Robotics and Autonomous Systems, 48 (2004), pp.31-40, Elsevier Science

A Multi-Modal System for Tracking and Analyzing Faces on a Mobile Robot

T. Wilhelm, H.-J. Böhme, H.-M. Gross

Ilmenau Technical University, Department of Neuroinformatics, P.O.Box 100565, 98684 Ilmenau, Germany

Abstract

This paper describes a user detection system which employs a saliency system working on an omnidirectional camera delivering a rough and fast estimate of the position of a potential user. It 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. The active vision head looks continously in the direction of the salient region. Thus, a high resolution image can be grabbed and analyzed with a face detector.

Key words: man-machine-interface, user tracking, color correction

1 Introduction

Localizing and tracking 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, see Fig. 1, for deployment in a home store [5]. The task is to actively contact potential users and to guide them as needed 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 the 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 be achieved by

Email addresses: Torsten.Wilhelm@tu-ilmenau.de (T. Wilhelm), Hans-Joachim.Boehme@tu-ilmenau.de (H.-J. Böhme), Horst-Michael.Gross@tu-ilmenau.de (H.-M. Gross).

Fig. 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.

using a panoramic image with a 360° field of view and low resolution, while for the second one a high resolution image of the user's face is needed.

Thus we decided to deploy a two step solution. First, a saliency system gives a rough estimate of the user's 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 robot's active vision head is turned to look towards this estimate of the position of the user's face. This is done for two reasons: to verify the presence of a person with the face detection system introduced by Viola and Jones [13] and to give the user a continuous feedback, which expresses the robot's attention during the communication process.

There are other known person detection systems which rely solely on a single sensor system such as laser scanner in [12]. Such systems might be sufficient for detecting the presence of a person in the robot's surroundings, but they cannot make any statement as to 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 [10] 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. Other multi-modal user detection systems have proven to be reliable and suitable for real-world applications. However, often expensive hardware, e.g. laser scanners are used [11], which seems to be a real handicap when it comes to serial production of robots for everyday use. Our research focuses on developing methods that work reliably with cheap and less accurate sensors, e.g. sonar sensors and cameras [5].

2 Saliency System

The saliency system estimates the likelihood of the presence of a person in the robot's surroundings and tracks this hypothesis over time. It is composed of two components, a vision and a sonar based saliency system.

2.1 Vision Based Saliency

2.1.1 Skin Color

A widely used method for finding faces in images is skin color classification. It lends itself for the use on mobile systems, since it is independent from ego-motion of the camera system. To represent skin color, we use the dichromatic r-g-color space ($r = R/(R+G+B)$, $g = G/(R+G+B)$, which is normalized in brightness and thus is widely independent from variations in luminance. This color space is well suited for representing skin color for a wide spectrum of different illumination conditions [15]. In principle, it is possible to use any color space which decorrelates brightness and color information.

The skin color model consists of a look up table with manually classified skin color pixels in the r-g-color space [10]. To avoid an unnecessary reduction of accuracy and an increase of processing time, the color model was not approximated by a Gaussian distribution as in [6]. Thus, it is necessary to ensure that the training data is sufficient, so the skin color model does not become holey. The resulting color model is depicted in Fig. 2(a). The skin color detector gives a value $w_{skin}(\underline{x})$ for the pixel at position $\underline{x}$ in the image.

2.1.2 Automatic Color Calibration

Despite the advantages of the skin color detection, it works only satisfactory as long as the illumination conditions are sufficiently similar when recording the training data and when using the model, which cannot be guaranteed when operating in an environment as diverse as a home store.

One solution to this problem is the continuous adaptation of the color model to the current illumination conditions [6][15]. To our experience, this approach is problematic, since the tracked target must be detected in every time step with a very high precision. When only slight positional errors occur, such adaptive models tend to drift away from the appropriate description of the target region. Some approaches try to prevent strong drifts by narrowing down the freedom of movement of the adaptive model by use of a general skin color model, recorded under a variety of different illumination conditions [2][8].

(a) (b) (c)

Fig. 2. (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 the occluded region corresponds to the floor just around the robot, it is not of interest in the context of user tracking.

There are also efforts that try to stabilize the localization of skin color by applying additional features. In [2] a face detector is used and the adaptation is only carried out, when a positive face detection result within the skin colored region is given. Such approaches reduce the mentioned problem but do not solve it in general, e.g. every face detector has a false-positive-rate above zero. Moreover, the use of additional features increases the complexity, which has a negative impact on the continuous tracking.

Another way of reducing the influence of illumination is to preprocess the image with color constancy algorithms. The task is to construct an image, such as it would look under a standard illumination from the given image grabbed under an unknown illumination only. In [4], different color constancy algorithms were tested on their suitability for color based object recognition. Despite the fact that the recognition rates could be improved significantly, a robust recognition under changing illumination was not possible. The problem here is that without knowledge of the current illumination any adaptation is bound to fail as long as no presumptions can be made on the observed scene.

Thus, to deal with the 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 to serve as white reference. Figure 2(b) shows the camera with omnidirectional mirror and white reference, and Fig. 2(c) shows an image grabbed with this camera containing the white reference on an inner radius. The surface of the white reference ring is not horizontal and flat, but has a slight convex curvature so that light coming from the side is also taken into account.

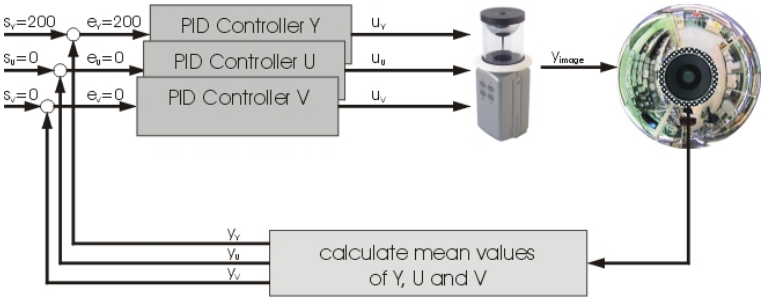


Fig. 3. The color calibration algorithm consists of a closed control loop with three discrete PID-controllers for Y, U and V of the camera-specific YUV-color space.

Fig. 4. Automatic white balance on images from the omnidirectional camera. A number of images from the beginning of a sequence is 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. It takes about 3 seconds to control the values for U and V to zero.

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 calculate the gain factors for the U and V channels of the white balance of the camera [14]. Besides that, the mean Y value is used to control the iris of the digital camera, such that a constant brightness (about 80% of maximum) is achieved, see Fig. 3. The effect of the color calibration on the skin color classification is shown in Fig. 4.

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:

Invalid measurements, i.e. distances larger than $22.5m$, are replaced by the previous measurements. A spatial low pass filtering of adjacent measurements and a temporal low pass filtering of successive measurements is applied to reduce the influence of noise.

We calculate a weighting factor for each direction c 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 ($2.0m$). For distances larger than d_{max} the weight is set to zero. The position of the maximum in the resulting weighting vector corresponds to the nearest object.

2.3 Sensor Fusion

2.3.1 Condensation Tracking

The basis of the saliency system is the condensation algorithm [7]. The task of calculating the probability of the presence of a face for every pixel and tracking the resulting density function over time is solved by an approximation of the density function $p(\underline{x}_t)$ by a relatively small number of samples $\underline{s}_t^{(i)}$:

$$p(\underline{x}_t) \propto \left\{ \underline{s}_t^{(i)} = \left\langle \underline{x}_t^{(i)}, w_t^{(i)} \right\rangle | i = 1, .., N \right\} \tag{1}$$

where each sample $\underline{s}_t^{(i)}$ has a position $\underline{x}_t^{(i)}$ and a weight $w_t^{(i)}$.

According to [7] the update formula of the recursive filter is as follows:

We begin with a sample set s representing the posterior density $p(\underline{x}_{t-1}|Y_{t-1})$ from the previous time step, were Y_{t-1} is the history of measurements $\{y_1, ..., y_t\}$. We propagate s according to a stochastic motion model, i.e. a gaussian distribution, accounting for unforeseen movements of the person and obtain the new sample set s' representing the prior density $p(\underline{x}_t|Y_{t-1})$. Then we apply factored sampling, i.e. we assign the new sample weights $w_t^{(i)}$ according to

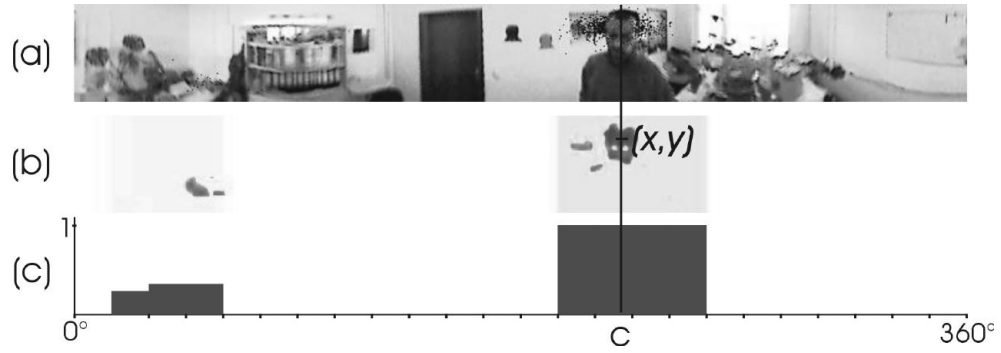


Fig. 5. Sensor fusion. (a) Panoramic image with condensation samples. (b) Color classification modulated by sonar weights. (c) Weighting factors calculated from sonar data. $\underline{x} = (x, y)$ is a position in the panoramic image, and c is a position in the vector of sonar weights.

the measurements from the saliency system in the current time step (see next section) and draw samples from s' with probability $w_t^{(i)}$. Sample set s'' then represents the posterior density $p(\underline{x}_t|Y_t)$.

Compared to a panoramic image with 76320 pixels the condensation algorithm calculates the feature extraction for only 500 samples and thus yields a reduction of computational cost to merely 0.655% while being able to track arbitrary multi-modal distributions. The center of the resulting distribution of samples is taken as hypothesis for the position of a user's face.

2.3.2 Sample Weights

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 $w_{sonar,t}(c)$ at position c in the scan to each position $\underline{x}$ in the image, see Fig. 5. This way, the sonar weights can be used to modulate the weights of the skin color detector $w_t^{(i)}(\underline{x}) = w_{skin,t}^{(i)}(\underline{x}) w_{sonar,t}(c)$. 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 whether 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 Fig. 6. 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 his face, so they cannot be distracted from him by another person, except they are standing very close to each other.

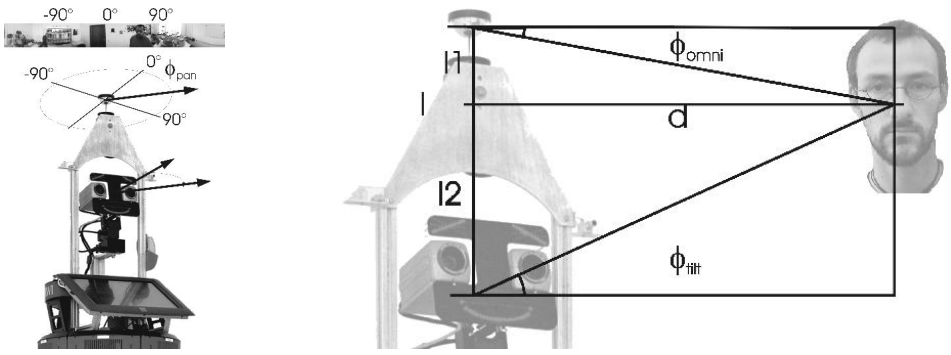
Fig. 6. Result of the fusion of sonar data with vision based tracking. (a) Pure vision based (skin color) saliency. Besides the face of the user, the skin color detector assigns high values also to the door and some other objects. The sample distribution is initially placed at an arbitrary position. The variance of the distribution would increase due to the stochastic movement of the samples until some skin colored region is detected. 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 from the color detection output and the sample distribution immediately concentrates on the face of the person.

There are other heuristics that try to eliminate skin colored image segments not stemming from faces (false positives), which evaluate the size and shape of these regions [10][6]. In such approaches only those segments are allowed, that have roughly the shape and size of a human face. The problem with these approaches is that when a face appears in front of a larger skin colored region in the background, the whole area is wiped out and the tracker loses the face. Due to the multi-modal nature of our approach, the color of the background does not matter as long as the user stays close to the robot.

3 Head Movement

In combination with the automatic white balance, our saliency system is already highly specific for skin colored image regions stemming from objects close to the robot. Still, it can not be guaranteed that it does not respond to other skin colored image regions that do not belong to users. 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 a neck for the head, need to be calculated. The horizontal angle can be taken from the output of the condensation algorithm directly, see Fig. 7(a), while the vertical



(a) (b)

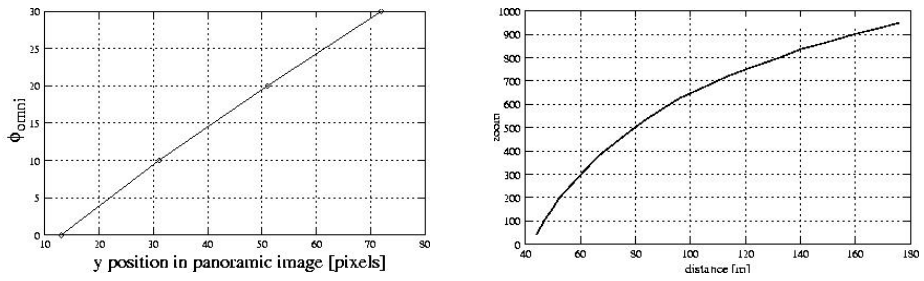
Fig. 7. (a) Upper part of our mobile service robot PERSES with omnidirectional camera and face with two cameras mounted on a pan-tilt-unit. The robot grabs high resolution images of the user, which are used to verify the hypothesis of the saliency system. The angle ϕ_{pan} corresponds to the position of the face in the omnidirectional image. (b) Geometrical illustration of the angles ϕ_{omni} and ϕ_{tilt}.

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 Fig. 7(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 Fig. 8(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's face giving him a direct feedback of the robot's attention during communication. If the tracking system loses the person, PERSES looks straight ahead and with a "sad" expression on its robotic face. On the other hand, if a person is found (i.e. the samples have high weights), PERSES generates a "happy" facial expression 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 too big.

Since we are only searching for human faces and we are able to 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 entire image. Figure 8(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 assure to get a high resolution image of the user's face independent of his distance to the robot up to about 2 meters.



(a) (b)

Fig. 8. (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 as prerequisite for subsequent processing steps.

Face detector	**Detection rate**	**False positive rate**
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. (In the approach from Viola and Jones the filters are resized instead of the image.) The false positive rate is the number of false positives divided by the number of all hypothetical positions in the test set.

4 Face Detection

The image obtained from the frontal camera is used to verify the hypothesis with the face detection system from Viola and Jones [13]. Among multiple implementations of face detectors, the one proposed by Viola and Jones appeared to be the fastest one, 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 [9], a Cascade-Correlation-Network [1], a system based on edge orientation matching presented by Fröba and Küblbeck [3] and the system from Viola and Jones [13] on image data from our home store. It should be mentioned that the used implementations might differ from the original face detectors and that the parameters of the single detectors were not tested systematically.

5 Multi Target Tracking

Theoretically, a condensation tracker is able to track multi-variate density distributions. However, as soon as the size of the skin color regions is not balanced, the sample distribution tends to collapse and track the largest region only. Thus, to be able to track more than one face, we use multiple condensation trackers that track a skin colored image region each. One of those regions is used as the current user hypothesis which is followed by the frontal camera as described above. If the underlying skin color probability drops below a threshold or there is no face detected for a certain amount of time, the current user hypothesis is switched to another condensation tracker. A tracker is erased when the underlying skin color probability is too low or when it gets too close to another tracker. A new tracker is created when there are skin color regions not currently tracked. This approach results in a behavior where the robot scans all the salient skin color regions sequentially until one of them contains the face of a potential user and then follows with the frontal camera as long as a face is found.

6 Experimental Results

Figure 9 shows a sequence with panoramic images and corresponding images from the face detector recorded in the home store. Although the person moves in front of the robot and changes his distance to the robot and there are other persons appearing in the robot's surroundings, the robot keeps tracking its current user. Due to the adaptation of the camera zoom, the face in the image from the frontal camera is mapped with approximately the same size and resolution over the sequence.

7 Summary and Outlook

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

Besides using the face detector for verification, we want to extract further information from the high resolution image of the user. This includes identity,

Fig. 9. Sequence of panoramic images and corresponding images from the face detector. Every 10th image from the sequence is shown. The centers of gravity of the sample distributions are marked with a white cross in the panoramic image. In frame 1, there are two targets, where the one on the right side is on a wooden shelf. However, because it is to far from the robot, it disappears in the second frame after the sonar based tracking was switched on. Even though the person moves away from the robot in frames number 2, 3, and 10, the face in the image from the frontally aligned camera appears with approximately the same size over the whole sequence. The face detector had false positive detections in frames 2 and 9. In frame 5 and 8, other people enter the robot's surroundings and get immediately targeted by the tracker. However, the robot maintains its focus on the current user although the other persons stand closer to the robot, see frame 9,10 and 11.

age, and gender, and hopefully could be used to adapt the man-machine interface to the needs of the current user. In our current work, we implement and analyze methods to localize facial feature points like eyes and the nose and a PCA/ICA based analysis of facial expressions and gender of users.

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] Fritsch, J., Lang, S., Kleinehagenbrock, M., Fink, G.A., and Sagerer, G. Improving adaptive skin color segmentation by incorporationg results from face detection. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN), Berlin, Germany*, pages 337–343, 2002.

[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] Funt, B., Barnard K., and Martin, L. Is machine colour constancy good enough? *Lecture Notes in Computer Science*, 1406:445–459, 1998.

[5] Gross, H.-M., Boehme, H.-J., Key, J, and Wilhelm, T. The perses project - a vision-based interactive mobile shopping assistant. *Künstliche Intelligenz*, 4:34–36, 2000.

[6] H.-J. Böhme, U.-D. Braumann, A. Brakensiek, A. Corradini, M. Krabbes, and Gross, H.-M. User localisation for visually-based human-machine interaction. In *Proc. 1998 IEEE Int. Conf. on Face and Gesture Recognition, Nara, Japan*, pages 486–491, 1998.

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

[8] Jang, G.-J. and Kweon, I.-S. Robust object tracking using an adaptive color model. In *Proc. of the IEEE Int. Conference on Robotics and Automation*, pages 1677–1682, 2001.

[9] 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.

[10] 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.

[11] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G.A. Fink, and G. sagerer. Providing the basis for human-robot-interaction: A multi-modal

attention system for a mobile robot. In *Proc. 2003 Int. Conf. on Multimodal Interfaces, Vancouver, Canada*, pages 28–35, 2003.

[12] 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.

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

[14] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. Automatischer Weissabgleich für eine omnidirektionale Kamera. In *Proc. 9. Workshop für Farbbildverarbeitung, Esslingen*, pages 43–50. Schriftenreihe ZBS, 2003.

[15] Yang, J. and Waibel, A. Skin-color modeling and adaptation. *Lecture Notes in Computer Science*, 1352:687–694, 1998.