] which is tracking the skin colored regions in the image, there are Gaussian hypotheses for directions of people which are limited in their distance by a map lookup.

To get a feature for disinguishing people, a color pattern is taken from the panorama image by determining the average UV value (luminance independent parts of YUV color space) and the variance of that pixels of three regions on the upper part of the body of people in the image. These are the first components $a_1, \ldots, a_{12}$ of the appearance vector. To find these regions, a parametric contour and skincolor model is fitted into the panorama image using a Monte-Carlo gradient descent. Following the list of modules in fig. 1, the next is the face detection and analysis. Here for detection the well known Viola and Jones detector [4] is used. For analyzing the face then an Active Appearance Model is used, which is tracking the frontal face while delivering parameters for position shape and texture of the face. The parameters for shape and texture are used as further components of the $\mathbf{a}$ vector. Furthermore, the direction of the face and the distance from local map are used to generate a Gaussian position hypothesis. By classifying the appearance parameters it is possible to classify the gender of the person tracked, which is used to generate a $p(G, \mathbf{x}, \mathbf{a})$ observation. A further sensor based cue is the sound analysis. Here sound source localization is utilized to

get a hint for the direction of a speaker, before a speech detection and a gender classification is done [2]. The classification is based on Mel Frequency Cepstral Coefficients and on the fundamental frequency of the speaker. Also using a lookup of the distance the result of that module is a probability of the gender and a given position.

In contrast to the other variables, the observations for D and I do not result from external sensoric inputs. For updating these parts of the model, information from the model itself is used. The property of being the user who is in dialog with the robot, is estimated from the users position (similar to our shopping assistant SCITOS [3]). Thus each person standing in or approaching the region in front of the robot is supposed to be in dialog with the robot. The need for interaction also is only extracted from the movement trajectory of people. For generating observation distributions, first for each object $i \in \{1, \ldots, n\}$ in the model the movement trajectory $p(\mathbf{x}, \Delta t | n = i)$ is infered. Then using a heuristic measurement model $p(D|\mathbf{x}, \Delta t)$ and $p(I|\mathbf{x}, \Delta t)$ is used to infere the current distribution of D and I. Together with the current position $p(\mathbf{x}|\Delta t = 0, n = i)$ a new observation for updating the model is composed.

4 Experiments and Results

Because it is difficult to quantify such a probabilistic model, experiments have been done for single parts of the model separately. To evaluate the accuracy of the position tracking system (fig. 2 left), a simple experiment with a given reference from a top view camera has been done. By comparing the tracking hypotheses to the reference, we got a distance error below $0.5m$. Fig. 3 shows some exemplary trajectories taken during the experiments. The soid line is the estimated path and the dotted one is the reference. The shape of graphs suggest that the movement trajectories are quite specific for people who are interested in an interaction (graphs a and b) and those who are not (graphs c and d). For evaluating the abilities of the model to estimate I currently a complex experimental session has been accomplished. The robot there was running an information terminal application, while capturing the movement tajectories of the people in its surrounding. After a person completed its interaction or when a person only passed the robot, a questionnaire has been filled during an interview. By asking, if people are in a hurry or not and if they had an intention to use the information terminal, we intend to create an empiric measurement model $p(I|\mathbf{x}, \Delta t)$. First findings on the labeled trajectories are disillusioning. There are many factors of influence, which are not included in the model yet. Thus for increasing the reliability of an empiric model, the

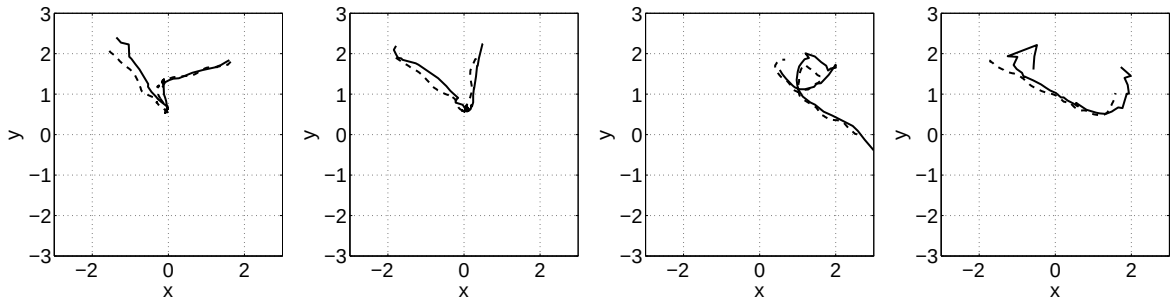

Figure 3: Exemplary trajectories of people moving in the surrounding of the robot, robot standing at $(0, 0)$ facing upwards, solid line: estimated path, dashed line: top down reference

specifics of the environment have to be taken into account. Depending on the position of doors and other points of interest in the room, the model will be suitable only for a fixed position. Therefore, higher effort will be necessary on encoding the trajectories for the small number of train data (157 non interacting people, 53 interactions at three different locations) to allow a satisfying generalization.

5 Conclusion

In this paper an overview of our probabilisic user model is given. We could show that tracking of positions works well, but evaluation of the model parts for user identification and estimation of user properties based on trajectories are still in progress.

References

[1] K. O. Arras, Ó. M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *Proc. of 2007 IEEE International Conference on Robotics and Automation*, Roma, Italy, 2007.

[2] R. Brueckmann, A. Scheidig, C. Martin, and H.-M. Gross. Integration of a sound source detection into a probabilistic-based multimodal approach for person detection and tracking. In *Proc. Autonome Mobile Systeme (AMS 2005)*, pages 131–137. Springer, 2005.

[3] S Müller, H.-M. Gross, A. Scheidig, and H.-J. Boehme. Are you still following me? In *Proc. of the 3rd European Conference on Mobile Robots (ECMR2007)*, 2007.

[4] P. Viola and M. Jones. Fast and robust classication using asymmetric adaboost and a detector cascade. In *NIPS 2001*, pages 1311–1318, 2001.

[5] Böhme H.-J Gross H.-M. Wilhelm, T. Classification of face images for gender, age, facial expression, and identity. In *Artificial Neural Networks: Biological Inspirations - ICANN 2005, LNCS 3696*, volume I, pages 569–574, 2005.

[6] T. Wilhelm, H.-J. Boehme, and H.-M. Gross. A multi-modal system for tracking and analyzing faces on a mobile robot. In *Robotics and Autonomous Systems*, volume 48, pages 31–40, 2004.

Dipl.-Inf. Steffen Müller, Dr.-Ing. Andrea Scheidig, Antje Ober,

Prof. Dr. Horst-Michael Gross

Ilmenau Technical University, Neuroinformatics and Cognitive Robotics Lab, POB 10 05 65, 98684 Ilmenau