Proc. 5th Int. Conf. on Multimodal Interfaces (ICMI-PUI' 03), Vancouver, pp. 297-300, acm press 2003

Towards An Attentive Robotic Dialog Partner

Torsten Wilhelm
Ilmenau Technical University
Dept. of Neuroinformatics
P.O.Box 100565, 98684
98684 Ilmenau
torsten.wilhelm@tu-ilmenau.de

Hans-Joachim Böhme
Ilmenau Technical University
Dept. of Neuroinformatics
P.O.Box 100565, 98684
98684 Ilmenau
hans-joachim.boehme@tu-ilmenau.de

Horst-Michael Gross
Ilmenau Technical University
Dept. of Neuroinformatics
P.O.Box 100565, 98684
98684 Ilmenau
horst-michael.gross@.tu-ilmenau.de

ABSTRACT

This paper describes a system developed for a mobile service robot which detects and tracks the position of a user's face in $3D$-space using a vision (skin color) and a sonar based component. To make the skin color detection robust under varying illumination conditions, it is supplied with an automatic white balance algorithm. The hypothesis of the user's position is used to orient the robot's head towards the current user allowing it to grab high resolution images of his face suitable for verifying the hypothesis and for extracting additional information.

Categories and Subject Descriptors: H.5.2 [User Interfaces]: Theory and methods

General Terms: Algorithms

Keywords: User Detection, User Tracking

1. INTRODUCTION

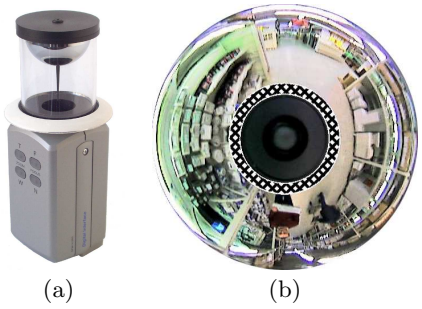
The ability to localize and track users is essential for every service robot which is supposed to serve people in special domains of everyday life. In case the user is inattentive or even absent, the robot should be able to interrupt the normal communication cycle or its current service and take appropriate actions. Moreover, it is desirable for the user to get a feedback of the tracking status from the robot, so she always knows whether the robot still pays attention to her or not. We develop our service robot PERSES, see figure 1, for deployment as shopping assistant in a home store [1]. The task of the robot is to get in contact with potential users and to guide them through the market area to products of interest or alternatively making contact to a qualified salesman through a wireless video conference system. Obviously, the users can not be introduced to the operation of the robot, so the interface has to be as natural, intuitive and appealing as possible. In experiments it turned out, that most customers are diffident when the robot approaches them, so it is not likely that a customer starts an interaction by herself. Thus, it is necessary for the robot to take the initiative and contact potential customers. Therefore, the robot looks for persons in its surroundings with its omnidirectional camera. In this way a rather large region can be processed thoroughly but with rather low resolution. However, this user position is very hypothetic, since it is extracted from skin-color and range measurements which are by no means person specific. To solve this problem, we decided to deploy a second processing step, using a high resolution image captured with one of the cameras mounted on the head of the robot.

Figure 1: The mobile robot PERSES (B21 from RWI IS Robotics) is equipped with an omnidirectional camera, two layers of sonar sensors, a touch display and a robotic face mounted on a pan-tilt-unit. The face can be used to express feelings like happiness, sadness, and anger.

The used vision based methods for face detection and tracking depend highly on constant illumination conditions. Since this constraint is not met in most real-world applications, we developed an automatic white balance algorithm for use with the omnidirectional camera.

Section 2 describes the user tracking system and section 3 describes how the tracking results are used to control the orientation of the robotic face in order to give the user an impression of dealing with an attentive dialog partner. Section 4 presents the results of an evaluation of different face detection systems.

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.



(a) (b)

Figure 2: (a) Camera with omnidirectional objective and coated aluminum ring. (b) Image taken with this camera, where the white reference appears near the center region of the image (marked with a checkered pattern). This region is not interesting in the context of user detection and tracking since it corresponds to the floor just around the robot.

2. TRACKING SALIENT REGIONS

First of all, the robot scans the market for potential customers. Therefore, it uses a saliency system which estimates the likelihood of the presence of a person in the robots surroundings and tracks this hypothesis over time. To achieve the necessary robustness, it uses a vision and a sonar based saliency system.

2.1 Vision Based Saliency

The basis of the vision based saliency system is the CONDENSATION algorithm [4]. The task of calculating the probability of the presence of a face 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 panoramic images from an omnidirectional color camera and uses a simple and fast skin color detection method to calculate the weights of its samples [7]. The center of the resulting distribution of samples is taken as hypothesis for the position of a user's face.

To deal with the problem of varying illumination conditions, which is crucial for skin color detection in realistic application scenarios, the omnidirectional camera was equipped with a coated aluminum ring serving as white reference. Figure 2 shows an image of the camera with omnidirectional objective and white reference and an image captured with this camera containing the white reference on the inner radius. The surface of the white reference ring is not horizontally oriented and flat, but shows a slight convex curvature so that light coming from the side is taken into account, too. The automatic white balance uses the facility of the digital camera (SONY DFW VL500) to set white balance parameters for U and V (YUV color space). We calculate the mean values for R, G and B from all pixels within the reference and transform these mean values to the YUV color space. From the difference of U and V from the target values $U = 0$ and $V = 0$ two discrete PID-controllers control the white balance of the camera. In addition, the mean Y value is used to control the iris of the digital camera, such that a constant brightness is achieved. The effect of the color calibration on the skin color classification is shown in figure 3.

Figure 3: A few images from the beginning of a sequence are shown together with the corresponding results from the skin color classification. The intensity of the output from the skin color detector rises from the first to the last image in the sequence because of the effects of the automatic white balance.

2.2 Sonar Based Saliency

The task of the sonar based saliency system is to assess the distance in every direction c around the robot. From the preprocessed sonar measurements a weighting factor $W_{sonar}(c)$ is calculated, which is 1 for objects in the immediate vicinity of the robot and decreases linearly to 0 with increasing distance between robot and object until some maximum distance [7]. Thus, the position of the maximum in the resulting weighting vector corresponds to the nearest object.

2.3 Sensor Fusion

Since the sonar scan as well as the image constitute a 360° description of the robot's surroundings, it is possible to assign a sonar weight at position c in the scan to each position $\vec{x}$ in the image, see figure 4. This way, the sonar vector can be used to modulate the weights of the CONDENSATION samples, see section 2.1. Thus, only those samples get a high weight, that are supported by a skin colored image pixel 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 saliency system eventually die out. As long as there is no region tracked, the sample distribution is initialized to places with high sonar weights at regular intervals. That means, the samples are placed on nearby objects, to check if they are skin colored or not. If so, the distribution concentrates on this position, if not, it diverges due to the stochastic movement of the samples, see figure 5. The person coming closest to the robot will initially attract its attention. However, once a person is tracked, the samples of the CONDENSATION algorithm are concentrated on her face, so they can not be distracted from her by another person, except they stand very close to each other. Due to the multimodal nature of our approach, the color of the background does not matter, as long as the user stays close to the robot.

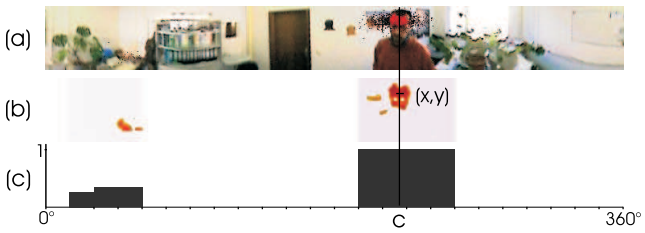
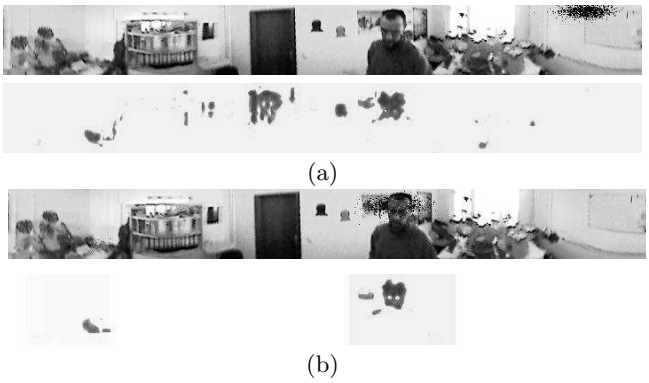


Figure 4: Sensor fusion. (a) Panoramic image with CONDENSATION samples. (b) Color classification modulated by sonar weights (c) Weighting factors calculated from sonar data. $\vec{x} = (x, y)$ is a position in the panoramic image and c is a position in the vector of sonar weights.



(a)

(b)

Figure 5: Result of the fusion of sonar data with vision based tracking. (a) Pure vision based (skin color) saliency. Besides the face of the user, the skin color detector assigns high values also to the door and some other objects. The sample distribution is initially placed at an arbitrary position. The variance of the distribution would increase due to the stochastic movement of the samples until some skin colored region is caught. This might be the face, but it could be the door as well. (b) As soon as the sonar based saliency system comes into play, most of the objects besides the face disappear and the sample distribution immediately concentrates on the face of the person.

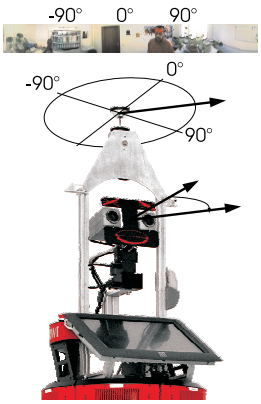


Figure 6: Upper part of our mobile service robot PERSES with omnidirectional camera and face with two cameras mounted on a pan-tilt-unit. Determination of the pan angle ϕ_{pan} from the position of the user's face in the panoramic image.

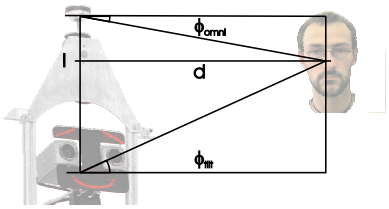


Figure 7: Calculation of the tilt angle ϕ_{tilt}.

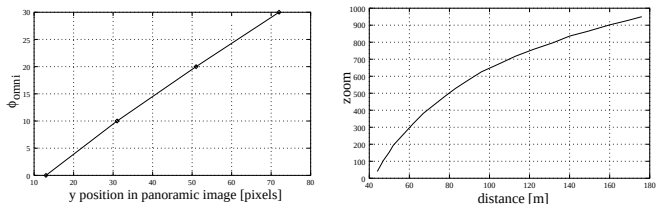
3. HEAD MOVEMENT

The saliency system estimates the position (x, y) of a user's face in the omnidirectional view. From this data it is possible to calculate the $3D$-position of the user's face in the robot's surroundings and finally the rotation angles of the pan-tilt-unit, which serves as neck for the head. The horizontal angle ϕ_{pan} corresponds to the x-position of the tracking result, see figure 6, while the vertical angle ϕ_{tilt} depends both on the y-position of the tracking result in the image and the distance to the corresponding object d, which can be determined by the sonar measurements, see equation 1 and figure 7.

$$\phi_{tilt} = \arctan \frac{l - d \cdot \tan \phi_{omni}}{d} \tag{1}$$

l is the distance between the omnidirectional and frontal cameras and d is the distance between camera and user, determined from the sonar measurements. The relationship between the y-position of the face in the panoramic image and the angle ϕ_{omni} depends on the shape of the mirror in the omnidirectional objective and was determined experimentally, see figure 8(a).

The angles ϕ_{pan} and ϕ_{tilt} are used to orient the robot's face towards the estimated position of the user giving her a direct feedback of the robot's attention during communication. If the tracking system looses the person, PERSES looks straight ahead and sad by means of its robotic face. On the other hand, if a person is tracked with a high probability, PERSES is happy and looks at the user continuously. This



(a) (b)

Figure 8: (a) Relationship between y-position of the face in the panoramic image and the angle ϕ_{omni}. (b) Dependance of zoom value from the distance between camera and object. The objective of this active zoom is to map a human face approximately with the same size for all distances.

head movement gives the user an impression of an attentive communication partner and can be accompanied by a rotation of the robot's body in case the angle ϕ_{pan} gets to large.

Moreover, the frontally aligned cameras on the robotic head can now be used to verify the hypotheses about the position of the user's face. This is useful since it can not be guaranteed, that the saliency system does not respond to some other skin colored region in the robot's proximity. In our home depot scenario, these can be tins with dye standing in a shelf just where the robot passes by. Thus, the robot takes a close look in the direction of its hypothesis with its frontally aligned cameras.

Since we are only searching for human faces and can determine the distance d to the object of interest, we can adjust the camera zoom such, that an average face would fill the whole image, see figure 8(b). Thus, we are able to capture a high resolution image of the users face, even if she stands relatively far from the robot. Additionally we are able to save processing time, since we don't need to use multiple resolutions for processing the high resolution image.

4. FACE DETECTION

The acquired high resolution images are used to verify the hypothesis with the face detection system proposed by Viola and Jones [6]. Among multiple re-implementations of face detectors made in our lab, the one presented by Viola and Jones turned out to be the fastest while, at the same time, it has the highest detection rates and very low false positive rates. First results of a comparative study of these face detectors are shown in table 1. We tested the face detector presented by Rowley [5], a Cascade-Correlation-Network [2], a system based on edge orientation matching presented by Fröba and Küblbeck [3], and the system from Viola and Jones [6].

Our test database was recorded in a home store under varying illumination conditions and consists of 118 images. A description of the experiments and the database can be fount at `http://cortex.informatik.tu-ilmenau.de/~wilhelm/research.html`.

face detector	detection rate	false positive
Cascade Correlation	18.26	0.00001379
Rowley	37.39	0.00472281
Edge Orientation	46.96	0.00091035
Viola and Jones	56.52	0.00024552

Table 1: Results of various face detectors on our test set. Each detector used a multi resolution image pyramid with 11 layers and a scaling factor of 0.707. The false positive rate is the number of false positives divided by the number of all hypothetical positions in the test set. An image has $640 \times 480 = 307200$ pixels. Thus the face detector from Viola and Jones detects less than 1 false positive per image.

5. SUMMARY AND OUTLOOK

We presented a person detection system consisting of two components: a fast saliency component, which uses skin color and sonar data to estimate and track the most probable position of a potential user. By means of an automatic color calibration, the skin color detector works independent from changes in illumination. In the second step, a high resolution image is checked for the presence of a human face.

Besides using the face detector for verification, we want to extract further information from the high resolution image of the user, such as identity, age, gender and mimic. This information is supposed to be used to adapt the man-machine-interface to the special needs of the current user.

6. REFERENCES

[1] Böhme, H.-J., Wilhelm, T., Key, J., Schauer, C., Schröter, C,, Gross, H.-M., and Hempel, T. An approach to multi-modal human-machine interaction for intelligent service robots. *Robotics and Autonomous Systems*, 44:83–96, 2003.

[2] Fahlman, S. E. and Lebiere, C. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems 2*, pages 524–532, 1990.

[3] Fröba, B. and Küblbeck C. Face detection and tracking using edge orientation information. *SPIE Visual Communications and Image Processing*, pages 583–594, 2001.

[4] Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.

[5] Rowley, H. A., Baluja, S., and Kanade, T. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[6] Viola, P. and Jones, M. Robust real-time object detection. In *Second International Workshop on Statistical and Computational Theories of Vision*, 2001.

[7] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. Looking Closer. *Proceedings of the European Conference on Mobile Robots*, 2003, to appear.