Detection, Tracking, and Pursuit of Humans with an Autonomous Mobile Robot

Stefan Feyrer and Andreas Zell

Wilhelm-Schickard-Institute, Department of Computer Architecture
University of Tübingen
Köstlinstr. 6, 72074 Tübingen, Germany
{feyrer, zell}@informatik.uni-tuebingen.de

Abstract

We present a system which is able to visually detect human faces, to track them by controlling a robot-head, and to pursue a detected person by means of driving movements. The detection is based on a multi-modal approach combining color, motion, and contour information. By using a stereo algorithm the position of the person in the scene is determined. Both the path of the person going ahead and a local environment map built by means of range sensor data are used to perform the navigation task. Stationary and dynamic obstacles are avoided during the process of pursuit.

1 Introduction

Acceptance of mobile service robots strongly depends on a man-machine interface which allows the user to communicate and to interact with the robot in an easy and natural way. Speech recognition and generation as well as the visual perception of the environment and in particular the perception of the human user are basic tasks within this context. The problem of integrating several individual basic behaviours into an overall system which is able to act autonomously and in an intelligent way in a dynamic and natural environment has been addressed by different research groups [6], [12]. Although the results are very promising, the problem is still far away from being solved satisfactorily.

In this paper we focus on just one single aspect of such an overall system: The visual detection and the pursuit of humans. Potential applications of this combination are e.g. the guidance of the robot by a human user to a certain operational area or to get the robot familiar with a previously unknown environment.

On a mobile platform there are a number of constraints to be taken into account. First, the system has to be highly robust with regard to changing illumination and scene structure. Furthermore it is important that the processing speed lies within a range that makes it possible to realize a natural reactive behaviour. In this context we have to keep in mind the limited processing power of a mobile robot. These constraints enforce the use of less general and sophisticated but more simple and efficient algorithms. In addition, to achieve real-time performance, image resolution has to be reduced respectively.

The high-level conception of the system has been influenced by these considerations. The two main components of the system – detection and gaze control on the one hand, navigation and motion control on the other – not only have to be considered as two separate entities on a logical and structural level. They are implemented as two independent processes that exchange information via a fast interprocess communication protocol. By this decoupled architecture a fast gaze fixation even during quick movements of the person can be accomplished, while the navigation component is able to independently perform the evaluation of range sensor data and the path planning in an appropriate processing frequency.

2 Vision-Based Person Detection

Earlier person recognition systems were often based on one single cue at the detection level, mainly motion [11], [13] or color [4], [10]. However, the claim for systems working in natural and complex environments requires a higher level of robustness and independence from constraints with respect to the application scenario. To comply with these requirements, many systems use a combination of multiple cues to segment a person from the background [2], [14]. In our approach we also combine several basic cues to increase the robustness of the system.

The first part of the detection process is based on the extraction of skin-colored areas and on the determination of image regions containing significant object movements in the scene. A fusion algorithm combines related and neighboured regions. Besides geometrical and structural cues, contour information is used to finally select one winner region among all detected regions of interest (ROIs). An overview of the detection component is shown in Fig. 1.

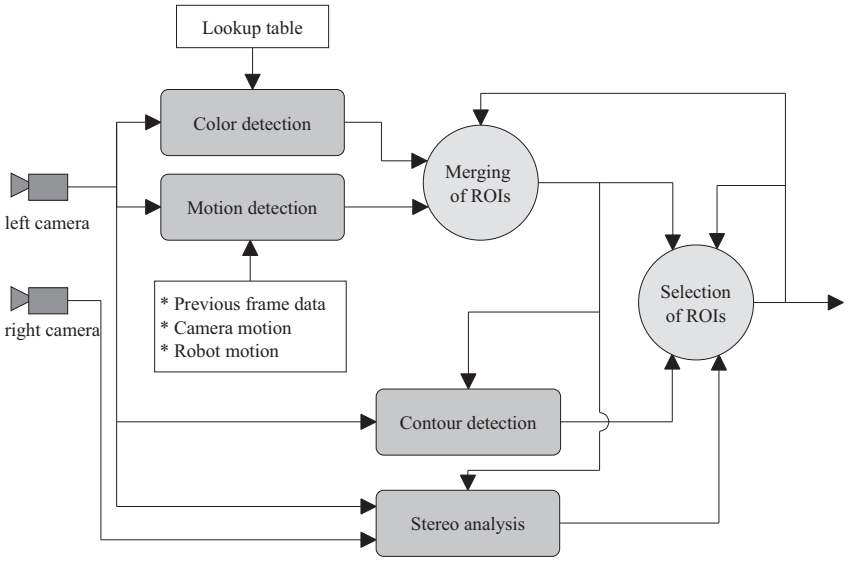

Fig. 1. Structure of the vision module

2.1 Color-Based Detection

To identify skin-colored image regions, we use the HSI color model. Since only pure color information is needed for classification we drop the intensity value. Within the resulting two-dimensional HS-subspace skin color values are defined by means of a lookup table which we have derived from manual segmentation of more than 50 image sequences recorded with 15 people under varying lighting conditions. We found that the resulting skin color areas in HS-space (Fig. 2) neither resemble a two-dimensional normal distribution nor could easily be approximated by geometric primitives. Consequently, in order not to lose too much accuracy, we directly use the resulting table for segmentation. To avoid the time-consuming calculation of the HSI values the lookup table is transformed back to RGB color space.

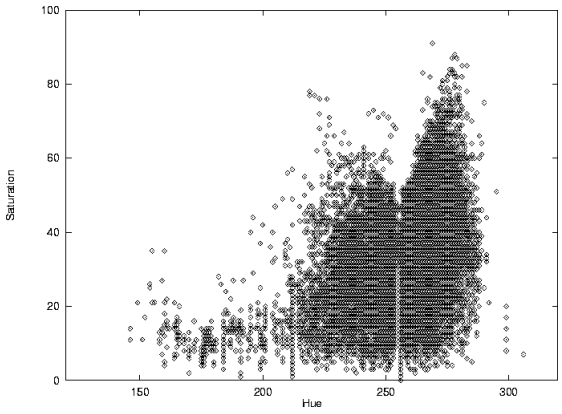

Fig. 2. Result of the manual skin color segmentation. Note that the full range of saturation values is [0..255].

2.2 Motion Detection

The problem of motion detection with a non-stationary camera is to distinguish between the stationary background and objects moving in the scene. Knowing the camera motion parameters and the depth value of each image pixel, compensation for ego-motion is possible but rather time consuming. For this reason we distinguish two cases.

In the case when the robot base does not move and only the pan-tilt-unit with the mounted-on cameras is moving, motion detection is performed by subtraction of two successive frames grabbed with temporarily stationary camera. This approach requires a saccadic camera motion but reduces the false detection rate considerably.

In case of camera and robot moving at the same time, compensation for ego-motion is performed using an approach described in [9]. First, the position and orientation change of the camera between two successive frames is computed in terms of a homogeneous transformation matrix T. Neglecting slipping effects we can easily derive this matrix since both camera and robot motion are controlled by the system itself and therefore are known by the program. Using (1) a predicted image is calculated and matched with the actually grabbed image. All non-matching areas represent moving objects.

$$\text{T} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad x_2 = f \cdot \frac{r_{11}x_1 + r_{12}y_1 + fr_{13} + f\frac{t_x}{Z}}{r_{31}x_1 + r_{32}y_1 + fr_{33} + f\frac{t_z}{Z}} \quad y_2 = f \cdot \frac{r_{21}x_1 + r_{22}y_1 + fr_{33} + f\frac{t_y}{Z}}{r_{31}x_1 + r_{32}y_1 + fr_{33} + f\frac{t_z}{Z}} \tag{1}$$

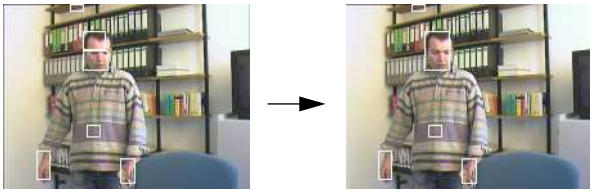

Fig. 3. Merging of ROIs

x_1 and y_1 are the image coordinates of the original image, x_2 and y_2 refer to the predicted image. f denotes the focal length of the camera. The value Z, which denotes the distance between the camera and the objects in the scene for each pixel, is derived by the stereo analysis described in Section 3. Calculation of distance values for all pixels increases the computational complexity and therefore requires a reduction of image resolution at this point.

2.3 Segmentation, Merging, and Selection of ROIs

Grouping of the extracted pixels is done by using an efficient combination of a region growing and a flood fill algorithm together with an integrated morphological opening operator.

It frequently occurs that a face in the image is covered by more than one of the resulting ROIs. Especially the eye region often separates two ROIs (see Fig. 3). Therefore a fusion algorithm must be applied. The algorithm is based on a depth-limited traversal of a neighbourhood graph of ROIs. Each combination of ROIs occurring during traversal is evaluated using the method described below. Those combinations gaining the highest rating are merged.

Face regions are distinguished from other segmentation clusters by a structure-based selection and evaluation process. Geometric criteria like the height-width ratio of the circumscribing rectangle or the compactness of the region are combined with statistical cues like the average of the relative frequencies of the HS skin color combinations within one region. Because of the temporal coherence of image sequences we can assume that there are merely small changes in size and position from one frame to the next. For this reason size and position changes with respect to the detected face region in the previous frame are also included in the selection process.

All these single cues, together with distance information determined by the stereo analysis and the results of the contour detection, are evaluated using a probabilistic framework which yields a confidence factor describing the likelihood of a ROI for containing a face. The mean values and correlation coefficients for each cue have been calculated based on a large set of face regions taken from our sample image sequences which have already been used to generate the color lookup table. The results of the motion analysis are taken into account by a constant additional value. The ROI with the highest confidence value is set to be the winning ROI for the next iteration.

2.4 Integration of Contour Information

Denzler and Niemann introduced a method for fast contour tracking of convex objects, called *active rays* [3]. Active rays are derived from active contours [1]. The main difference is that instead of an energy minimization in the 2D image plane the minimization is reduced to a 1D search problem. Furthermore the approach shows any-time behaviour which is important with respect to real-time applications like ours. In our system we use the active ray approach with a slightly modified energy minimization term.

Starting at a reference point within the object boundaries, which is provided here by the color segmentation module, grey levels are examined along straight lines passing through this point. An energy minimization yields a chain of contour points approximating the object boundaries. During the minimization process our algorithm not only uses the grey level values of the image pixels but also includes information previously derived by the motion analysis. Subsequently, the detected chains of contour points are matched with a statistically determined average head contour. Fig. 4 shows the determined contour points connected to a polygonal chain in a sample image sequence.

3 Stereo Analysis

Stereo analysis plays an important role in our application because localization of a person in the three-dimensional scene and thus determination of a path is only possible when depth information is provided. In addition, depth information makes a valuable contribution to the ROI selection process, assuming that there are no significant changes of distance between the robot and the pursued person between two subsequent frame captures. Both cases – path determination and plausibility check – only require distance information within an extracted ROI, whereas the compensation for ego-motion needs dense depth maps and therefore is considerably more time-consuming.

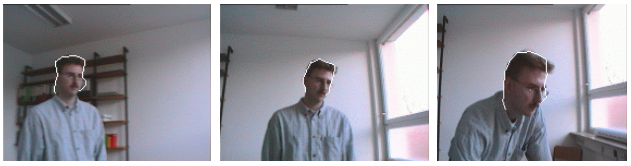

Fig. 4. Extraction of contours

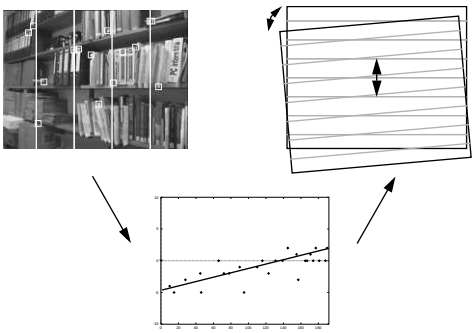

Fig. 5. Calibration of the stereo geometry

The stereo algorithm we use is based on a standard stereo geometry, i.e. the optical axes are parallel and the scan lines of both images are aligned with the optical center axis. This kind of geometry facilitates the correlation task since it suffices to search for correspondences along the image lines. One drawback however is, that the physical geometry is difficult to adjust precisely and is very sensitive to vibrations or other mechanical disturbances. Therefore it is necessary to have a fast working auto-calibration algorithm, which recalibrates the external camera parameters in certain time intervals automatically.

Our auto-calibration algorithm works as follows. First, the image is separated into columns of equal width. Within each column significant image points are determined using the Moravec feature detection operator [8]. Subsequently, correspondences for these points are searched within a stripe consisting of several image lines. By choosing the line containing the best correspondence we get a vertical offset between the left and the right image at each feature position. The maximum and minimum offsets within each column are dropped and the remaining values are used to calculate a linear regression line which describes the position and orientation of the epipolar lines. Instead of searching for correspondences along image lines the subsequent correlation analysis is performed along the resulting regression line. Thus, vertical offset errors as well as rotational errors with respect to the optical axes can be compensated (see Fig. 5).

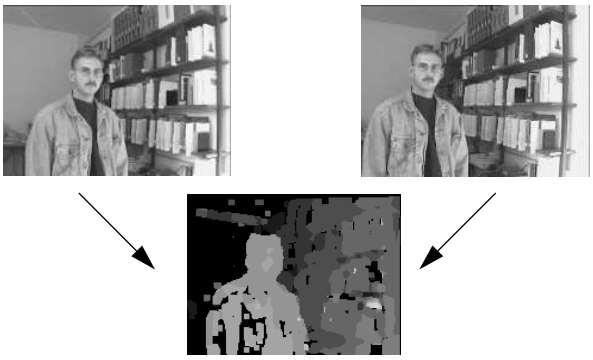

Fig. 6. Stereo image pair and resulting depth image

Fig. 7. Camera platform with pan-tilt-unit

There are a number of different techniques for correlation analysis. For performance reasons we have chosen a method which correlates normalized grey level areas of size 5x5 [5]. The correlation measure is defined by

$$s = \max(0, 1 - c), \tag{2}$$

where

$$c = \frac{\sum_{i,j}((L(x+i, y+j) - \bar{L}) - (R(x+dx+i, y+dy+j) - \bar{R}))^2}{\sqrt{(\sum_{i,j}(L(x+i, y+j) - \bar{L})^2) \cdot (\sum_{i,j}(R(x+dx+i, y+dy+j) - \bar{R})^2)}}. \tag{3}$$

L and R denote the left resp. the right image matrix, $\bar{L}$ and $\bar{R}$ are the average grey level values of the respective correlation window.

By restricting the search to an interval heuristically depending on the disparity of the previous pixel and by dropping areas with insufficient grey level variance we further reduce computation time. To eliminate false correspondences to the greatest possible extent, the depth images are postprocessed by several morphological operations. Fig. 6 shows the resulting depth image of a stereo image pair.

4 Camera Control and Navigation

The camera platform with the two CCD cameras has two degrees of freedom described by a pan angle ψ and a tilt angle θ. In order to facilitate motion detection we use a saccadic control of the pan-tilt-unit. Assuming a pinhole camera model we can easily determine the rotation angles necessary to center the detected face in the image:

$$\psi = \Delta x \cdot \frac{\psi_{max}}{w} \qquad \theta = \Delta y \cdot \frac{\theta_{max}}{h} \tag{4}$$

w and h denote the image dimensions in pixels. Δx and Δy are the pixel offsets of the detected face with respect to the image center. The two maximum angle values describe the opening angle of the camera. The error we make due to an offset between the axes of rotation and the lens centers is smaller than the accuracy of measurement of the vision module and thus is negligible.

The task of the navigation module is to convert the results of the vision module into robot movements, i.e. to pur-

sue a detected person resp. to approach a person up to a certain minimum distance. At the same time obstacles should be avoided. To accomplish this, we use the potential field method [7]. This method works on an artificial potential field U, which is defined as the sum of an attractive potential U_{attr} pulling the robot towards the goal position and a repulsive potential U_{rep} pushing the robot away from the obstacles. Since we have no global obstacle map, the raw distance values ρ_k provided by the laser scanner are used to calculate the repulsive potential:

$$U_{rep} = \sum_{k=1}^{r} U_k \cdot \sin\varphi_k$$

$$U_k = \begin{cases} \frac{1}{2}\eta\left(\frac{1}{\rho_k} - \frac{1}{\rho_0}\right)^2 & \text{for} \quad \rho_k \le \rho_0 \\ 0 & \text{else} \end{cases} \tag{5}$$

ρ_0 denotes the distance of influence, i.e. the maximum distance to an obstacle being used for the calculation of the potential. The attractive potential is defined as follows:

$$U_{attr} = \frac{1}{2}\xi\rho_{goal}^2 \tag{6}$$

ρ_{goal} is the current distance to the goal. η and ξ are scaling factors describing the relationship between the attractive and the repulsive potential. The resulting direction of motion is determined by an artificial force induced by the potential field:

$$\vec{F} = -\vec{\nabla}U \tag{7}$$

The translational velocity depends on the distance between the robot and the current goal. When the distance falls below a minimum value of approximately 1 m, the velocity is set to zero.

Navigation as such is performed in an iterative fashion. At each iteration step, a new goal position is determined. This is done by checking which of the remembered path positions of the person are within direct reach. If there are reachable positions, the most current one is chosen as the new goal. If all positions are out of direct reach, the last position at which a person was detected is used. Whenever a remembered path position is reached by the robot, the list entry and all preceding entries are removed. By remembering the person's path positions and dynamically determining the goal position, dead end situations can be avoided, provided that the dimensions of the robot allow the navigation to all positions the person has passed through.

5 Experimental Results

The presented system works on a RWI B21 mobile robot (Fig. 8). The robot is equipped with two dualprocessor 200 MHz PentiumPro PCs. The main function of one of them is image processing, while the other one is mainly used for range sensor evaluation and navigation. Image acquisition is done by two CCD color cameras mounted on a pan-tilt-unit. For navigation a built-in 2D laser range finder is used.

Fig. 8. The mobile service robots Robin and Colin

Right now we achieve a processing frequency of 8 frames per second at an image resolution of 192x144 pixels in case of a stationary robot base. When the robot is moving, the performance decreases to 6 frames per second due to the time-consuming compensation for ego-motion. Nevertheless even during motion, tracking and pursuit of humans not moving unusually quickly is possible. The velocity varies up to a maximum of 0.5 m/s depending on the environment and in particular on the distance to obstacles and to the person.

Recognition performance in terms of correct detection rates is hard to determine because of the strong dependence on scene structure and scene dynamics. However, as a qualitative result, analysis of several image sequences showed that we get considerably higher recognition rates when we add motion analysis to simple color detection. Additionally integrating contour information leads to a relatively modest improvement of the recognition performance. Tab. 1 exemplarily shows the detection rates for one selected image sequence. The values are calculated dividing the number of images in which the head has been detected correctly by the total number of images in the sequence.

	Contour detection off	Contour detection on
Motion detection off	92.5%	95.0%
Motion detection on	97.9%	98.9%

Table 1. Head recognition rates with respect to a selected sample image sequence

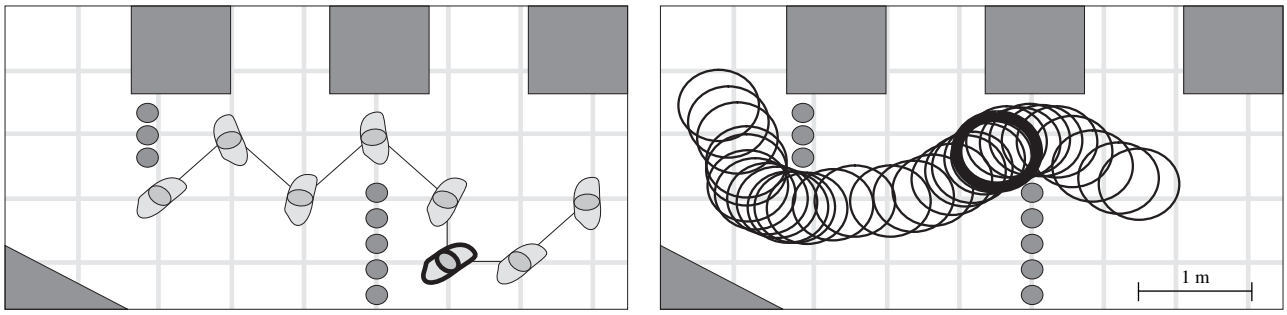

Fig. 9. Path of a person and path of the pursuing robot drawn manually from trace data recorded on the real robot

To illustrate the behaviour of pursuit, the path of a person walking backwards along marking lines on the floor and the path of the pursuing robot are depicted in Fig. 9. The temporal relationship of the two paths can be read from the marked positions where the person has already passed the obstacle while the robot is still located behind it.

6 Conclusion

We have presented a system which is capable of following a person with a mobile robot in real-time. The use of several cues during the detection process results in a high robustness needed for acting in complex natural environments. The navigation component combines obstacle avoidance based on range sensor data with the task of pursuing a person walking ahead. Experiments have shown that the system successfully handles difficult situations like highly dynamic environments as well as narrow doorways.

The main limitation of the current system is that the person has to face the robot constantly and therefore has to walk backwards. To overcome this restriction future work will be on the realization of a more flexible detection module, which, after an initial phase, will then be able to track the person even when the face has been turned away. Other major issues of future research will be an improved image sequence analysis and the integration of a prediction component regarding the position of the person in the three-dimensional scene.

References

[1] A. Blake and A. Yuille (eds.), *Active Vision*, MIT Press, Cambridge, Massachusetts, 1992.

[2] J. L. Crowley, F. Berard, "Multi-Modal Tracking of Faces for Video Communications," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 640-645, 1997.

[3] J. Denzler and H. Niemann, "Active Rays: A New Approach to Contour Tracking,", *Int. Journal of Computing and Information Technology*, vol. 4, no. 1, pp. 9-16, 1996.

[4] P. Fieguth and D. Terzopoulos, "Color-Based Tracking of Heads and Other Mobile Objects at Stereo Frame Rates," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 21-27, 1997.

[5] P. Fua, "A Parallel Stereo Algorithm That Produces Dense Depth Maps and Preserves Image Features," *Machine Vision and Applications*, vol. 6, no. 1, pp. 35-49, 1993.

[6] Y. K. Hwang, S. C. Kang, S. Lee, S. M. Park, K. R. Cho, H. S. Kim, and C. W. Lee, "Human Interface, Automatic Planning, and Control of a Humanoid Robot," *Int. Journal of Robotics Research*, vol. 17, no. 11, pp. 1131-1149, 1998.

[7] O. Khatib, "Real-Time Obstacle Avoidance for Manipulators and Mobile Robots," *Int. Journal of Robotics Research*, vol. 5, no. 1, pp. 90-98, 1986.

[8] H. Moravec, "Towards Automatic Visual Obstacle Avoidance," in *Proc. 5th Int. Joint Conf. on Artificial Intelligence, Vision-I*, p. 584, 1977.

[9] D. Murray and A. Basu, "Motion Tracking with an Active Camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 449-459, 1994.

[10] Y. Raja, S. J. McKenna, and S. Gong, "Tracking and Segmenting People in Varying Lighting Conditions," in *Proc. 3rd Int. Conf. on Automatic Face and Gesture Recognition*, pp. 228-233, 1998.

[11] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94-115, 1994.

[12] C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, and R. Wörz, "Integrating Vision Based Behaviours with an Autonomous Robot," in *Proc. Int. Conf. on Vision Systems*, pp. 1-20, 1999.

[13] A. Shio and J. Sklansky, "Segmentation of People in Motion," in *Proc. IEEE Workshop on Visual Motion*, pp. 325-332, 1991.

[14] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780-785, 1997.