Proc. Eur. Conf. on Mobile Robots (ECMR 2003), Warsaw, Poland, pp. 65-70 Zturek Press Warschau 2003

Looking Closer

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. This paper describes a user detection system which employs a saliency system delivering a rough and fast estimate of the position of a potential user. This saliency system consists of a vision (skin color) and a sonar based component, which are combined to make the estimate more reliable. To make the skin color detection robust under varying illumination conditions, it is supplied with an automatic white balance algorithm. Finally, the hypothesis of the users position is verified with a highly specific face detector.

1 Introduction

Localization and tracking of users is a basic working task for every service robot which is supposed to serve people in special domains of everyday life. We develop our service robot PERSES for deployment in a home store. The task is to actively contact potential users and to guide them through the market area. Therefore the robot needs to detect users in a wide operation area while at the same time it is desirable to get information like identity, gender and age of the user to adapt the dialog management accordingly. These two tasks are more or less oppositional: the first one should analyze the complete surroundings of the robot, which can only be achieved by a panoramic image with a field of view of 360° but low resolution, while for the second one a high resolution image of the users face is needed. Thus we decided to deploy a two step solution. First a saliency system gives a rough estimate of the users position. This system consists of a sonar and a vision based tracking component. The panoramic image used by the vision based tracking is automatically white balanced to cope with varying illumination conditions. Second, the robots active vision head is turned to look towards this estimate of the position of the users face. This is done for two reasons: to verify the presence of a person by use of the face detection system introduced by Viola and Jones [7], which analyses a high resolution image taken from one of the front cameras, and to give the user a continuous feedback, which expresses the robots attention during the communication process.

Other known person detection algorithms rely solely on a single sensor system such as a laser scanner in [6]. Such systems might be sufficient for detecting the presence of a person in the robots sourroundings but they can not make any statement of whether this person is facing the robot as an expression of his will for interaction. In our opinion, visual cues are indispensable when an intuitive human-machine-interaction is to be achieved. On the other hand, in [5] only visual cues are used, which together with the lack of a white balance algorithm makes the system fragile for changing illumination conditions. Moreover the saliency system uses only frontally aligned cameras and thus only has a very limited field of view.

2 Saliency System

The saliency system estimates the likelihood of the presence of a person in the robots surroundings and tracks this hypothesis over time.

2.1 Vision Based Saliency

The basis of the vision based saliency system is the condensation algorithm [3]. 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. Compared to a panoramic image

with 720×106 pixels calculating the feature extraction only for 1000 samples yields a reduction to merely 1.31%. The center of the resulting distribution of samples is taken as hypothesis for the position of a users face.

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 [5]. To prevent the color model from getting holey because of insufficient training data, there is a small Gaussian placed around each skin color pixel, see figure 1(a). The color detection can be calculated very fast because it consists only of a lookup operation, but on the other hand it is highly dependent on illumination color and variations in hue.

(a) (b) (c)

Fig. 1.: (a) Skin color lookup table in the dichromatic r-g-color space. (b) White reference in-between camera and objective. (c) Image taken with this camera, where the white reference appears near the center region of the image (marked with a checkered pattern). Since this region corresponds to the floor just around the robot, it is not interesting in the context of user tracking.

Automatic Color Calibration To deal with the mentioned problem of varying illumination conditions, we developed an automatic white balance algorithm operating on the images from the omnidirectional camera. For this purpose the camera was equipped with a coated aluminum ring serving as white reference. Figure 1 shows an image of the camera with omnidirectional objective and white reference and an image grabbed with this camera containing the white reference on an inner radius. The surface of the white reference ring is not horizontally and flat, but shows a slight convex curvature so that also light coming from the side is taken into account.

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 white 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 separate discrete PID-controllers control the white balance of the camera. Besides that, the mean Y value is used to control the iris of the digital camera, such that a constant brightness is achieved, see figure 2. The effect of the color calibration on the skin color classification is shown in figure 3.

2.2 Sonar Based Saliency

The task of the sonar based saliency system is to measure the distance in every direction around the robot. 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:

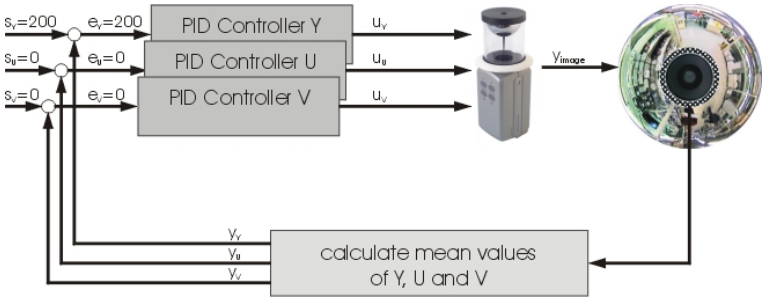


Fig. 2.: Color calibration algorithm.

(a) (b)

Fig. 3.: Automatic white balance on images from the omnidirectional camera. (a) Some images from the beginning of a sequence. (b) 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. It takes about 3 seconds to control the values for U and V to zero.

1. replacement of invalid measurements: distances larger than $22,5m$ are considered 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 $W^{(c)}_{sonar} = 1 - d^{(c)}_{sonar}/d_{max}$, where $d^{(c)}_{sonar}$ 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. The robot could be aligned towards this maximum, preconditioned that it is in accordance with the user hypothesis provided by the visual tracking system [8]. However in this setup we use an approach, where the sonar data is used to support the visual tracking procedure.

2.3 Sensor Fusion

Since the sonar scan as well as the image constitute a 360° description of the robots surroundings, it is possible to assign a scan measurement at position c in the scan to each position $\mathbf{x}$ in the image, see figure 4. This way, the sonar vector can be used to modulate the sample weighting in the condensation algorithm, see equation 1. Only those samples get a high weight $W^{(i)}_{sample}(\mathbf{x})$, 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, see figure 5.

$$W^{(i)}_{sample}(\mathbf{x}) = W^{(i)}_{skincolor}(\mathbf{x}) W_{sonar}(c) \tag{1}$$

There are other heuristics that try to eliminate skin colored image segments not stemming from faces (false positives), which depend on the size and shape of these regions [5][**?**]. In such approaches only those segments are allowed, that have roughly the shape and size of a human face. The problem with this approach is that when a face appears in front of some larger skin colored region, the whole area is wiped out and the

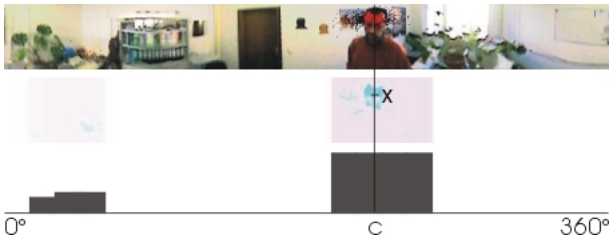
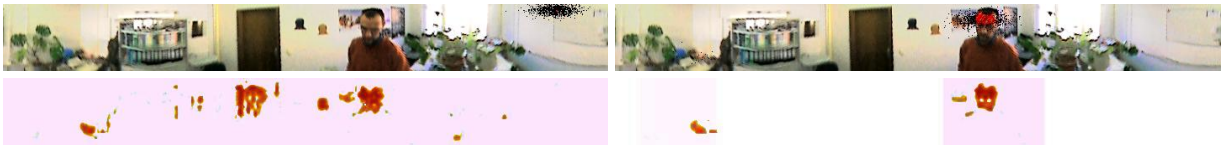


Fig. 4.: Sensor fusion. From top: panoramic image with condensation samples, color classification modulated by sonar weights, and sonar weights.



(a) (b)

Fig. 5.: Result of the fusion of sonar data with vision based tracking. The pure vision based (skin color) saliency is shown in (a). Besides the face of the user, there are assigned high values also to the door and some other objects. As soon as the sonar based saliency system comes into play, most of the objects besides the face disappear and the sample distribution immediately clings to the face of the person (b).

tracker looses the face. The presented multimodal approach can cope with this situation. As long as the user stays near the robot, the color in the background does not matter. The presented approach is intuitive in that the person coming closest to the robot will initially get his attention. However, once a person is tracked, the samples of the condensation algorithm are concentrated on his face, so they can not be distracted from him by another person, except they stand at close quarters.

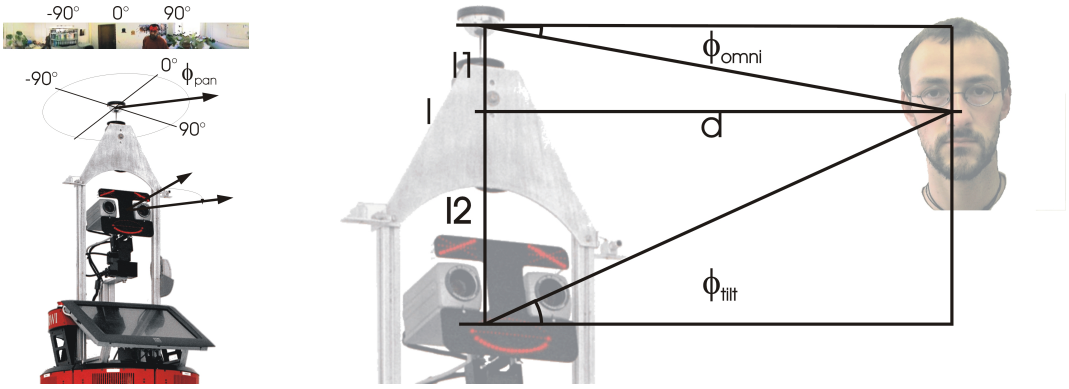
3 Looking Closer

By means of the automatic white balance and in combination with the sonar based saliency our system is already highly specific for skin colored image regions stemming from objects near the robot. But still it can not be guaranteed, that it does not respond to some other skin colored image region. In our home store scenario these can be tins with dye standing in a shelf just where the robot passes by. Thus before contacting a potential user (e.g. by speech output), the robot takes a close look in the direction of its hypothesis with its frontally aligned cameras.

Therefore the rotation angles of the pan-tilt-unit, which serves as neck for the head, need to be calculated. The horizontal angle can be taken from the output of the condensation algorithm directly, see figure 6(a), while the vertical angle ϕ_{tilt} depends both on the vertical position of the target in the image and the distance d to the corresponding object, which can be determined by the sonar measurements, see equation 2 and figure 6(b). The angle ϕ_{omni} is a function of the vertical position of the target in the image and depends on the shape of the used mirror in the omnidirectional objective. This function was determined experimentally and is shown in figure 7(a).

$$\phi_{tilt} = \arctan \frac{l - d \cdot \tan \phi_{omni}}{d} \tag{2}$$

The angles ϕ_{pan} and ϕ_{tilt} are used to orient the face towards the estimated position of the user giving her a direct feedback of the robots 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 head movement gives the user an impression of an attentive communication partner and can be accompanied by a rotation of the robots body in case the angle ϕ_{pan} gets to big.



(a) (b)

Fig. 6.: (a) Upper part of our mobile service robot PERSES with omnidirectional camera and face with two cameras mounted on a pan-tilt-unit. The robot grabs high resolution images of the user, which are used to verify the hypothesis of the saliency system. The angle ϕ_{pan} corresponds to the position of the face in the omnidirectional image. (b) Geometrical illustration of the angles ϕ_{omni} and ϕ_{tilt}.

Since we are only searching for human faces and can determine the distance to the object of interest from the sonar measurements, we can set the zoom of the frontally aligned camera such, that an average face would fill the whole image. Figure 7(b) shows the dependence of the camera zoom from the distance between camera and object. In the same way as the zoom the focus can be set according to the distance (the cameras do not have autofocus). Thus we can asure to get a high resolution image of the user independent of his distance to the robot. This image is used to verify the hypothesis with the face detection system from Viola and Jones [7]. Among multiple reimplementations of face detectors the one presented by Viola and Jones appeared to be the fastest while at the same time it has high 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 [4], a Cascade-Correlation-Network [1], a system based on edge orientation matching presented by Fröba and Küblbeck [2] and the system from Viola and Jones [7] on image data from the home store.

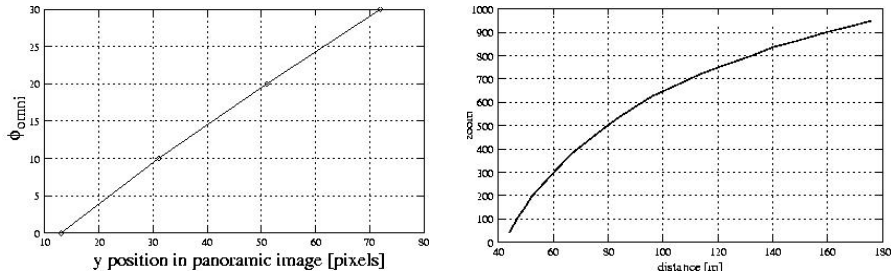
face detector	**detection rate [%]**	**false positive rate**
Rowley	37.39	20635
Cascade Correlation	39.13	5850
Edge Orientation	46.96	109847
Viola and Jones	56.52	407299

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 hypothetical positions in the test set divided by the number of all false positives. Since an image has 307200 pixels, the face detector from Viola and Jones detects less than 1 false positive per image.

4 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 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.

Our experiments were carried out using the sonar data for measuring the distance for controlling the camera zoom. Because of the high noise level in these measurements, the results are not satisfying, with the



(a) (b)

Fig. 7.: (a) Relationship between the vertical 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.

zoom changing often due to false measurements. A possibility to solve this problem arises from the topology of the robots head, consisting of two parallel oriented cameras, which lend themselves perfectly for the use of a stereo algorithm. To avoid the extra cost of a stereo algorithm, we also want to test a so called autofocus sensor mounted on the robots head. This type of sensor is widely used in camera technology and can thus be considered both cheap and reliable.

To cope with multiple users we use a number of condensation trackers that exclusively track skin colored image regions. One of those regions is used as user hypotheses as long as it shows a sufficient amount of skin color and as long it is containing a face as evaluated by the face detector. If one of these conditions is not met anymore, the next tracked skin colored image region is used as hypotheses and the face is turned in that direction. This approach results in an behavior, where the robot scans all the salient skin color regions sequentially until one of them contains a face of a potiential user. However this is still to be considered as work in progress.

Besides using the face detector for verification, we want to extract further information from the high resolution image of the user. These include identity, age, and gender, and hopefully could be used to adapt the man-machine-interface to the special needs of the current user.

References

1. Fahlman, S. E. and Lebiere, C. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems 2*, pages 524–532, 1990.
2. Fröba, B. and Küblbeck C. Face detection and tracking using edge orientation information. *SPIE Visual Communications and Image Processing*, pages 583–594, 2001.
3. Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.
4. 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.
5. S. Feyrer and A. Zell. Detection, tracking, and pursuit of humans with an autonomous mobile robot. In *International Conference on Intelligent Robots and Systems (IROS ´99)*, pages 864–869, 1999.
6. Schulz, D., Burgard, W., Fox, D., and Cremers, A.B. Tracking Multiple Moving Targets with a Mobile Robot using Particle Filters and Statistical Data Association. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2001.
7. Viola, P. and Jones, M. Robust real-time object detection. In *Second International Workshop on Statistical and Computational Theories of Vision*, 2001.
8. Wilhelm, T., Böhme, H.-J., and Gross, H.-M. Sensor fusion for vision and sonar based people tracking on a mobile service robot. In *Proc. of the Int. Workshop on Dynamic Perception 2002, Bochum*, pages 315–320. IOS Press, infix, 2002.

This article was processed using
the TEX macro package with ECMR2003 style