Proceedings of the 1999 IEEE
International Conference on Robotics & Automation
Detroit, Michigan • May 1999

# A Person Following Behaviour for a Mobile Robot

H. Sidenbladh[1], D. Kragić[1] and H. I. Christensen[2]

[1]Computer Vision and Active Perception Lab
Dept of Numerical Analysis and Computing Science
Royal Institute of Technology
SE–100 44 Stockholm, Sweden

[2]Centre for Autonomous Systems
Dept of Numerical Analysis and Computing Science
Royal Institute of Technology
SE–100 44 Stockholm, Sweden

## Abstract

*In this paper, a person following behaviour for a mobile robot is presented. The head of the person is located using skin colour detection. Then, a control loop is fed with the camera movements required to put the upper part of the person in the center of the image. The algorithm was tested in different rooms of a research lab. It performed well in all lightings except in direct sunlight. Since the background and lighting cannot be controlled, the vision algorithm must be robust to such changes. However, since the computing power is quite limited, the algorithm must have as low complexity as possible.*

## 1 Introduction

Fixating and following humans or objects with the eyes is a low–level function that is essential for human perception. This implies that a robot that is going to interact with the surrounding world should possess this skill. The discipline of active vision has started to develop in the 80's by [4, 2, 5]. The idea of active vision is that the information extracted from visual perception should serve a certain purpose, dependent of the application. In other words, a vision system should not spend time on obtaining the maximum information from the image but rather concentrate on particular information related to the given task [7].

This paper presents work on person following implemented on a mobile platform in an office environment. Tracking a moving object in general has attracted a significant amount of attention in the past few years. Many methods presented employ optical flow [14, 13, 10], stereo correlation [1, 9, 3] or colour information techniques. For many methods that exploit optical flow computational time needed for the computational complexity of flow segmentation is high and, therefore, this problem is far from being solved. Skin colour have been used extensively for person segmentation and tracking [11, 16, 8], since it is distinct and quite robust to lighting and the person's race.

The Intelligent Service Robot project aims at constructing a mobile, semi–autonomous, behaviour–based robot which will assist people in their homes. Suitable tasks could be to wash the dishes, help elderly people to rise from a chair, find the TV remote control or vacuum the floor. It is equipped with sonars, a laser scanner and two cameras on pan–tilt units. The sonars and the laser scanner are used for navigation, surveying and obstacle detection [18], while the cameras are used for more fine–tune navigation, object recognition, person following and gesture recognition.

The person following behaviour will help the robot to fixate on a person so that it could recognise the gestures of the person or what he is doing, in order to function like an attentive, human–like assistant [6]. The fixation can also have a psychologically positive effect on the person interacting with the robot [12].

The mobility of the robot opens up for new problems within the vision field. Background and lighting can not be controlled, and the robot's movements must be compensated for by movements of the camera. Because of this, robustness is a much larger issue than in desktop applications. The computing power is also limited, which puts constraints on the complexity of the algorithms.

## 2 Outline of the paper

The system presented in this paper consists of a vision module, which calculates the current position of the head in the camera image. This information is sent in parallel to a fixation module which adjusts the pan and tilt of the camera, and a following module which controls the wheels of the robot on which the camera is mounted. The fixation is faster than the

0-7803-5180-0-5/99 $10.00 © 1999 IEEE

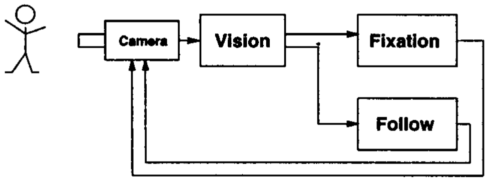


Figure 1: Overview of the system. The vision algorithm receives images from the camera. The fixation algorithm steers the pan–tilt unit on which the camera is mounted. The following algorithm steers the wheels of the robot.

following, which simplifies control. The two modules will indepentent of each other strive to center the person in the image. The architecture of the system is outlined in Figure 1.

In Section 3, the vision algorithm is described. The camera and robot control algorithms are presented in Section 4. Sample experimental results are shown in Section 5. Finally a summary and issues for future research are presented.

# 3 Vision algorithm

The person is detected in the image using skin colour segmentation. A graphical representation of the vision algorithm can be seen in Figure 2. As for now, only one camera is used. However, since the mobile robot is equipped with two cameras it is possible to use disparity cues for segmentation as well as colour cues.

## 3.1 Initialisation of head position

The head of a person is almost never covered by clothing, and it is usually not subject to occlusion. Therefore, it is convenient to use the head as the initial cue for localisation. The detection of the head is described below.

First, the image is transformed from RGB to HSV (Hue–Saturation–Value) space:

$$H = \arccos\left[\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right] \quad (1)$$

$$S = 1-\frac{3}{(R+G+B)}\min(R,G,B) \quad (2)$$

$$V = \frac{1}{3}(R+G+B) \quad (3)$$

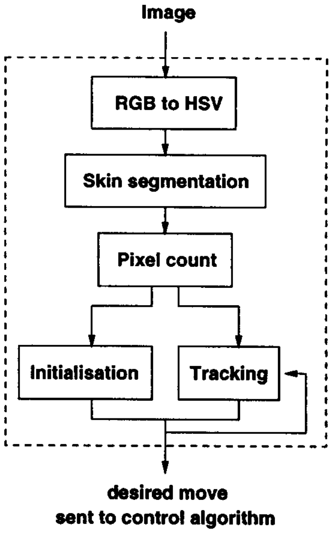


Figure 2: Flow scheme for the visual algorithm

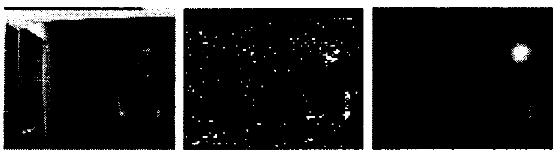
Then, it is segmented into skin colour and non–skin colour using hue, saturation and value thresholds (Figure 3b). The thresholds have previously been trained manually to provide robust segmentation in different rooms of the CVAP laboratory.

If the amount of skin in the image is below a certain threshold (dependent of the colour thresholds), the image is considered not to contain a person, and no following is initiated.

The skin colour subspace is kept quite small, to ensure that very few non–skin pixels are classified as skin. The result is a quite noisy segmentation image, where skin areas have a higher density of skin pixels than other areas. To get a measure of the density of the pixels classified as skin, the number of skin pixels within a certain distance from each pixel is counted (Figure 3c). The distance within which the pixels are counted corresponds to a typical skin area size of a person whose head, arms and upper body are visible in the image. This depends on the intrinsic parameters of the camera, and on the distance to the person.

The head is considered to be the largest skin area in the image. The largest area corresponds to the highest value in the skin pixel count image (Figure 3c). The head can also be assumed to be high up in the image since it is the highest part of the person. Thus, the head is assumed to be located at the maximum of the upper half of the skin pixel count image.

Since the person follow behaviour is going to be



(a) Original image

(b) Skin segmentation

(c) Count of adjacent skin pixels

Figure 3: Example of skin segmentation. The original image is in colour. Besides the person, it contains a wooden bookshelf and walls in a colour close to skin.

used together with gesture recognition and detection, the arms and head of the person must be visible. Therefore, the cameras should be directed so that the center of the image is positioned a short distance below the head center. This distance is proportional to the person size mentioned above.

### 3.2 Tracking of the head

Once the head of the person is found, it is tracked in subsequent frames. The position $\mathbf{p_k}$ in frame $k$ is estimated using a simple first order Taylor approximation, i.e.:

$$\mathbf{p_k} = \mathbf{p_{k-1}} + \mathbf{\Delta p} \pm \mathbf{d} \tag{4}$$

where $\pm\mathbf{d}$ corresponds to a square of size $2 * d$ (large $d$ gives a large search area, i.e. the head can make sudden accelerations without the system losing track) and $\mathbf{\Delta p} = \mathbf{p_{k-1}} - \mathbf{p_{k-2}}$.

If no large enough skin pixel count is found within this area, the tracker is re-initialised (Figure 2).

### 3.3 Output to control modules

The output sent to the control algorithm is the desired camera movement to place the fixation point just below the head. (Figure 2).

## 4 Active tracking

### 4.1 Camera model

The camera is modeled as a pinhole. In our current implementation we convert image coordinates into angular measures under perspective projection (Figure 4).

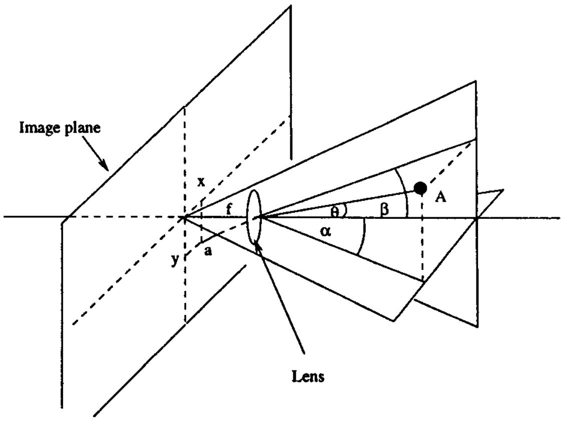


Figure 4: Under perspective projection the 3D point with a tilt angle $\beta$ and pan angle $\alpha$ in the camera referential projects on the image plane at the 2D point of coordinates (x,y).

### 4.2 Control algorithm

Under perspective projection the velocity of a point $(X, Y, Z)$ in space can be associated to the velocity of a point $(x, y)$ in the image space as [15, 17]:

$$\begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-f}{Z} & 0 & \frac{x}{Z} & \frac{xy}{f} & \frac{-f^2-x^2}{f} & y \\ 0 & \frac{-f}{Z} & \frac{y}{Z} & \frac{f^2+y^2}{f} & \frac{-xy}{f} & -x \end{bmatrix} T_c \tag{5}$$

where

$$T_c = \begin{bmatrix} \frac{dX_c}{dt} & \frac{dY_c}{dt} & \frac{dZ_c}{dt} & \omega_{Xc} & \omega_{Yc} & \omega_{Zc} \end{bmatrix}^T \tag{6}$$

Here, $f$ is the focal length of the camera's lens, $T$ is a vector of 3 translational and 3 rotational velocities and subscript $c$ denotes the relation to the camera frame. If we consider the inverse situation where the target is state and we have to control two angles of the camera rotation (i.e. pan and tilt angle), we get:

$$\begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix} = \begin{bmatrix} \frac{xy}{f} & \frac{-f^2-x^2}{f} \\ \frac{f^2+y^2}{f} & \frac{-xy}{f} \end{bmatrix} \begin{bmatrix} \omega_{Xc} \\ \omega_{Yc} \end{bmatrix} \tag{7}$$

This yields:

$$\begin{bmatrix} \frac{dq_{pan}}{dt} \\ \frac{dq_{tilt}}{dt} \end{bmatrix} = \begin{bmatrix} -\omega_{Xc} \\ \omega_{Yc} \end{bmatrix} \tag{8}$$

where $q$ is the pan and tilt angles. The control system uses visual measurements obtained from the image to drive the pan and tilt angels. The goal is to keep the

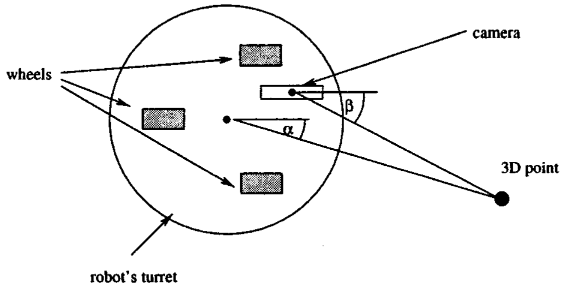


Figure 5: The offset between the camera axis and wheels axis of rotation cause the difference in angles used to control the direction of the pan and wheels respectively.

torso of the person centered in the image. The error signal is defined as the difference between the target image position and some reference position which is in our case the center of the image $(x_0, y_0)$:

$$e = [\ x - x_0 \quad y - y_0\ ]^T \tag{9}$$

where $x$ and $y$ the components of the target position vector. We are using the P controller to recenter the target in the image. The controller can be defined as:

$$\Delta q = [\ \Delta q_{pan} \quad \Delta q_{tilt}\ ]^T = K_p e \tag{10}$$

where $K_p$ is p–type controller constant. Visual measurements are also used to control the robot's wheels in the same manner as the pan angle is controlled. Since there is an offset between axis of rotation for the camera pan angle and the wheels' angle of rotation, we have to compensate for it (Figure 5). To achieve smooth pursuit a dead band is introduced in the visual module, i.e. we are ignoring small angles resulting from changed lightning conditions, size of the blob due to the rotation of the person relative to the camera, etc.

## 5 Experiments

### 5.1 Experimental Setup

The platform used for experiments was a Nomad 200, which is an integrated mobile robot system with wheels and turret. It has an onboard Pentium 133MHz for sensor and motor control and for host computer communication.

The camera used was a Sony XC-999 CCD color video camera with chip area of 1/2", automatic white balancing, electronic shutter and a 6 mm C-mounted lens. The camera is mounted on a pan-tilt unit on the robot.

The pan-tilt unit, delivered by Directed Perception, has two degrees of freedom (pan and tilt angles) with a maximum rotational speed of 60° $s^{-1}$. For image acquisition a PCI-based board from Fujitsu was used.

### 5.2 Results

The person tracking was tested in different rooms of the CVAP laboratory. The walls and the floor has a colour somewhat close to skin, which sometimes creates problems if the person is too far away, i.e. the skin areas are too small.

The program ran in 5 Hz with input images of size 60 × 80 pixels. They were segmented according to the criterion $((hue < 30)\&(30 < saturation < 150)\&(60 < value < 220))$ which corresponds to orange–red, not grey, not too deeply coloured, not black and not white. The counting of adjacent skin pixels was performed on a 12 × 12 area around each pixel.

#### 5.2.1 Pan–tilt experiments

These two experiments show the performance of the fixation of the person with the camera's pan–tilt unit. The wheel and turret following has been turned off. In the first experiment, a person is standing still 1.5 m in front of and 0.75 m to the left of the robot. The start values of the pan and tilt of the camera was zero, i.e, the robot looked straight ahead. Figure 6 shows the camera movement during the first 3 seconds of the program session. We can see an initial phase about 0.5 seconds long, when the robot turns the camera to center the person in the image. In the next experiment, the robot and the person is in the same configuration, but then the robot moves to the left with a speed of $0.2 m/s$. This causes the tilt to change constantly to keep the camera fixated on the person. The results can be seen in Figure 7. The "stepping" behaviour is due to the dead–band introduced into the controller. The initial phase here includes a small oscillation in the tilt direction, which makes it longer, about 4 seconds. This causes no big problems, since the person moves very little in vertical direction.

#### 5.2.2 Control experiment

The third experiment demonstrates the fixation module's robustness to noise. A person moves back and forth in front of the robot. The horisontal and vertical dislocation of the head from the ideal position is

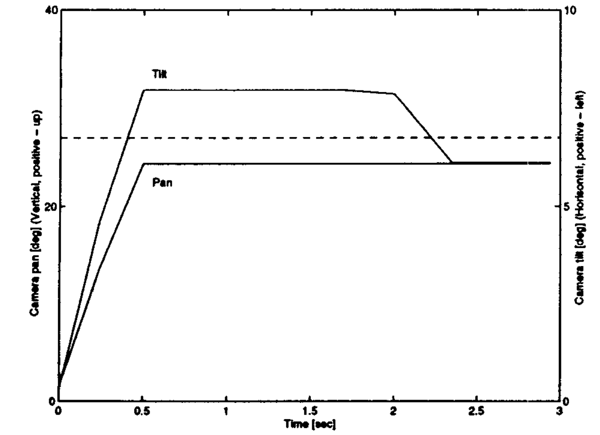


Figure 6: Plotting of the camera's pan and tilt angles for pan–tilt experiment 1. The dotted line are the ideal angle values, i.e. the angles for which the person would have been in the center of the image.

interpreted by the controller as the tilt and pan speed in *deg/s* in which the camera should move. This is the in–signal to the controller (dotted curve in Figure 8). Figure 8 shows that the control of the pan is slower than the control of the tilt, and therefore more insensitive to momentary errors in tracking of the head. Smoothing of the in–signal would improve the tilt controls robustness to noise.

# 6 Conclusion

## 6.1 Summary

A behaviour for person following with a camera mounted on a robot was presented. The position of a person is estimated using skin colour detection, and the pan and tilt of the camera is controlled in order to fixate on the upper part of a moving person. To enable following of the person both the pan-tilt unit and the actual platform are controlled. The control ensures fixation on the person, while it maintains a certain distance to the person. Experiments demonstrate a good performance in a laboratory environment, where lighting and background varies significantly. A solid basis is thus available for future work on interpretation of the gestures of a human operator.

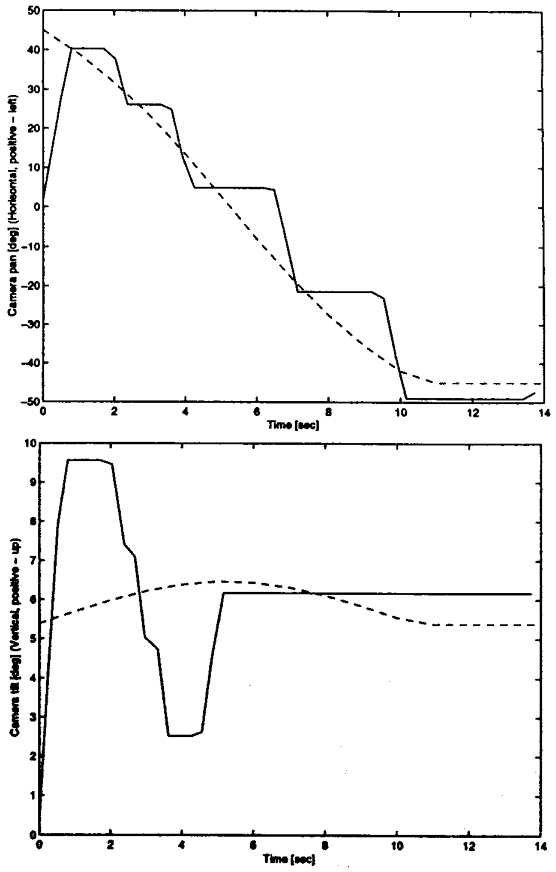


Figure 7: Plotting of the camera's pan and tilt angles for experiment 2. The dotted lines are the ideal angle values, i.e. the angles for which the person would have been in the center of the image.

## 6.2 Future work

The work presented in this report could be extended in several directions.

As mentioned in Section 3, a faster initialisation of the head position estimator would improve the robustness so that re-initialisation would become less frequent. Adding a Kalman filter would also allow more robust estimation of the position of the head, which in turn would allow smoother control. Furthermore, the area of search for the head could be reduced, which would speed up the system.

However, the most critical part is the skin colour segmentation. By combining segmentation results from colour and disparity, the effects from skin coloured background will be reduced. However, a

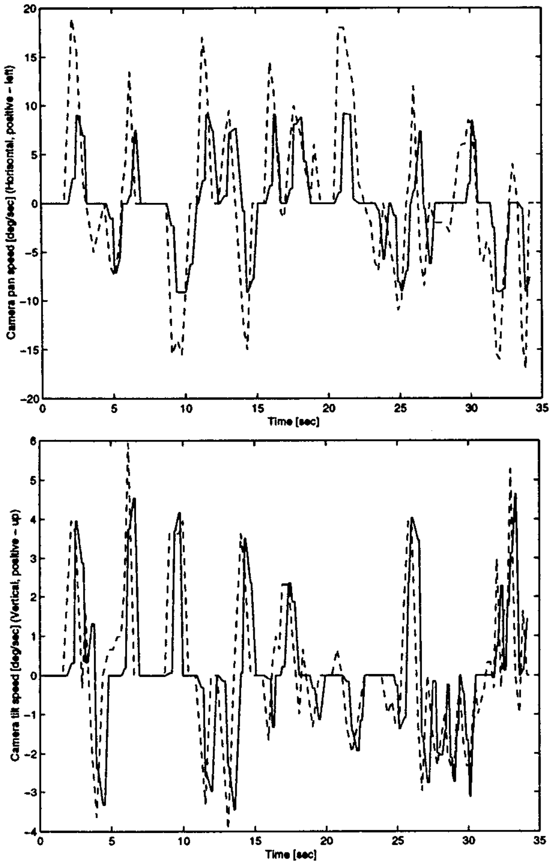


Figure 8: Plotting of the in–signal and the out–signal ($deg/s$) of the fixation control system. The dotted lines are the required pan and tilt speed, the non–dotted lines are the actual speed of the pan–tilt unit.

more complex segmentation algorithm would slow down the program.

The thresholds for segmentation are presently fixed. To allow for added variations in lighting it is desirable to introduce adaptive thresholding of the colour space.

Since the platform we use has a turret the system can be extended to allow control of pan and turret angles simultaneously. With that we will be able to cover 360° around the platform.

In addition, threshold variables deciding whether to trigger a saccade or not were empirically set. A low distance threshold can result in unstable control. To reduce the problem it might be necessary to introduce automatic focusing to extend the operating range of the system.

## References

[1] A.Arsenio and J.Santos-Victor. Robust visual tracking by an active observer. *Proc. IROS*, 3:1342–1347, 1997.

[2] J.Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *IJCV*, 1(4), January 1988.

[3] A.Maki, T.Uhlin, and J.O.Eklundh. Phase-basd disparity estimation in binocular tracking. *Proc. SCIA*, pages 1145–1152, 1993.

[4] R. Bajcsy. Active perception. *IEEE Proc.*, 76(8):996–1006, August 1988.

[5] D. Ballard. Animate vision. *Artificial Intelligence*, 48(1):1–27, February 1991.

[6] R. Cipolla and A.Pentland, editors. *Computer Vision for Human-Machine Interaction.* Cambridge University Press, 1998.

[7] J.L. Crowley and H.I. Christensen. Integration and control of active visual pocesses. *IROS*, August 1995.

[8] J.L. Crowley and J. Coutaz. Vision for man machine interaction. *EHCI*, August 1995.

[9] I.D.Reid and D.W.Murray. Active tracking of foveated feature clusters using affine structure. *IJCV*, 18:41–60, 1996.

[10] J.Santos-Victor and G.Sandini. Visual behaviours for docking. *Computer Vision and Image Understanding*, 1997.

[11] R. Kjeldsen and J. Kender. Finding skin in color images. In *IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 312–317, 1996.

[12] P. Maes. *Designing Aoutonomous Agents: Theory and Practice from Biology to Engineering and Back.* Bradford Books/MIT Press, 1991.

[13] D. Murray and A. Basu. Motion tracking with an active camera. *IEEE PAMI*, 16(5):449–459, 1994.

[14] P. Nordlund and T. Uhlin. Closing the loop:detection and pursuit of a moving object by a moving observer. *Image and Vision Computing*, 14(4):265–275, 1996.

[15] P. Y. Oh and P. Allen. Design of a partitioned visual feedback controller. *IEEE ICRA*, 2:1360–1365, 1998.

[16] Y. Raja, S. J. McKenna, and S. Gong. Tracking and segmenting people in varying light conditions using colour. In *IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 228–233, 1998.

[17] S.Hutchinson, G.D.Hager, and P.I.Corke. A tutorial on visual servo control. *IEEE TRA*, 12(5):651–670, 1996.

[18] O. Wijk, P. Jensfelt, and H. I. Christensen. Triangulation based fusion of ultrasonic sonar data. In *IEEE Int. Conference on Robotics and Automation*, 1998.