Active People Recognition using Thermal and Grey Images on a Mobile Security Robot

André Treptow
Department of Computer Science
University of Tuebingen
Tuebingen, Germany
treptow@informatik.uni-tuebingen.de

Grzegorz Cielniak and Tom Duckett
AASS, Department of Technology
University of Oerebro
Oerebro, Sweden
{grzegorz.cielniak, tom.duckett}@tech.oru.se

***Abstract*— In this paper we present a vision-based approach to detect, track and identify people on a mobile robot in real time. While most vision systems for tracking people on mobile robots use skin color information, we present an approach using thermal images and a fast contour model together with a Particle Filter. With this method a person can be detected independently from current light conditions and in situations were no skin color is visible (the person is not close or does not face the robot). Tracking in thermal images is used as an attention system to get an estimate of the position of a person. Based on this estimate we use a pan-tilt camera to zoom to the expected face region and apply a fast face tracker in combination with face recognition to identify the person.**

***Index Terms*— Robot vision, real-time people tracking, thermal images**

I. Introduction

Vision-based detection, tracking and identification of humans on mobile robots is a challenging task. The ability to interact with people in populated environments is important for robots that fulfill tasks in cooperation with humans (e.g., service robots, inspection tasks, surveillance). Recently, systems for human-robot interaction that are able to locate the position of a person facing the robot have been developed. However, these approaches assume that people are close to the robot and face toward it so that methods based on skin color and face detection can be applied: Wilhelm et al. [12] track regions in the image which have skin color and combine this information with sonar data to get an estimate of the position of a person that is close to the robot. In a second step they use a face detector to get the position of the face in the image. Barreto et al. [6] describe a human-robot interface that relies purely on a face detector in combination with face recognition based on PCA. Similar work can be found in [7] where a detected face region is tracked with skin color information. Lang et al. [11] combine several cues including sonar, laser scanner, sound localisation and color image processing.

The work presented here is part of a robotic security guard project, where one task for the mobile robot is to identify people in the building while patrolling. In this scenario the robot must be able to detect a person even from larger distances and it cannot be assumed that the person faces the direction of the robot. Therefore skin color cannot be used as a cue for the position of a person in the image. In this paper we address this problem and introduce a new method to detect and track a person in thermal images. This information is used to get a first estimate of the position of a person relative to the robot. While tracking a person in the thermal image, the robot tries to get closer to identify the person. Identification is performed using grey value images. Our experimental platform is an ActivMedia PeopleBot mobile robot that is equipped with several sensors including a thermal camera and a pan-tilt camera unit (see figure 1).

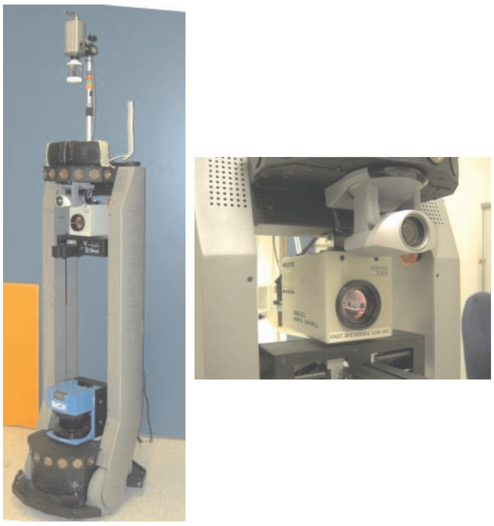

Fig. 1. ActivMedia Peoplebot, thermal camera (NEC Thermal Tracer TS7302) and pan tilt camera.

II. Method

Our approach to identify people in real time on a mobile robot is shown in figure 2. The system can be divided into 4 parts. First of all, the robots starts in the search mode where it tries to detect a person based on the information from the thermal camera. If a person is detected in the thermal image the robots drives toward the person while tracking. This part is the attention system where the robots tries to get a rough estimate of the person's position based on thermal images. If the robot is close to a person we use grey value images from the pan tilt camera to track the face. While tracking the face, images from the face tracker are fed into the recognition system to update an estimate of the identity of the person.

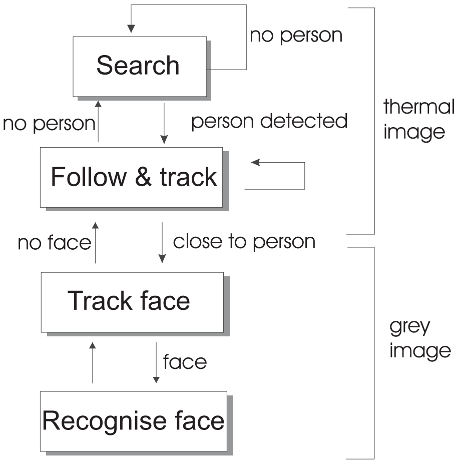

Fig. 2. Overview over the proposed system.

A. Tracking people in thermal images

The advantage of using sensor information from a thermal camera is that a person in the thermal image has a very distinctive profile so that the person can be clearly separated from the background. In figure 3 one can see that in the color image there is hardly any skin color visible if the person is further away, even though the person faces toward the camera. On the other hand one can easily detect the person in the same scene shown by the thermal image. However, apart from the work published in [3], where Cielniak and Duckett use image segmentation based on thresholding, noise filtering and morphological operations, there is hardly any published work on using thermal sensor information to detect humans on mobile robots until now. Infrared sensors have been applied to detect pedestrians in a driving assistance system: Bertozzi at al. [8] use a template based approach while Nanda and Davis [4] apply different image filtering techniques. Meis et al. [13] also filter the whole image and classify based on the symmetry calculated for gradients. Xu et al. [2] employ a classification method based on a support vector machine. However, template based detection as well as SVM classification and image filtering over the whole image is time consuming. Xu et al. reported a frame-rate of their system of about 5Hz and the frame rate of system proposed in [4] lies between 3Hz and 11Hz depending on the image resolution. To track a person in the thermal image we use a Particle Filter and a simple elliptic model which is very fast to calculate. Particle Filters [1] have become quite popular in recent years for estimating the state of a system at a given time based on current and past measurements. The probability $p(X_t|Z_t)$ of a system being in the state X_t given a history of measurements $Z_t = \{z_0, ..., z_t\}$ is approximated by a set of N weighted samples:

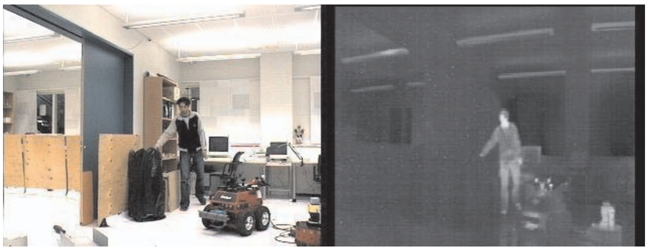

Fig. 3. Person in color and thermal image.

$$S_t = \{x_t^{(i)}, \pi_t^{(i)}\},\ i = 1...N. \tag{1}$$

Each $x_t^{(i)}$ describes a possible state weighted with $\pi_t^{(i)}$ which is proportional to the likelihood that the system is in this state. Particle Filtering consists of three main steps:

1) Create new sample set S_{t+1} by resampling from the old sample set S_t based on the sample weights $\pi_t^{(i)}, i = 1...N$
2) Predict sample states based on the dynamic model $p(x_{t+1}^{(i)}|x_t^{(i)}), i = 1...N$
3) Calculate new weights by application of the measurement model: $\pi_{t+1}^{(i)} \propto p(z_{t+1}|X_{t+1} = x_{t+1}^{(i)}), i = 1...N.$

The estimate of the system state at time t is the weighted mean over all sample states:

$$\hat{X}_t = E(S_t) = \sum_{i=1}^{N} \pi_t^{(i)} x_t^{(i)}. \tag{2}$$

To increase robustness of the system to outliers, instead of calculating the estimate from all samples we use 20% of the samples with the highest weights. 10% of samples with the lowest weights are reinitialised in each iteration. For each sample we use an elliptic contour measurement model to estimate the position of a person in the image: one ellipse describes the position of the body part and one ellipse measures

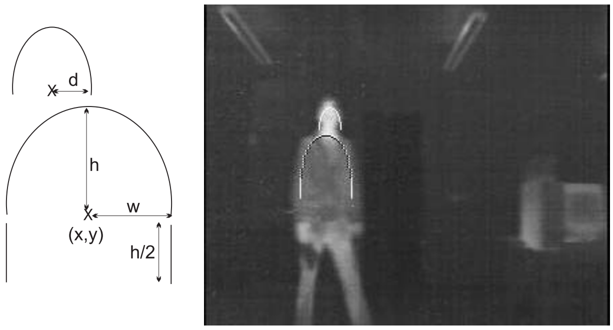

Fig. 4. The elliptic measurement model in thermal images.

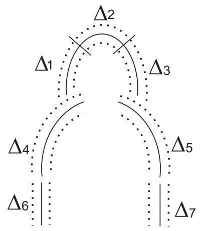

Fig. 5. Elliptic model divided into 7 sections.

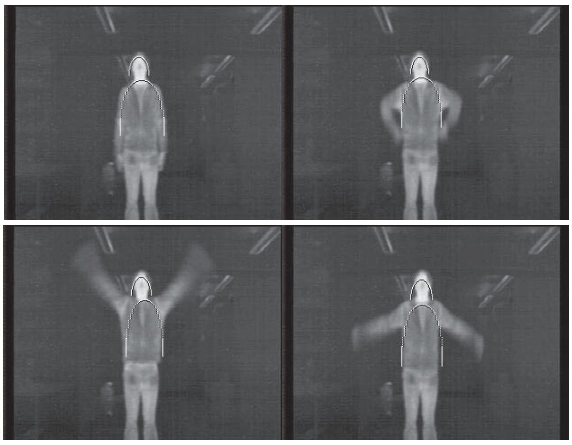

Fig. 6. Tracking with different arm positions.

the position of the head part. Therefore, we end up with a 9-dimensional state vector: $x_t = (x, y, w, h, d, v_x, v_y, v_w, v_h)$ where (x, y) is the mid-point of the body ellipse with a certain width w and height h. The height of the head is calculated by dividing h by a constant factor. The displacement of the middle of the head part from the middle of the body ellipse is described by d. We also model velocities of the body part as (v_x, v_y, v_w, v_h). The elliptic contour model can be seen in figure 4. To calculate the weight $\pi_t^{(i)}$ of a sample i with state $x_t^{(i)}$ we divide the ellipses into different regions (see figure 5) and for each region j the image gradient Δ_j between pixels in the inner part and pixels in the outer part of the ellipse is calculated. The gradient is maximal if the ellipses fit to the contour of a person in the image data. A fitness value $f^{(i)}$ for each sample i is then calculated as the sum of all gradients multiplied with a penalty factor W to reduce the total fitness in the case that a low or negative gradient exists in certain region:

$$f^{(i)} = W \cdot \sum_{j=1}^{m} \Delta_j \tag{3}$$

with

$$W = \sum_{j=1}^{m} w_j, \quad w_j = \begin{cases} 0 & : \text{ if } \Delta_j < \tau \\ \alpha_j & : \text{ otherwise} \end{cases} \tag{4}$$

The value τ defines a gradient threshold and the weights α_j sum up to one and are chosen in a way that the shoulder parts have lower weight to minimize the measurement error that occurs due to different arm positions (see figure 6). The weight of each sample is calculated as the normalised fitness over all samples and the tracker claims a detection if the weighted mean of the fitness of the 20% of the best samples lies above a threshold. The dynamic model that we use for the Particle Filter is a simple random walk: we model a movement with constant velocity plus small random changes. Our approach to track the contour of a person in the image is similar to the work by Isard and Blake [5] for tracking people in a grey image. However, they use a spline model of the head and shoulder contour which cannot be applied in our case because in situations where the person is far away or visible in a side view, there is no recognisable head-shoulder contour. The elliptic contour model is able to cope with these situations. The second advantage of using our contour model is that it can be calculated very quickly due to the fact that we measure only differences between pixel values on the inner and outer part of the ellipse.

In figure 7 one can see the results of tracking a person under different views at different distances. Starting with a frontal view the person turns to a side view, back view and again to a frontal position at the end.

B. Face tracking

After the robot has been able to drive close to the person we switch to the pan-tilt camera and zoom to the expected face region in the image based on the information from the thermal camera. This can be done due to the fact that positions in the thermal image can be transformed to coordinates in the grey image by applying an affine transformation (due to the close proximity of the two sensors, see figure 1).

To detect a face we use the algorithm proposed by Viola and Jones [10], which is considered to be one of the fastest

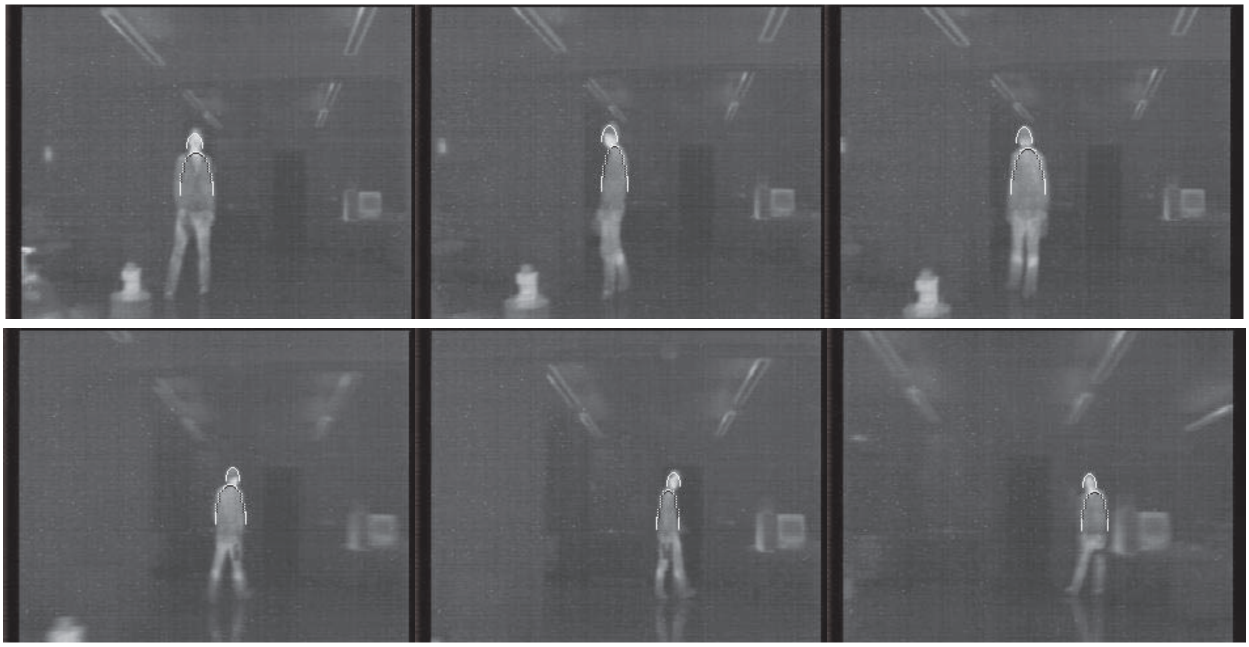

Fig. 7. Tracking under different views.

systems to detect objects in grey value images. With this approach, classifiers that consist of simple grey value features are learned offline on a given training set. Each so-called "strong classifier" is a linear combination of a number of "weak classifiers" which are simple threshold classifiers based on a single grey value feature. The features can be calculated very quickly on a so-called integral image: an integral image II over an image I is defined as $II(x, y) = \sum_{x' \leq x, y' \leq y} I(x', y')$. Good features that are able to discriminate between positive and negative object examples are selected with a boosting mechanism to build the final strong classifiers (for details see [10]).

We train a single strong classifier and instead of scanning the classifier over the whole image at every location and every scale to detect a face (as done in, e.g., [12] or [7]) we use Particle Filtering again: each sample describes a possible face located at position (x, y) and having the scale s. Therefore, the state vector for face tracking becomes $x_t = (x, y, s)$. To calculate the weight $\pi_t^{(i)}$ the classifier is evaluated at the particle's position. Instead of using the binary output of the classifier, we rate each sample according to the weighted sum of all T features which are part of the strong classifier: $\pi_t^{(i)} = \delta \sum_{j=1}^{T} \alpha_j h_j(x_t)$ where $\alpha_j,\ h_j$ are the weighted weak classifiers (see [10]). The dynamic model is again a movement with constant velocity plus small random changes. The face tracker is trained to detect faces under slightly different views and the detected region can also contain parts of the background. Due to the fact that the Eigenface recognition approach is sensitive to different positions of the face center within the located face region, we scan this region to crop out a close area that contains only facial features (see figure 8).

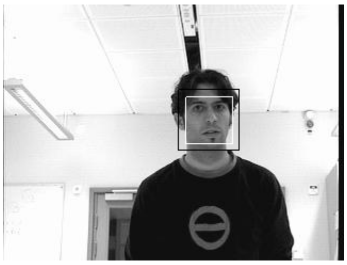

Fig. 8. Face detection.

C. Face recognition

To identify the person we use a face recognition algorithm based on the well-known Eigenface approach [9]. Face regions that are extracted by the face tracker are used to update the probability of the person's identity. Therefore, each face region is rescaled, normalised and projected onto the face-space. The Euclidean distances to each face from the database in the face space is used to calculate the probabilities for each identity. Instead of recognising each frame independently from the next frame (still-to-still recognition) we use each frame to update the identity probability with a Bayesian update rule. If the probability exceeds a certain threshold, the robot announces the estimated identity using its speech synthesizer. Figure 9 shows the face recognition process.

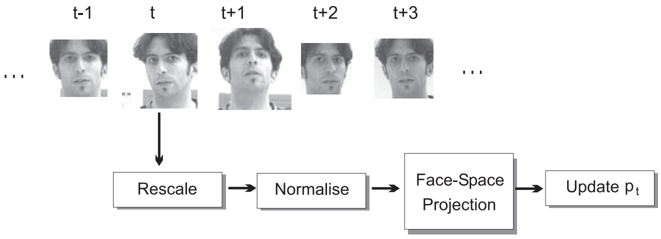

Fig. 9. Face recognition.

III. Experiments

To test the performance of our approach we recorded test sequences with 17 different persons. In each sequence a person stood 4 to 5 meters from the robot in an unconstrained indoor environment, and the robot was started facing away from the person so that it had to turn around and search for the person in the thermal image. While tracking the person in the thermal image, the robot approached and used its pan-tilt camera to extract the face region using the method described in section II-B. After the face had been detected, the person walked to a different position behind the robot and the robot approached a second time, so that we recorded each person under two different light conditions. The length of the recorded sequences varied from 700 to 1400 frames per person with an image resolution of 320×240 for thermal and grey images.

A. Tracking in thermal images

To get ground truth information about the position of a person in the thermal images we used a semi-automatic method to segment the sequences: the result from a segmentation based on a flood-fill algorithm was corrected by hand to extract the exact region in the thermal image containing a person.

In the Particle Filter we used a total of 300 particles, and the weighted mean of the best 20% of all particles of the tracker was compared to the ground truth data for all test sequences. If a person was detected in a frame and the person was visible in the ground truth segmentation, we calculated a detection accuracy d_{acc} as follows:

$$d_{acc} = \frac{2 \cdot n_{overlap}}{n_{detected} + n_{real}}, \tag{5}$$

where $n_{overlap}$ is the number of overlapping pixels between the box around the true person position and the detected position. $n_{detected}$ is the number of pixels in the box, which the tracker returns and n_{real} is the number of pixels in the rectangle around the true person position. Based on the detection accuracy we calculated the following values on each test sequence to evaluate the performance of the tracker:

- False positive rate $FPR = \frac{N_F}{N_N}$ with N_F=number of frames where a person was detected but not visible in ground truth, N_N=total number of frames where no person was visible.
- Detection rate $DR = \frac{N_D}{N_P}$ with N_D= number of frames where person was detected and visible in ground truth, N_P=total number of frames where a person was visible.
- Classification rate CR: percentage of all frames which where correctly classified ($d_{acc} > 0.6$ or no person in ground truth and no person detected).

Figure 10 shows the results of the evaluation. As one can see, CR was in the range from 81% to 95% with a mean over all test sequences of 88.9%. False detections mainly occurred if a person was very close to the robot so that large parts of the image are covered or no head was visible. False detections in this case did not influence the performance of the whole system, due to the fact that in situations where the person is close to the robot, we do not use the information that comes from the thermal image, but rely on the results of the face tracker.

Based on the fact that the tracker claimed a detection only if the weighted mean of the fitness of 20% of the best samples lied above a certain threshold, in some frames the person was not detected. However, those “missed” single frames did not interfere with the tracking process. Using 300 samples we are able to achieve a mean tracking frame-rate of 80Hz on an AthlonXP 1600 processor which leaves enough computational resources for other high-level tasks such as planning, navigation, face recognition etc.

B. Face detection and recognition

The face classifier was trained offline using 4846 images of faces and 7474 non-faces. The final strong classifier consists of 150 features. To test the recognition ability of the system we collected a face database consisting of 8 faces per person that where extracted by the face tracker in the second part of the 17 sequences. The first part had been used as test sequences. We used 500 particles to track the face and in all sequences the face region was successfully detected and tracked. Due to the fact that Particle Filtering is a stochastic process face tracking and recognition experiments had been repeated 5 times on every test sequence. In 41% of all test iterations the face could not be recognized correctly which is mainly due to two problems with using the Eigenface approach:

- Different light conditions: Some images in the training set had strong light from one side.

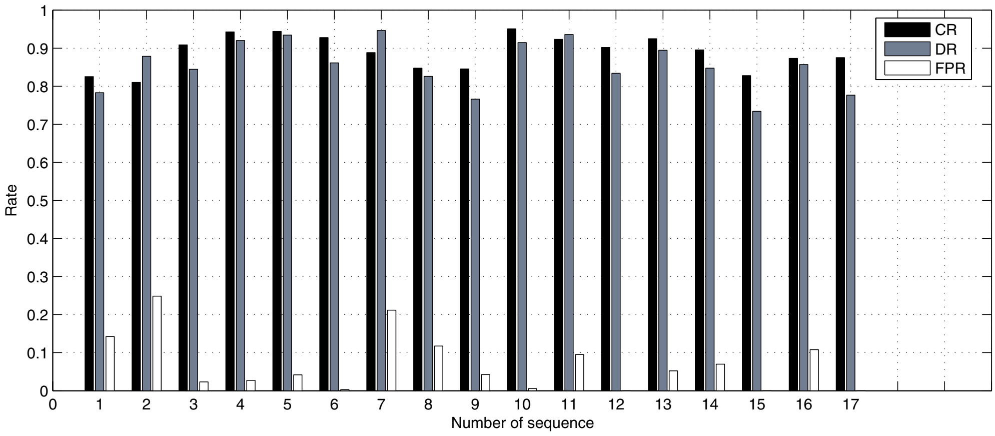

Fig. 10. Tracking result on thermal data.

- Recognition rate depends on the viewing angle and very accurate located and cropped face region.

However, the main focus in this paper lies on detection and tracking in the thermal image so that the improvement of the recognition step by e.g. using a larger database which covers more different light conditions is left for future research.

IV. Conclusion and future work

In this paper we presented a purely vision based approach to track and identify people based on the information from thermal and grey value images. The main contribution of this paper is the application of a thermal camera together with a novel contour measurement model to detect and track people that are further away from the robot and cannot be detected by skin color. Special attention is payed to the real-time ability of this approach. Face detection and recognition is used to identify a person that is close to the robot. In this case we propose the usage of Particle Filtering in combination with a fast face classifier to accumulate evidence about the identity over time, instead of scanning each image independently from the previous one.

Until now, the tracker will always lock onto a single person (the person that has highest measurement probability in the thermal image) but we are currently extending our approach to multiple persons using multiple clusters of particles. To improve and evaluate the person identification part, more experiments with a larger database and different face recognition approaches have to be done. Another direction for future research would be to select actions based on the information provided by our system. For example, if the robot is in front of a person but there is no face visible, it could learn a suitable sensing strategy to get a better look at the face.

References

[1] N. de Freitas A. Doucet and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[2] F. Xu, X. Liu and K. Fujimura. Pedestrian Detection and Tracking with Night Vision. *IEEE Transactions on Intelligent Transportation System*, 5(4), 2004.

[3] G. Cielniak and T. Duckett. Person Identification by Mobile Robots in Indoor Environments. In *Proc. IEEE Int. Workshop on Robotic Sensing (ROSE 2003)*, Örebro, Sweden, 2003.

[4] H. Nanda and L. Davis. Probabilistic Template based Pedestrian Detection in Infrared Videos. In *IEEE Intelligent Vehicle Symposium*, Versailles, France, 2002.

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

[6] J. Barreto, P. Menezes and J. Dias. Human-Robot Interaction based on Haar-like Features and Eigenfaces. In *Proc. of the 2004 IEEE International Conference on Robotic and Automation (ICRA 04)*, pages 1888–1893, New Orleans, LA, 2004.

[7] L. Brèthes, P.Menezes, F. Lerasle and J.Hayet. Face Tracking and Hand Gesture Recognition for Human-Robot Interaction. In *Proc. of the 2004 IEEE International Conference on Robotic and Automation (ICRA 04)*, pages 1901–1906, New Orleans, LA, 2004.

[8] M. Bertozzi, A. Broggi, P.Grisleri, T.Graf and M. Meinecke. Pedestrian Detection in Infrared Images. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 662–667, Columbus, USA, 2003.

[9] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[10] P. Viola and M.J. Jones. Robust real-time object detection. In *Proc. of IEEE Workshop on Statistical and Theories of Computer Vision*, 2001.

[11] S. Lang, M. Kleinhagenbrock, 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. Int. Conf. on Multimodal Interfaces*, pages 28–35, Vancouver, Canada, 2003.

[12] T. Wilhelm, H.-J. Böhme and H.-M. Gross. A Multi-Modal System for Tracking and Analyzing Faces on a Mobile Robot. *Robotics and Autonomous Systems*, 48(1):31–40, 2004.

[13] U. Meis, W. Ritter and H. Neumann. Detection and classification of obstacles in night vision traffic scenes based on infrared image. In *Proc. IEEE Intelligent Transportation Systems*, pages 1140–1144, Shanghai, China, 2003.