A&QT-R 2002 (THETA 13)
International Conference on Automation, Quality and Testing, Robotics
May 23 – 25, 2002, Cluj-Napoca, Romania

ADDRESSING AND TAKING PHOTOS OF PASSERSBY BY FUSING COLOR IMAGES AND LASER RANGE SCANS

Axel Walthelm

Medical University of Lübeck
Institute of Computer Engineering
Ratzeburger Allee 160
D-23538 Lübeck, Germany
walthelm@iti.mu-luebeck.de

Abstract: Service Robots have to interact with humans, but detecting and identifying humans is difficult. Before for example image based face recognition algorithms can be applied, the human has to be in front of the camera at a roughly known position in the image. With this paper we present the results of a project, which actively turns the camera toward human beings, using a 180° laser range scanner, and a color camera both mounted on top of a pan tilt unit. Sequences of laser range scans are segmented and the segments are classified whether they possibly belong to a standing or walking human, resulting in tracked hypotheses of human beings. Color information from the camera is used to classify skin-like image regions. Classifications from both sensors are unreliable, but fusing them together results in a much better performance. The system was demonstrated successfully at an exhibition.

Keywords: Robotics, Image Processing, Laser Range Scanner, Sensor Fusion

1. INTRODUCTION

Service robots are robots which work in an environment not specifically modified to suit the needs of the robot and which is partially unknown. This is often referred to as an *unstructured environment*. Therefore service robots have to use complex sensor systems to interact successfully with their environment. In many cases, service robots will work among human beings. Possible applications range from (relatively) simple floor cleaning tasks [6] to complex scenarios of aiding elderly or handicapped persons [2]. For any interaction with human beings, it is a precondition that the human being is localized and recognized by the robot based on its sensory input [1,3,4].

This paper presents a successful combination of algorithms, which are able to localize a standing or walking human being a few meters around the service robot MAVERIC (see fig.1). MAVERIC

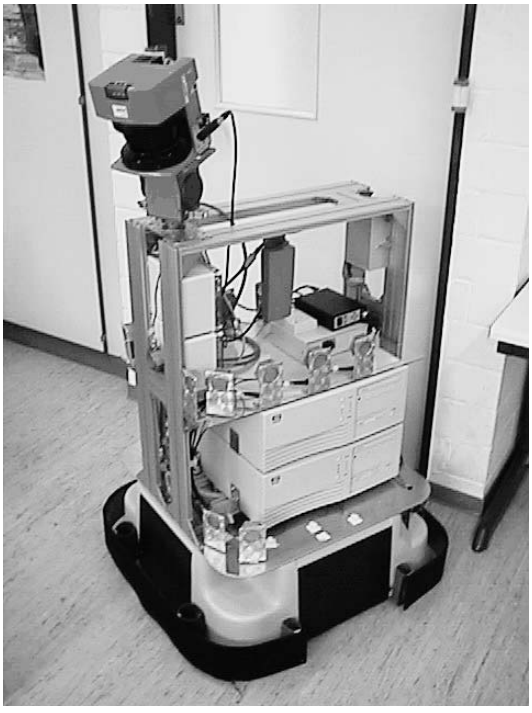

Fig. 1: MAVERIC – Mobile Autonomous Vehicle to Experiment upon Robotic Indoor Chores

Fig. 2: example of a camera image. Top: grayscale image. Bottom: image of fully saturated colors, classified skin regions (white) and the two bounding boxes of skin-colored blobs (rectangles).

has a variety of different sensors: 16 ultrasonic distance sensors, a color CCD video camera, a SICK laser range scanner and two contact sensitive areas (bumpers). For recognizing humans, the laser range scanner and the camera are used. Since both sensors are mounted together on a pan tilt unit, they can be used to actively search and look at humans.

Section 2 presents the algorithm to extract skin-collared regions from the video camera image. Section 3 shows how to segment the laser range scan and how to classify these segments. In section 4 the way to fuse the results from sections 2 and 3 over time is discussed. Section 5 will give an impression of the capabilities and limitations of the implementation on MAVERIC.

2. CAMERA IMAGE

Since detection of faces within a gray-scale image is a complex and time consuming task, only a simple feature is used to detect humans: the pure color of skin. Since black and white are no colors, this detection scheme is tolerant to a surprisingly wide range of skin colors. Blood is always red.

Video images are very sensitive to lighting conditions. Using only color information has the obvious advantage of avoiding any brightness normalizations. The pure color information could be obtained by transforming from RGB to HSI color space, which has the dimensions hue, saturation and intensity, where hue is a value encoding the pure color. But to avoid angular, trigonometric and even floating point computations, the following color normalization is applied, where R, G, B are unsigned 8 bit values, f is an unsigned integer and "<<" and ">>" are C-like bit shift operations:

```
if (max(R,G,B)>thresh_color [1])
   f=(255<<16)/max(R,G,B);
   R=(f*(R-min(R,G,B))>>16;
   G=(f*(G-min(R,G,B))>>16;
   B=(f*(B-min(R,G,B))>>16;
else // no color
   (R,G,B)=(0,0,0); // black
```

The resulting pure color vector has one zero component, and one component of 255[2]. The name of the remaining component indicates on which of the three sectors of the color circle the color is located and the value is monotonic related to the hue values of this sector. Evaluation of several test images with daylight, neon light and halogen light lead to the following classification rule[3] for our DSP Micro Head CCD Color Camera CV-M1250K:

```
if (R ≈ 255 and B = 0 and …
    G ≥ 9 and G ≤ 100)
      (R,G,B)=(255,255,255);
```

[1] $thresh_{color}$ avoids division by zero and defines how saturated a color has to be to be valid. The darker the worst acceptable lighting is, the lower this value should be chosen. We use a value of 7.

[2] Due to computational inaccuracies the actual value sometimes is 254.

[3] Note that changes in blood circulation do change the pure color of the skin too.

Every pixel with pure color from almost red to orange is considered to be skin-like. These pixels become white.

Since color information from the camera is very noisy, the half-image is reduced in width and height by a factor of two before the above transformation is applied. Finally, the resulting image of size 384 by 143 is binarized in white and non-white pixels and a standard blob analysis is done, which finds contiguous areas of white pixels fast. Considering only blobs which have a bounding box wider and higher than 16 pixels and with at least 16^2 pixels reduces remaining noise problems.

Still, of course, these areas in the camera image, represented by their bounding boxes, are not necessarily part of a human face. More features have to be considered, which we will derive from the laser range scan.

3. LASER RANGE SCAN

A laser range scanner is a sensor which measures distances by time of flight of a near infrared laser beam, which is directed by a rotating mirror. The SICK laser range scanner [5] has a reading resolution of 0.5° and about 1 cm at closer ranges. The resulting scan is in polar coordinates. It is transformed to Cartesian coordinates to do distance and shape measurements, but the order of the scanned points is preserved.

A standing or walking human being in the laser range scan has a shape, which is different from e.g. walls or doors. At first, the scan is segmented. Two subsequent points belong to different segments if and only if they are more than a threshold of 15 cm apart. For every segment a number of features is computed and a classification is done. Several features have been computed and evaluated for about 3000 manually labeled segments from test scans. The features finally used are:

points: number of measurement points of the segment.

width: distance in cm between the first and the last point of the segment.

depth: extension in cm of the point cloud of the segment orthogonal to the base line of the segment, which is the line through the first and last point of the segment.

curved: *length*/*width*, where *length* is the length of the line segment, i.e. the sum of the distances of subsequent points.

square: *width*/*depth*; a value of 1.0 means the bounding box of the segment is approximately a square.

linminmax: the segment is split at the point closest to the laser scanner and for both halves the maximal distance in cm of a point from the line connecting start and end point of the half is computed. *linminmax* is the smaller one of these errors.

linminavg: similar to *linminmax*, but instead of the maximal error the average error is used.

ellminmax: similar to *linminmax*, but instead of a line a quarter of an ellipse is fit into each of the half segments. Start and end point of the half segment are on the ellipse; the center and one axis of the ellipse are on the base line.

These features are not sufficient to identify humans reliably. Therefore classification is done into three classes, which are ‘not human’, ‘potentially human’ and ‘probably human’. Classification for the last class was tuned so that a moving human facing the robot typically was classified as ‘probably human’ within about 20 laser range scans, i.e. within about a second.

‘potentially human’ =
$$\begin{aligned} & \textit{points} \in [5, 131] \wedge \\ & \textit{width} \in [8, 100] \wedge \\ & \textit{depth} > 1.5; \end{aligned}$$
‘probably human’ =
$$\begin{aligned} & \textit{points} \in [18, 63] \wedge \\ & \textit{width} \in [10, 70] \wedge \\ & \textit{depth} > 1.5 \wedge \\ & \textit{curved} \in [1.23, 1.88] \wedge \\ & \textit{square} \in [1.75, 5.8] \wedge \\ & \textit{linminmax} \in [3.9, 26.1] \wedge \\ & \textit{linminavg} \in [-8.63, 18.9] \wedge \\ & \textit{ellminmax} \in [0.0266, 0.166]; \end{aligned}$$

Obviously, 'potentially human' is a superset of 'probably human'. For future extensions, it might be useful to split 'probably human' into one class to recognize heads and one class to detect torsos.

4. TRACKING AND SENSOR FUSION

If a human observer analyses recorded laser range scans visually, the observer too has problems to classify scans of humans from a single laser range scan. But from a sequence of laser range scans it is easy to detect a human as soon as it moves. Therefore we decided to implement tracking of 'potentially human' segments. Only those segments are tracked, which at least look slightly as they could belong to a person.

Matching a newly recorded set of segments to the set of tracked segments is done by looking for the closest match, using the position of the middle indexed point as the position of a segment: if a new segment and its closest old segment are less than 50 cm apart, they are identified and the old segment is updated. Else the new segment is inserted into the list of tracked segments. Segments that have not been observed for more than a second are removed from the list of tracked segments. This simple but successful tracking strategy relies on a fast scan rate of about 20 laser range scans per second.

It would have been advantageous to use movement information to finally classify a tracked potentially human laser scan segment as being human. But laser range scans and position readings of the robot base and the pan tilt unit of MAVERIC cannot be done exactly at the same time. This induces an error into the positions in world coordinates computed for the tracked segments. Furthermore, unstable segments – which are classified for tracking in one scan but not in the next scan – are often mismatched to neighboring unstable segments. All these errors look like movements. Since we want to identify

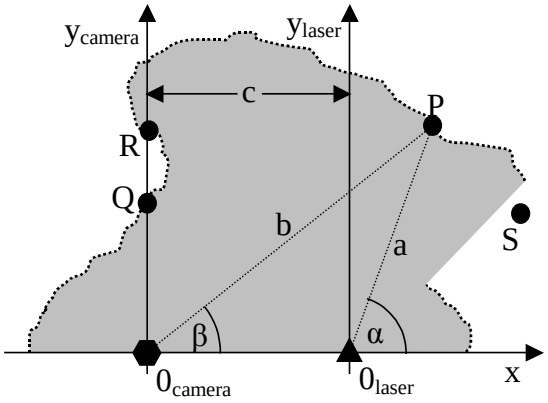

Fig. 3: transforming laser-centric polar coordinates into camera centric coordinates.

marginally moving and standing humans too, this criterion was not used.

To fuse the information from the camera into this model, we have to transform skin area coordinates to world coordinates. Fig. 3 illustrates how this is done. Point P has laser range scanner polar coordinates (a,α), camera polar coordinates (b,β) and Cartesian camera coordinates (x,y), where

$$(x,y) = (a \cos(\alpha) - c , a \sin(\alpha))$$

$$(b, \beta) = (\sqrt{x^2 + y^2} , \text{atan2}(y,x))$$

When transforming every laser range scanner measurement to camera centric coordinates, apart from quantization inaccuracies, special care has to be taken for ambiguities and singularities. Points Q and R are ambiguous; Q is preferred, because it is closer and therefore probably more interesting. Point S is undefined.

From the resulting transformed, camera-centric laser range scan, the distance of a skin like region in the camera image can be deduced in most situations where the region is sufficiently vertically centered in the camera image. The angle β of a skin like region depends on the camera geometry and is approximated by a linear transform of the horizontal pixel distance of the skin like regions center to the center of the image.

MAVERIC tries to point the camera to skin colored regions by using the pan tilt unit. Higher regions are preferred, since the lower regions are probably hands and

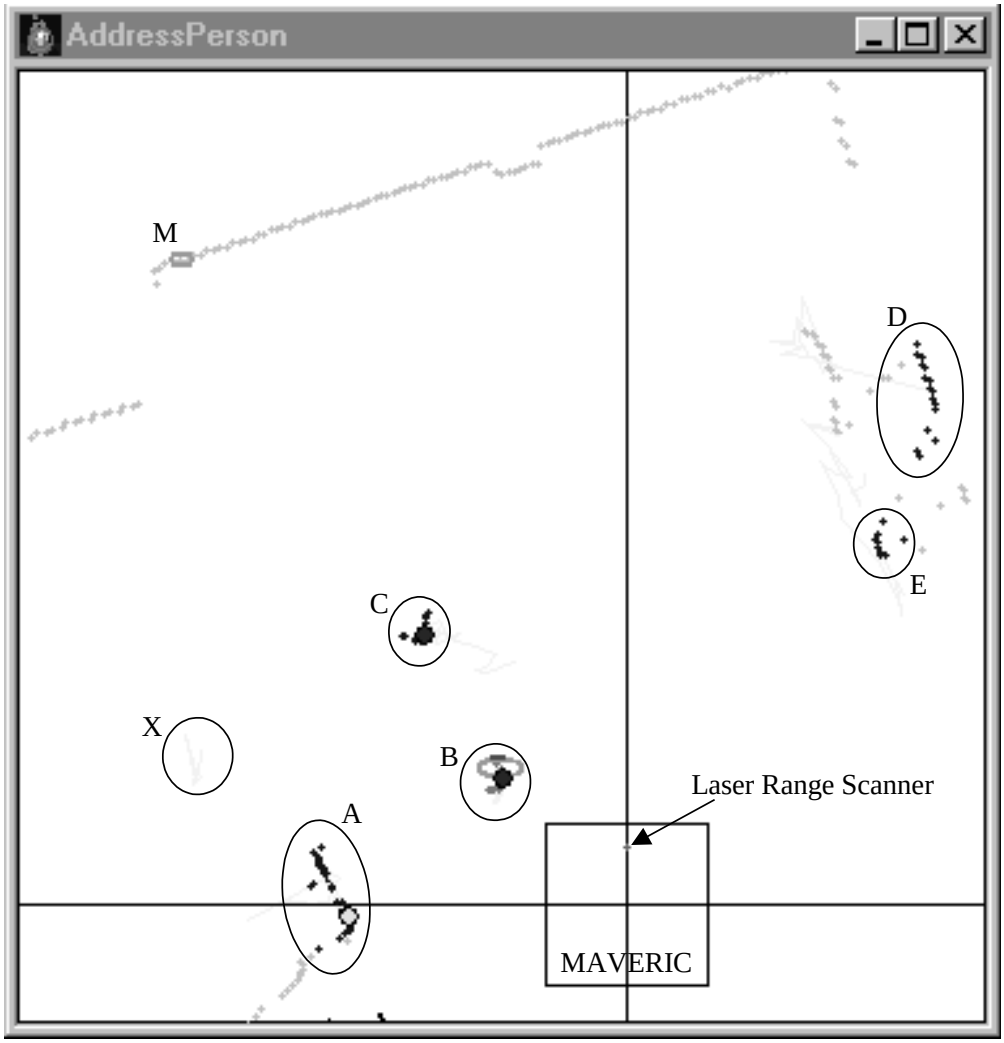

Fig. 4: Visualization of the tracking results corresponding to the camera image of figure 2.

the upper regions are more likely to be faces. If a skin colored region corresponds to a tracked laser range segment, which at least once was classified 'probably human' and which has a world coordinates height of approximately the height of a human face, this segment will be considered a human to address. As soon as the camera is directed toward the person, one of several audio greetings is played and the original camera image is taken as a picture of the person. For viewing, the picture is transferred via wireless network to a PC with a large display.

5. RESULTS

Figure 4 gives an example of the integrated results from tracking laser range scanner segments and skin color region analysis. This integration is done in 3D robot coordinates, from which we see the top-view from above. Height information is only used to exclude unreasonable high or low tracked hypotheses.

The situation of figure 4 corresponds to the camera image of figure 2. The segments marked by B and C are correctly classified 'probably human' (dark big dot). B currently has a matching skin color region hypothesis (ellipse), while the skin color hypotheses of the moving human at C is mismatched to the background at M. Still C had a correct match of a skin colored region in a previous tracking step.

More segments are currently being tracked, marked by A, D, E and X. Of all the segments, only segment B is currently

classified as 'probably human' (brighter color of the laser scan points). The segment at A did have an erroneous classification as 'probably human' once, but never had a matching skin colored region (gray big dot). Bright gray lines indicate movement information of the tracked segments for the last 20 scans, i.e. the last second. The Person at B is standing and the person at A is moving to the left. These interpretations are correct. But movement information for segments A, D, and E are completely due to the errors mentioned in section 4. At X, we see movement information from a segment which is not observed in the current laser range scan, probably because of a changed tilt gaze direction of the pan tilt unit. It will be removed from the list of tracked segments soon.

Figure 5 shows MAVERIC while slowly driving through a crowd of visitors at an exhibition, addressing people. It is surprising how well this set of simple behaviors is suited to attract the attention of passersby.

6. CONCLUSION AND OUTLOOK

A successful implementation of a system using sensor fusion to address people was presented. Although the algorithms used are simple, real sensor inputs generate a variety of behavior that is appealing to the human observer.

Future possible improvements of the systems capabilities might include automatic calibration of the optimal skin color range, which varies slightly with lighting and addressed person, and a more purposeful strategy to search for people. An open question is whether the results would be more reliable, if the modeling and tracking was done only with the angular component of the polar coordinate system of the laser range scanner. Despite the obvious modeling errors, this approach might be more robust to erroneous sensor interpretations and provide better tracking capabilities.

Figure 5: MAVERIC taking photos of and talking to visitors of an exhibition.

7. LITERATURE

[1] T. Darrell et al., "Integrated Person Tracking Using Stereo, Color, and Pattern Detection", Conf. Computer Vision and Pattern Recognition, IEEE CS Press, Los Alamitos, Calif., 1998, pp. 601-608

[2] M. Hans, W. Baum: "Concept of a Hybrid Architecture for Care-O-bot", Proc. of the IEEE International Conference on Robot and Human Interaction, RO-MAN 2001, 18.-21. Sept. 2001, Bordeaux-Paris, France, pp. 407-411, http://www.morpha.de

[3] E. Prassler, J. Scholz, P. Fiorini: "Navigating a robotic wheelchair in a railway station during rush hour", International Journal of Robotics Research 18(7): 711-727, 1999

[4] C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, R. Wörz: "Vision Based Person Tracking with a Mobile Robot", Ninth British Machine Vision Conference, BMVC '98, September '98, Southampton, UK, pp. 418-427

[5] SICK laser range finder: http://www.sick.com, http://195.145.243.169/de/en/products/categories/safety/laserscannerpls/lasers cannerpls.html

[6] Siemens supermarket cleaning robot: http://www.ad.siemens.de/sinas