Wilhelm, T., Boehme H.-J., Gross, H.-M.: Sensor Fusion for Vision and Sonar Based People Tracking on a Mobile Service Robot. in: Proc. Int. Workshop on Dynamic Perception 2002, Bochum, pp. 315-320, IOS Press, infix 2002

Sensor Fusion for Vision and Sonar Based People Tracking on a Mobile Service Robot

T. Wilhelm, H.-J. Böhme and H.-M. Gross

Ilmenau Technical University, Department of Neuroinformatics, P.O.Box 100565, 98684 Ilmenau
{wilhelm, hans, homi}@informatik.tu-ilmenau.de

Abstract. Service robots intended to interact with people must be able to localize and continuously track their users. A method is described which integrates information from visual and sonar based tracking pathways while updating hypotheses about the position of the robot's human user. Each tracking method uses information from the other to generate a more robust measure of the user's position, and thus a more robust behavior generation is achieved.

1 Introduction

A service robot, which is designed to serve people in special domains or to help them in their everyday life, must be able to localize and continuously track its users. If the user breaks the interaction off, there is no need for the robot to continue to produce any outputs. Lacking these capability would result in a robot, which is trying to contact arbitrary things or which is proceeding to offer its services even when the user already left the operation area. The authors consider the knowledge about the position of the user as fundamental for a smart appearance of any service robot. On the other side, the price determines the economical success of any service robot application, so it seems favorable to use cheap hardware whenever possible, which has consequences on the complexity of any people tracking algorithm.

Our experiments were carried out in a home store, where our service robot is to operate as a mobile shopping assistant, guiding customers to desired products in the store [1]. A major problem concerning people tracking in this environment are the varying illumination conditions from natural to artificial lighting, which imply a multimodal approach to the problem, not only relying on visual cues.

2 Tracking

Tracking of users can be realized by using different sensor systems. The distance to an object can be measured by means of sonar or laser data, and there are methods that extract hypotheses about the position of people in the robot's surroundings from laser data [6]. In contrast to laser scanners, the resolution and accuracy of sonar sensors give only a vague hint about the nature of the object,

and it seems that these methods can not be assigned to cheap sensor systems such as sonars. Moreover, the used features are not very person-specific and could detect other objects as potential users as well. Cameras can be considered as cheap sensors compared to laser scanners, and visual data can be used to solve ambiguous situations and to discriminate people from arbitrary objects. Thus the proposed tracking method consists of a sonar and a vision based tracking module.

2.1 Sonar Based Tracking

The task of the sonar based tracking is to always keep contact to the user by moving the robot according to its mode of operation and the position of the user. Our experiments were carried out on a B21 mobile robot (RWI IS Robotics) equipped with two layers of sonar sensors with 24 sonars respectively. The raw sensor data is noisy and depends on the orientation and the material of the objects around the robot. Therefore the raw data is preprocessed as follows:

1. replacement of invalid measurements: distances larger than $22,5m$ are considered as invalid and are replaced by the previous measurements
2. local spatial low pass filtering of adjacent measurements
3. temporal low pass filtering of successive measurements
4. calculation of a weighting factor in each direction which is inversely proportional to the measured distance $W_{Sonar}^{(c)} = 1 - d_{sonar}^{(c)}/d_{max}$, where $d_{sonar}^{(c)}$ is the preprocessed sonar measurement at position c in the scan and d_{max} is the maximum distance $(1,5m)$; for distances larger than d_{max} the weight is set to 0

The position of the maximum in the resulting weighting vector corresponds to the nearest object (see Figure 2e) and is used to generate an appropriate behavior, depending on the robots mode of operation:

1. *communication:* orient the touch interface mounted on top of the robot to the position of the maximum, thus allowing the user to make inputs
2. *guide user:* keep the distance to the user small and stay in front of him, while driving towards a goal position in the market
3. *follow user:* keep distance to user small and try to stay behind him

The advantages of the sonar based tracking are its low computational costs and thus its ability to continuously track the user and align the robot appropriately. It generates an adequate behavior as long as the nearest object is really the user, otherwise the robot reacts to any object in its surroundings and tries to interact with it. This drawback can be encountered by integrating information from a vision based tracking module, which is able to distinguish people from any other objects in the area.

2.2 Vision Based Tracking

The basis of the vision based tracking procedure is the condensation algorithm [4]. The task of calculating the probability of the existence of a person for every pixel and tracking the resulting density function is solved by an approximation of the density function by a relatively small number of samples. The condensation algorithm operates on the panorama images from an omnidirectional color camera and uses different feature extraction methods to calculate hypotheses about humans faces and the upper part of the human body. Compared to a panoramic image with 720×106 pixels calculating the feature extraction only for 200 samples yields a reduction to merely 0.262%.

Skin Color A widely used method for finding faces in images is skin color classification. Here the dichromatic r-g-color space ($r = R/(R+G+B)$, $g = G/(R+G+B)$) is used, which is widely independent from variations in luminance. The color model consists of a look up table with manually classified skin color pixels in the r-g-color space [3]. To prevent the color model from getting holey because of insufficient training data, there is a small Gaussian placed around each skin color pixel. The skin color model is depicted in Figure 1. The color detection can be calculated very fast but it is highly dependent on illumination color and variations in hue and often fails in back light situations.

Head-Shoulder-Contour The second method uses a contour model which describes the mean head-shoulder-contour of a person [2]. The model Λ was derived from a number of images containing frontally aligned persons. On the mean gray level image, the local orientations were calculated with a structure tensor [5]. The same tensor is used during head-shoulder-contour detection to calculate the gray value orientation in a local surrounding around each sample, and the template matching is carried out for every sample according to equation 1, where o is the orientation in the image and λ is the orientation in the contour model. Figure 1 depicts the head-shoulder-contour model Λ of size 20×20.

$$W_{hsc}(x,y) = \frac{\sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \frac{1}{2} \left[\cos(2|\lambda_{i,j} - o(x-i, y-j)|) + 1\right]}{card(supp(\Lambda))} \tag{1}$$

The head-shoulder-contour is computational more expensive and not as person specific as the skin color detection, but it yields good results in back light situations, where any other gray value or color based face detector fails.

Combination of the Vision Based Cues Although both cues are person-specific, it can happen that they do not detect a user or give false alarms. Therefore both cues are combined by a fuzzy min-max-operator ($minmax(a,b) = \gamma \min(a,b) + (1-\gamma)\max(a,b)$), which can be configured between a pessimistic and an optimistic fusion. Pessimistic (min, $\gamma = 1$) means that an user which was not detected by at least one cue is not accounted for at all, while using the

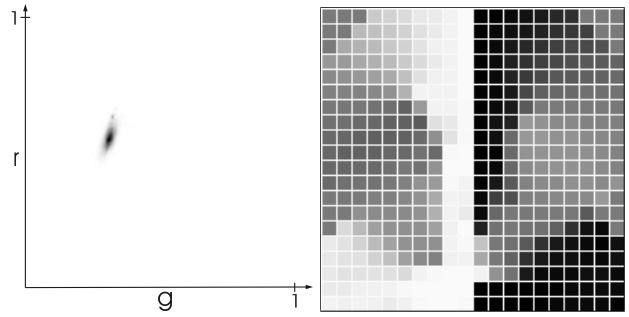

Fig. 1. Models used for vision based tracking. *Left*: Color model in the dichromatic r-g-color space. *Right*: Head-shoulder-contour model, local orientations are represented by gray levels, where white and black pixels code horizontal, and gray pixels code vertical edges respectively

max ($\gamma = 0$) fusion results in a behavior, where all false positive matches from one cue are considered valid. See Figure 2 for results of the single cues and their combination.

3 Sensor Fusion

As mentioned before, vision based tracking shall now be used to prevent the sonar based tracking from interacting with arbitrary non-human objects. On the other hand, the vision based tracking can benefit from the sonar based method by using it as third cue for calculating the sample weighting.

Support of Vision Based Tracking by Sonar Data Since the sonar scan as well as the image constitute an 360° description of the robots surroundings, it is possible to assign a scan measurement at position c in the scan to each position $\boldsymbol{x}$ in the image. This way, the sonar vector can be used to modulate the sample weighting in the condensation algorithm, equation 2 and 3. Thus only those samples get a high weight, that are supported by the vision based cues and, at the same time, lie in a direction with a short distance measured from the sonar sensors. Samples that are only supported either by the vision or the sonar based tracking eventually die out (Figure 3).

$$W^{(i)}_{Sample}(\boldsymbol{x}) = minmax\left(W^{(i)}_{skincolor}(\boldsymbol{x}), W^{(i)}_{hsc}(\boldsymbol{x})\right) W_{sonar}(c) \tag{2}$$

$$P^{(i)}_{Sample}(\boldsymbol{x}) = \frac{W^{(i)}_{Sample}(\boldsymbol{x})}{\sum_i W^{(i)}_{Sample}(\boldsymbol{x})} \tag{3}$$

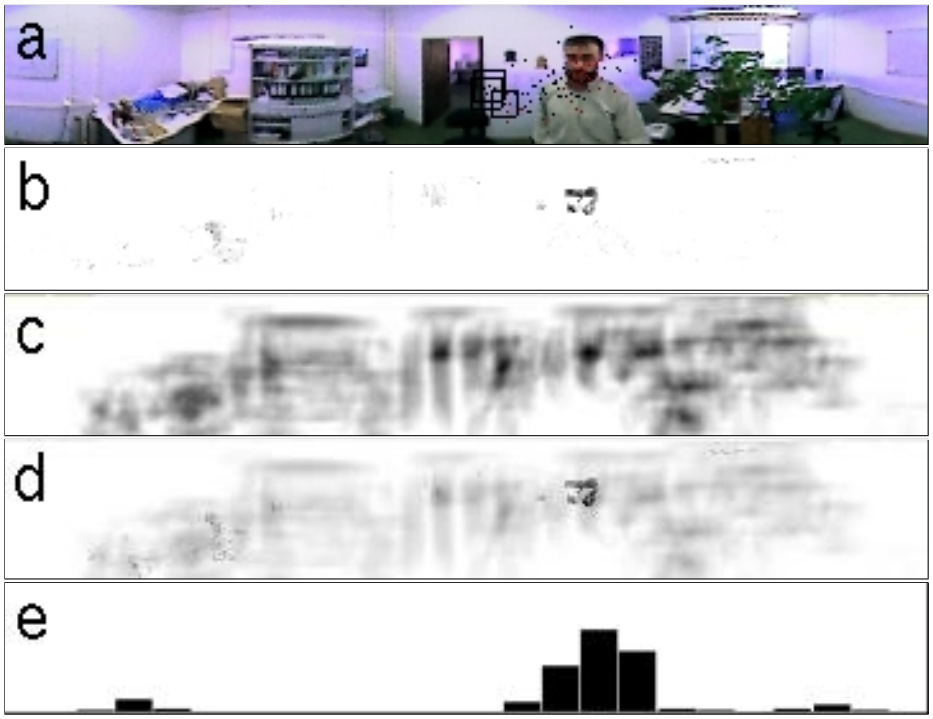

Fig. 2. Results of the single vision based tracking modules and the sonar based tracking: a) original panoramic image; b) skin color classification; c) head-shoulder-contour detection; d) $MinMax$-fusion ($\gamma = 0.7$), note that at the position of the users head, both cues give the largest contribution; e) weighting factors W_{hsc} calculated from the sonar scan

Support of Sonar Tracking by Vision Based Data Since only the sonar based tracking is responsible for behavior generation, the case where vision based data supports sonar tracking is more important. The camera image is divided into columns corresponding to the single sonar measurements. In every column c, the sum of the sample weights is calculated, resulting in a vector with high values on those positions where most likely the user is. For behavior generation, the positions of the maxima in the sonar and vision based scan are compared. If they are aligned, the motor commands are executed, otherwise all actions are suppressed. Thus, other people can approach, without the robot turning away from its current user.

4 Summary

The paper presents the integration of a sonar and a vision based user tracking pathway into a robust tracking procedure, which was applied successfully on a mobile service robot in a home store.

Fig. 3. Comparison of pure vision based tracking (*left*) and vision based tracking with sonar support (*right*). Every 10th image in the sequence is shown; the user moves around the robot (sometimes the robot is turning to the user based on sonar tracking). While at the left many samples get stuck on other objects, the tracking with sonar support does not loose the user

5 Outlook

In our current work, we investigate possibilities of automatic camera color calibration to get the skin color classification independent from variations in illumination color. In addition to that, we analyse the performance of other feature extraction and face detection methods, such as cascade correlation neural networks for the vision based tracking pathway. Furthermore, a robotic face with two cameras was designed, which is always oriented towards the currently tracked person. High resolution images from these frontally alligned cameras can be used to recognize a user who was lost from the omnidirectional view during tracking.

References

1. Boehme, H.-J., Wilhelm, T., Key, J., Schroeter, Ch., Hempel, T., and Gross, H.-M. An Approach to Multimodal Human-Machine Interaktion for Intelligent Service Robots. In *EUROBOT'01 - the fourth Euromicro Workshop on Advanced Mobile Robots*, volume 86 of *Lund University Cognitive Studies*, pages 17–24. Lund University, 2001.
2. Corradini, A., Boehme, H.-J., and Gross, H.-M. A Hybrid Stochastic-Connectionist Approach to Gesture Recognition. *International Journal on Artificial Intelligence Tools*, 2000(9):177–204, 2000.
3. Feyrer, S. *Detektion, Lokalisierung und Verfolgung von Personen mit einem mobilen Serviceroboter.* PhD thesis, Eberhard-Karls-Universität Tübingen, 2000.
4. Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.
5. B. Jähne. *Digitale Bildverarbeitung.* Springer-Verlag, Berlin Heidelberg, 3. Auflage, 1993.
6. Schulz, D. and Burgard, W. Probabilistic state estimation of dynamic objects with a moving mobile robot. *Robotics and Autonomous Systems*, 34:107–115, 2001.