in: Proc. 52nd Int. Scientific Colloquium (IWK), 2007, Technische Universität Ilmenau, Germany, Vol. I, pp. 59-64, Universitätsverlag 2007

F.-F. Steege / C. Martin / H.-M. Groß

Recent Advances in the Estimation of Pointing Poses on Monocular Images for Human-Robot Interaction

1 Introduction and Motivation

In recent years, a lot of research work has been done to develop intelligent mobile robot systems, which can interact even with non-instructed users, making the robots suitable for applications in everyday life. Besides the verbal communication also the non-verbal communication plays a very important role in a dialog between humans. To the knowledge of the authors, only a few projects have already successfully integrated non-verbal communication parts in an interactive dialog on their mobile robots. In the work presented in this paper, we show how a basic non-verbal communication (more precisely: the problem, of instructing a mobile robot by the use of pointing gestures/poses) can be realized on a mobile robot system.

Some approaches allready exist which focus on integrating gesture recognition into Man-Machine-Interfaces. In the works of Rogalla et al.[1], Paquin and Chohen [2] and Triesch and v.d. Malsburg [3] different approaches to detect and classify human gestures and poses are presented. However, most of this work concentrates on distinguishing different gestures, creating a command alphabet for robot control. A much more intuitive and smoother way to direct the robot is through pointing directly at the target position on the ground. In [4, 5] for the first time we presented an approach, which allows to direct a mobile robot to a certain position by means of such pointing poses. The system presented in [4, 5] was capable of estimating the target point of the pointing gesture on the floor with a low error, but could only operate in environments with unstructured background and ideal lighting conditions. Besides a computation time of 3-4 seconds was required for the estimation of a single target. These constraints conflict with the requirements for the usage of this approach in robotic real world applications. Therefore, in this paper we present several improvements on this approach making it possible to estimate the target point of a pointing pose in highly structured environments with variable lighting conditions with a computation time of only 80ms.

This paper is organized as follows: After this introduction, Section 2 describes the ground truth of the training data and the robot used for the experiments. Section 3 explains, how the pointing poses can be estimated and how our entire system is designed. In Section 4 the experiments and results will be presented. The papers ends with conclusions and a short outlock in Section 5.

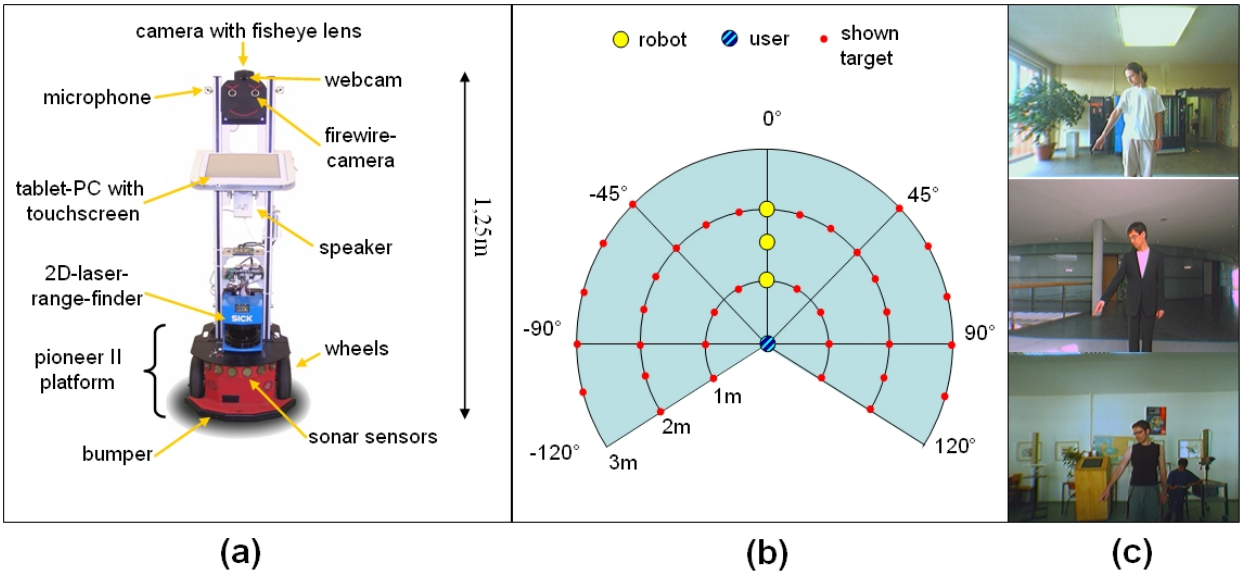

Figure 1: Image (a) displays our robot HOROS used for experimental imvestigation of the pointing pose estimation is displayed. The images for the estimation of the pointing target were taken with the front camera (loacted in the right eye). Image (b) displays the configuration used for recording the ground truth training and test data. The subject stood in front of the robot and pointed at one of the marked targets on the ground. The distance of the robot to the subject varied between 1 m and 2 m. Image (c) displays some examples of pointing poses recorded with the camera of the robot.

2 The Robot HOROS and the Ground Truth

The approach described in this paper was developed and tested on our mobile robot HOROS (**HO**me **RO**bot **S**ystem). HOROS' hardware platform is an extended Pioneer II based robot from ActivMedia (see Fig. 1(a)). Because one objective of our project is the development of a low-cost prototype of a mobile and interactive robot assistant, we are especially interested in vision technologies with a good price-performance ratio. Therefore, the two low-cost frontal cameras were utilized instead of a high-end stereovision system. We were interested if it would be possible to robustly estimate a target position at the floor from a pointing pose using only inexpensive hardware and monocular images.

A labeled set of images of subjects pointing to target points on the floor was required to train the system. We encoded the target points on the floor as (r, φ) coordinates in a subject-centered polar coordinate system (see Fig. 1) and placed the robot with the camera in front of the subjects. Moreover, we limited the valid area for targets to the half space in front of the robot with a value range for r from 1 to 3m and a value range for φ from $-120°$ to $+120°$. Figure 1 shows the configuration we chose for recording the training data. The subjects stood at distances of 1, 1.5 and 2m from the robot. Three concentric circles with radii of 1, 2 and 3m are drawn around the subject, being marked every 15°. The subjects

were asked to point to the markers on the circles in a defined order and an image was recorded each time. All captured images are labeled with distance, radius and angle, thus representing the ground truth used for training and for the comparing experiments with human viewers (see Section 4). This way, we collected a total of 2.340 images of 26 different interaction partners (90 different poses for each subject). This database was divided into a training subset and a validation subset containing two complete pointing series (i.e. two sample sets each containing all possible coordinates (r, φ) present in the training set). The latter was composed from 7 different persons and includes a total of 630 images. This leaves a training set of 19 persons including 1710 samples.

3 Estimation of Pointing Poses

Since the interaction partners standing in front of the camera can have different body height and distance, an algorithm had to be developed that can calculate a normalized region of interest, resulting in similar subimages for subsequent processing. We use an approach suggested by [4, 5] to determine the region of interest (ROI) by using a combination of face-detection (based on the Viola & Jones Detector cascade [9]) and some empirical factors. With the help of a multimodal tracker [4, 5] implemented on our robot, the direction and the distance of the robot to the interacting person can be estimated. The cropped ROI is scaled to 160*100 pixels for the body and the arm and 160*120 pixels for the head of the user. Additionally, a histogram equalization is applied to improve the feature detection under different lighting conditions. The preprocessing operations used to capture and normalize the image are shown in Fig. 2. To reduce the effects of different backgrounds, in the improved version we used a simple Background Subtraction algorithm and tested its influence on the pose estimation result in comparsion with our approach in [4, 5] where no Background Subtraction was used. On the normalised image regions a feature extraction is used for the approximation of the target position the user is pointing to. In our work Gaborfilters of different orientations and frequencies, bundled in Gaborjets that are located on several fixed points in the selected ROIs, are used. The several steps of preprocessing and feature extraction used in our comparison are shown in Fig. 2.

In [4, 5] a cascade of several Multi-Layer Perceptrons (MLP) was used to estimate the target point from the extracted features. However other techniques are also often used for the estimation of certain human poses (but mostly not on mobile robots but under predefined observation conditions). Nölker and Ritter [10] used a Local Linear Map (LLM) and a Parametrized Self-Organizing Map (PSOM) to estimate the target of a pointing pose on a screen the user is pointing to. Krüger and Sommer [7] utilized Gaborfilters and a LLM to estimate the head pose, while Stiefelhagen [8] presented a system that works on edge-filtered images and uses a MLP for head pose estimation. Therefore, for this paper we implemented and compared several selected neural approaches, which all were trained and tested with the same sets of training and test data. This way we are able to give an

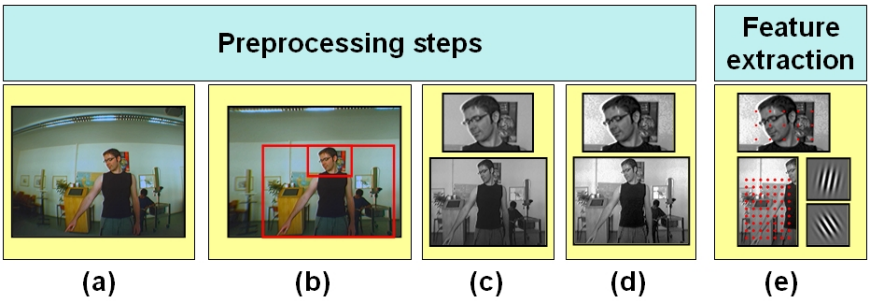

Figure 2: Steps of preprocessing and feature extraction: the raw distorted image of the lowcost camera in the robot's eye (a) is transformed into an undistorted image and the face of the user is detected by means of [9] (b). Based on the height of the face in the picture and the distance of the user, two sections of the image are extracted and transformed into grayscale images (c). On these images a histogram equalization is used (d). Subsequently, features are extracted by Gaborfilters placed at pre-defined points of the image (marked as red dots in (e)). A Background Subtraction can optionally be used between steps (d) and (e).

overview of the suitability of the different approaches for the task of estimating a pointing pose on a monocular image. We compared a k-Nearest-Neighbour method (kNN), a Neural Gas network (NG, [11]), a Self-Organizing Map (SOM, [12]), a Local Linear Map (LLM, [13]) and Multi-Layer Perceptrons (MLP, [4, 5]).

4 Experiments and Results

To have a simple reference for the quality of the estimation, 10 human subjects were asked to estimate the target point of a pointing pose on the floor. At first, the subjects had to estimate the target on a computer screen where the images of the training data set were displayed. The subject had to click on the screen at the point where they estimated the target. Thus, the subjects were estimating the target on the images having the same conditions as the different estimation systems. Second, we determined the estimation result the subjects achieved under real world cirumstances. Here, each subject had to point at a target on the ground and a second subject had to estimate the target. At first the recognizing person used both of their eyes to estimate the target, later we blindfolded one of the eyes and the person estimated the target again under monocular conditions. The results of the human based reference experiments are included in Fig. 3. The label *Human (srceen)* refers to the experiments on the computer screen and the labels *Human (2 eyes)* and *Human (1 eye)* refer to the results under real world conditions.

The results of the several approaches for estimating the target position are shown in Fig. 3. As described in Sect. 2 the ground truth data is a tuple (r, φ) with the target radius r and the target angle φ. For the correct estimation of the target point, r as well as φ had to be estimated correctly. We defined the estimation result to be correct if r differed less than 50cm from the ground truth radius and φ differed less than 10° from the ground truth

target point estimation (correct radius *and* correct angle)					
correct samples in %	**k-NN**	**NG**	**SOM**	**LLM**	**MLP**
Gaborfilters	11,12%	4,70%	6,70%	11,76%	29,16%
Gaborfilters and BG Subtraction (BGS)	22,28%	17,72%	15,34%	23,53%	44,90%
Gaborfilters and Discriminant Analysis	17,69%	9,38%	11,66%	16,04%	28,14%
Gaborfilters, BGS and Discriminant Analysis	34,72%	22,66%	23,66%	31,74%	50,63%

Human (2 eyes)
62,90 %

Human (1 eye)
40,75 %

Human (screen)
37,50 %

Figure 3: The results for the estimation of the target point of the pointing pose. The target point is determined by the radius r and the angle φ. For each method the percentage of the targets estimated correctly is determined. The results of the human viewers (on computer screen, and in reality (with both eyes "Human (2 eyes)" and with one eye blindfolded "Human (1 eye)")) are given for comparison. Methods that achieve a result equal to that of the human viewers are marked with a shaded background with different colors.

angle. Figure 3 shows the results for a correct estimation of both values.
Every of the five selected approaches was trained and tested on the same training data set. For each system, we used four different feature extraction strategies: first only Gaborfilters were utilized, second we combined Gaborfilters with an additional Background Subtraction to reduce the effects of the different cluttered backgrounds in the images. Third, we used only those Gaborfilters that had a high discriminant value extracted by means of a Discriminant Analysis executed over all predefined Gaborfilter positions. Fourth, we combined Gaborfilter, Background Subtraction and utilized only the relevant features extracted by the Discriminant Analysis mentioned above.
The results demonstrate, that a cascade of several MLPs as proposed in [4, 5] is best suited to estimate the target position of a user's pointing pose on monocular images. A Background Subtraction and the information delivered by a Discriminant Analysis can be used to improve the results. The best system is capable of estimating r as good as humans with their binocular vision system in a real world environment and even better than humans estimating the target on 2D screens. The estimation of φ does not reach equally good values. The system is able to reach a result equally to humans on 2D screens or humans with one eye blindfolded, but it is not able to estimate the angle as good as humans in a real world setting using both eyes. This is because the estimation of the depth of a target in a monocular image is difficult for both, human and function approximators.

5 Conclusion and Outlook

In this paper we presented an extension to our approach introduced in [4, 5]. The major

problems of the old approach (bad results in environment with structured background and a computation time which exceeds real-time requirements) could be solved. Extensive experiments with different neural function approximators have shown, that the MLP-based approximator leads to the best result. The realized approach is able to estimate a referred position on the ground based on monocular images with an accuracy nearly equal to humans and work in real-time (80ms). This enables the user to direct a mobile robot system into a target position based on pointing gestures only.

References

[1] Rogalla, O. and Ehrenmann, M. and Zöllner, R. and Becher, R. and Dillmann, R.: *Using Gesture and Speech Control for Commanding a Robot Assistant.* In In Proc. of the 11th IEEE Int. Workshop on Robot and Human Interactive Communication, ROMAN. (2002) 454–459.

[2] Paquin, V. and Cohen, P.: *A Vision-Based Gestural Guidance Interface for Mobile Robotic Platforms.* In Proc. of the Workshop on HCI, Computer Vision in Human-Computer Interaction, ECCV (2004) 39–47.

[3] Triesch, J. and von der Malsburg, C.: *A System for Person-Independent Hand Posture Recognition against Complex Backgrounds* In IEEE Trans. Pattern Anal. Mach. Intell, Vol. 23, Number 12 (2001) 1449–1453.

[4] Gross, H.-M. and Richarz, J. and Müller, S. and Scheidig, A. and Martin, C.: *Probabilistic Multi-modal People Tracker and Monocular Pointing Pose Estimator for Visual Instruction of Mobile Robot Assistants.* In: Proc. of the IEEE World Congress on Computational Intelligence, WCCI (2006) 8325–8333.

[5] Richarz, J., Martin, C., Scheidig, A., Gross, H-M.: *There You Go! - Estimation Pointing Gestures in Monocular Images for Mobile Robot Instruction.* In: RO-MAN 2006 - 15th IEEE Int. Symposium on Robot and Human Interactive Communication (2006) 546–551.

[6] Takahashi, K., Tanigawa, T.: *Remarks on real-time human posture estimation from silhouette image using neural network.* In: Proceedings of the International Conference on Systems, Man and Cybernetics: The Hague (2004) 370–375.

[7] Krüger, V., Sommer, G.: *Gabor wavelet networks for efficient head pose estimation.* In: Image and Vision Computing, Vol. 20, Number 9-10, August (2002) 665–672.

[8] Stiefelhagen, R.: *Estimating Head Pose with Neural Networks - Results on the Pointing04 ICPR Workshop Evaluation.* In: Pointing 04 ICPR Workshop, Cambridge, UK (2004).

[9] Viola, P. A., Jones, M. J.: *Rapid object detection using a boosted cascade of simple features.* In: Proc. of the Conf. on Computer Vision and Patter Recognition, Munich, Germany (2001) 511–518.

[10] Nölker, C., Ritter, H.: *Illumination Independent Recognition of Deictic Arm Postures.* In: Proc. of the 24th Annual Conference of the IEEE Industrial Electronics Society, Aachen, Germany (1998) 2006–2011.

[11] Martinetz, T., Schulten, K.: *A Neural-Gas Network Learns Topologies.* In: Proc. of the the ICANN 1991, Helsinki, Finnland (1991) 397–402.

[12] Kohonen, T.: *Self-organized formation of topologically correct feature maps.* In: Biological Cybernetics, Vol. 43 (1982) 59–69.

[13] Ritter, H.: *Learning with the Self-Organizing Map, Eds.: T. Kohonen et al., Artifical Neural Networks.* (1991)

Author Information:

Dipl.-Inf. Frank-Florian Steege

Dipl.-Inf. Christian Martin

Prof. Dr.-Ing. Horst Michael Groß

Neuroinformatics and Cognitive Robotics Lab, Ilmenau Technical University

G-Kirchhof-Str. 2, 98693 Ilmenau, Germany

Tel: ++49 +3677 69 2858

Fax: ++49 +3677 69 1665

E-mail: frank-florian.steege@stud.tu-ilmenau.de, christian.martin@tu-ilmenau.de