

# Image Processing Methods for Interactive Robot Control

Christoph Theis<sup>1</sup>, Ioannis Iossifidis<sup>2</sup> and Axel Steinhage<sup>3</sup>

<sup>1,2</sup> Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany

<sup>3</sup> Infineon Technologies AG, Corporate Research, Systems Technology, München, Germany

<sup>1</sup> Phone +49 (0)234 32 27969, Email: Christoph.Theis@neuroinformatik.ruhr-uni-bochum.de

<sup>2</sup> Phone +49 (0)234 32 25567, Email: Ioannis.Iossifidis@neuroinformatik.ruhr-uni-bochum.de

<sup>3</sup> Phone +49 (0)89 234 55181, EMail: Axel.Steinhage@infineon.com

## Abstract

*In this paper we describe a straight forward technique for tracking a human hand based on images acquired by an active stereo camera system. We demonstrate the implementation of this method on an anthropomorphic assistance robot as part of a multi-modal man-machine interaction system: detecting the hand-position, the robot can interpret a human pointing gesture as the specification of a target object to grasp.*

Keywords *image processing, man-machine interaction, robot control, tracking*

## 1 Introduction

In the field of industrial robotics, the interaction between man and machine typically consists of programming and maintaining the machine by the human operator. For safety reasons, a direct contact between the working robot and the human has to be prevented. As long as the robots act out preprogrammed behaviors only, a direct interaction between man and machine is not necessary anyway. However, if the robot is to assist a human e.g. in a complex assembly task, it is necessary to have means of exchanging information about the current scenario between man and machine in real time. For this purpose, the classical computer devices like keyboard, mouse and monitor are not the best choice as they require an encoding and decoding of information: if, for instance, the human operator wants the robot to grasp an object, he would have to type in the object's coordinates (if these are known at all) or move the mouse pointer to an image of the object on a computer screen to specify it. This way of transmitting information to the machine is not only unnatural but also error prone. If the robot is equipped with a camera system, it would be much more intuitive to just point to the object to grasp and let the robot detect its position visually. Observing two humans in the same situation reveals another interesting effect: by detecting the partner's gaze direction the person who points to an object can immediately control whether

his intention has been interpreted correctly. If the partner looks at the wrong object, this becomes obvious immediately. Therefore, the movement of the head fulfills two functions: first, it is an efficient exploitation of the sensor equipment by shifting the interesting objects into the focus of view. Second, it can be used as a communication channel to provide information about the current behavioral state. In a robot system, this function can be implemented by providing the robot with a dynamic camera head that actively tracks the human hand position. To guarantee a smooth interaction between man and machine a task like this requires that the visual processing, the transmission of the position information to the camera mechanics and the movement of the camera head itself are very fast. In the following, we will describe a system which fulfills these requirements (compare with [1]). Before we go into details about the vision processing methods, we shortly describe our anthropomorphic assistance robot CORA on which we have implemented our method.

## 2 The Anthropomorphic Assistance Robot Cora

CORA(=Cooperative Robot Assistant, see Fig. 1) was built as a prototype of a service robot system to assist a human partner in industrial assembly tasks. The research on this platform ranges from investigating the behavioral aspects of man-machine interaction and the representation and interpretation of scenes to the generation and organization of behavior.

In our opinion successful man-machine interaction requires that both, the robot and the human partner, possess a similar sensor- and effector equipment. Therefore, we designed CORA anthropomorphic: the seven DoF manipulator is connected to a one DoF trunk which is fixed on the edge of a table. Above the trunk we assembled a two DoF pan/tilt unit carrying a stereo color camera system and microphones.

By turning the trunk joint, the robot can change its configuration from left to right handed. Two of the manip-

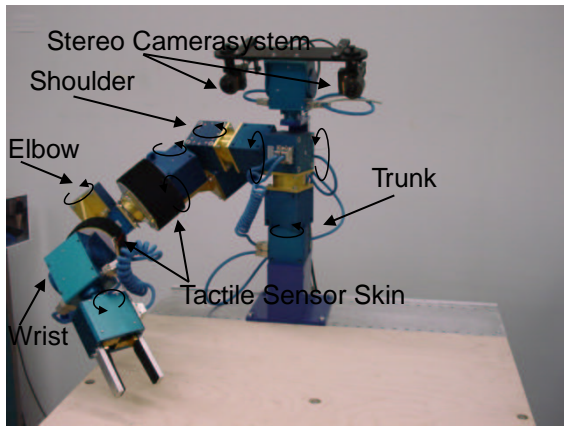


Figure 1: The service- and assistance robot CORA. A seven DoF manipulator arm is mounted on a one DoF trunk which is fixed on a table. The robot possesses a two DoF stereo camera head with microphones.

ulator arm modules are covered with a touch-sensitive so-called *artificial skin*. When a human partner is sitting at the opposite side of the table, the robot and the human partner share the same eye-level. Relying on the stereo camera, the microphones and the artificial skin, CORA uses similar sensor channels as those available to the human partner. The restriction to audio, vision and touch and the redundant configuration of the arm sets high demands on the control structure of the robot. The goal of our research is a robot system that can perform the following tasks: a) visual identification of objects presented by a human teacher, b) recognition of these objects in a cluttered scene, c) grasping and handing over objects, d) acting out simple assembly tasks. All behaviors should be realized under visual, acoustical and haptic control by the human partner. This user interaction consists of pointing gestures, speech commands and corrections of the manipulators configuration via touch.

In many applications specifically optimized vision processing methods for the detection of the human hand from camera images have been developed. One possibility is to detect the hand by means of its movement in the image. Under the assumption that the rest of the scene is static, this method enables a fast tracking. However, in the man-machine interaction task, the assumption of a static scene does not hold: grasped objects, the arm of the operator and even the robot manipulator moves within the scene. In addition, we want to be able to detect a static pointing hand too and to move the head at any time.

A second possibility is to compare the actual image with elements from a model database. By storing models of

hands in different poses, even a hand gesture recognition can be achieved by this method. However, to be flexible and robust, this approach requires either that a large number of different hand models are stored or that a complex visual pre-processing reduces the variability of the image to gain a prototypic representation of the hand. For our purpose, this model based method is not appropriate for the following reasons: First, we do not need to recognize hand poses but just the general pointing direction. Second, we cannot afford a time consuming pre-processing as the interaction with the human happens in real time. Third, we would like to recognize any human hand without the need to have a model of this hand in the database before.

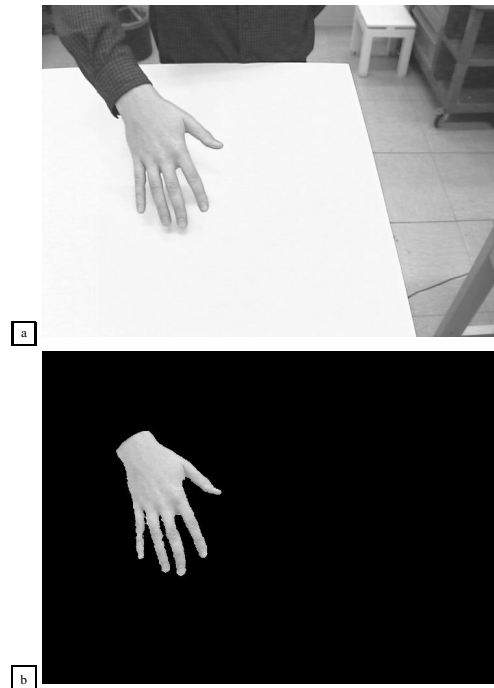


Figure 2: Figure (a) shows the human hand in a typical calibration phase. In figure (b) only that region got left, which represents the human skin color. These color samples are used for further operations, the calibration is complete.

### 3 Hand detection

Keeping all this in mind, we have decided to implement another very simple recognition method: we detect the hand on the basis of its typical skin color. In most situations, this parameter turned out to be sufficiently robust for our purpose. However, we have implemented a dynamical initialization phase in which the robot can "learn" the skin color of the specific operator (see Fig. 2). In the following, we will describe the recognition method in detail.

### 3.1 Skin Color Segmentation

Our purpose is to segment the quantized color image according to skin color characteristics using the HSI (Hue, Saturation, Intensity) color model. Skin color patches have been used in order to approximate the skin color subspace. Evaluating several skin color samples, covering a large range of skin color appearances (different skin colors, different lighting conditions) one can observe that skin color samples form a quite compact cluster in the HSI color space.

In order to locate the user's hand, skin color areas must be segmented and classified. Their position in space are calculated by means of a stereo-algorithm from their disparity between the two camera images. Using our knowledge about the restricted workspace of CORA, we make the assumption that one of the biggest skin colored clusters nearest to the cameras must belong to the user's hand. Thereby, the disparity of a few points of this cluster are sufficient to approximate the hand's mean distance.

### 3.2 Stereo-based Localization

Our test scenario aims at an assembly situation in which man and machine should solve a given task sharing the same workspace is located between the robot and the human.

After the localization of the skin color cluster belonging to the operators hand by selecting the nearest cluster to the cameras with a sufficient size (see Fig. 3), we calculate the disparity of the pixel area at the bottom of this preselected hand cluster. This area typically represents the fingertip of the user for a normal pointing gesture in this environment. The disparity provides the 3-D position by a simple transformation. This limitation on such a small pixel area enables a very fast determination of the hand position.

### 3.3 Tracking

By restricting the time consuming calculation of the disparity on this single point, the described method for determining the hand localization is very fast. Therefore we can obtain a new hand position so frequently that a fast tracking of the hand is possible (for fast trackers see [4] and [2]). By a simple coordinate transformation we translate the 3-D coordinate of the finger-tip into a 2-D pan-tilt angle for the motors of the stereo camera head such that the gaze direction of the robot follows the current hand position. As stated in the introduction, this behavior allows for a very natural interactive control of the detected hand position. During the tracking of the hand, human and robot become a coupled system. If the operator recognizes that the robot has lost track of the hand, he can simply move the hand into the focus of view so that the tracking mechanism "snaps in" again.

## 4 Object recognition

The purpose of hand-position-recognition is the specification of an object in the workspace by pointing to it. The detection of the hand position serves as a pre-selection of the *region of interest* (ROI) in the image. In our case, the ROI is defined as a triangular area in the bird's eye view of the table which covers a sector with  $30^\circ$  opening angle in front of the fingertip. Within this sector the object which is nearest to the fingertip is considered to be specified by the user's pointing gesture. The specification might be supported by a verbal command to avoid ambiguities. At this point the robot has simply to detect or, if there are further ambiguities, to recognize the specified object.

### 4.1 Learning phase

First, the robot must be enabled to recognize a specified object. Therefore, the robot must learn some of the object's special features. In our case these features are the color distribution and the calculated bird's eye view of the object. To obtain these features, all components of the image which are not part of the table are masked. This is done by a simple geometric calculation on the basis of the known spatial relation between the camera head and the table. In a second step, the color of the table is determined by assuming the table to represent the biggest cluster in the image. Image areas with similar colors are also masked as we assume them to belong to the table's surface.

Thereby the scene gets restricted to its substantial part. On the basis of these limited data a color analysis and a determination of the disparity of each pixel follows. The color analysis provides the characteristic color values of the object within the HSI space. The disparity values provide enough data to calculate the three dimensional position of each pixel. These positions are transformed into a view of the scene from above. This bird's eye view is used as an object representation which is independent of the direction of view of the cameras and does not change its size, according to different distances. Based on this representation, the object can be found in more complex scenes later on (compare with [9] and [8]).

### 4.2 Searching phase

In the searching phase the learned object characteristics enable a successful search for this object. Initially the scene gets reduced to its objects as done in the learning phase.

At this time a search for the disparities could already take place, but due to the fact that this search is very timeconsuming, we implemented a data-reduction in a pre-processing step. This gets accomplished by using the color values, already known from the learning phase, to mask the extracted objects. Only areas with object clusters that are sufficiently well filled with the learned color values are used in the further image processing. The calculation of the



Figure 3: Figure (a) shows a typical situation during a man machine interaction. By using the color samples, acquired during the calibration phase, we get all regions, which possibly are skin areas (figure (b)). Finally we check the median disparity of adequate big clusters, to get the pointing hand, which should be the nearest cluster to the camera (figure (c)).

bird's eye view on the basis of the disparity information is again the same as in the learning phase.

At this time the bird's eye view of the learned object and the bird's eye view of the scene are present. In a process, similar to the Hausdorff Tracker described in [5], the bird's eye view gets mutually correlated with a blurred copy of the bird's eye views of the scene. The maximum of this correlation, verified by an inverse correlation, specifies the most probable position of the object in the scene. Because the object can be rotated in the two dimensional plane, the correlation must be accomplished for a number of possible rotation angles. In the application on CORA rotation steps between  $15^\circ$  and  $30^\circ$ , depending on the object size, turned out to be sufficient.

## 5 Grasping

As soon as the position of the object is known, its 3-d coordinates and orientation, calculated by a Hough transformation, are transmitted to the manipulator control to initiate the grasping process. The generation of the grasping trajectory is a complex problem since CORA's manipulator has seven degrees of freedom like the human arm. As only six degrees of freedom are needed to specify the position and orientation of the end-effector, the additional joint angle introduces a redundancy into the kinematic problem. For detailed information we would like to direct the reader to [6] [3].

## 6 Integration of other sources of information

The major goal of our research is the design of a robot system which does not only have an anthropomorphic body structure but which is able to communicate with the operator on the basis of natural communication channels. Recognizing the human's hand position to identify a target object for grasping is only one of these channels. Within the current project we design several other means of man-machine interaction. We have, for instance, implemented a module for the recognition of the human's gaze direction. This module can identify the position of the human's eyes so that a rough estimate of the focus of attention is possible. In turn, the operator can see the gaze direction of the robot head so that a natural way of identifying the current region of interest on the table is possible.

In addition we have built a speech recognition system which can identify spoken keyword commands [7]. By means of a speech synthesizer, the robot can express its behavioral state. On the basis of natural commands it is possible to guide the robot's manipulator to a certain target or to terminate an incorrect behavior (such as the selection of a wrong object) by a spoken command.

Another communication channel is based on touch. We used a touch sensitive *artificial skin* by means of which

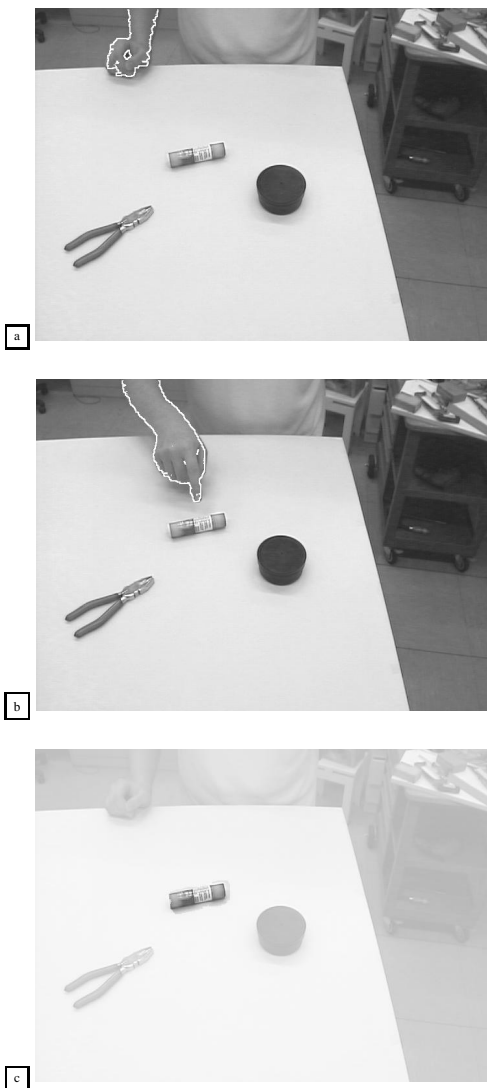


Figure 4: The operator determines, by pointing, the search-region for the selected object: In (a) the initial scene is shown. In (b) the operator points to the object and finally in (c) the detected object is highlighted.

the operator can correct the posture of the robot arm. In addition, unintended contact with obstacles can be avoided.

The goal of our research is the integration of all these different communication channels in a complete man-machine interaction scheme. Within this scheme, the redundancy of the different channels is transformed into accuracy. For example, the operator can specify an object by simultaneously pointing and looking to it and calling out the object's name. In effect, the combination of these different sources of information will enable the robot to overcome errors in single sensor channels.

## 7 Results

The image processing method described so far has been used to track the operator's hand and to extract and interpret the pointing gesture. In this section we present our results by means of figure 4.

In figure 4 (a) the initial scene with three objects is shown and the operator's hand has been detected. In (b) the moving hand is tracked and the pointing direction is extracted. This pointing direction of the operator's hand serves as a pre-selection of the *region of interest* in the image. In (c) the highlighted object is detected as the only possible selection. For the case that two or more objects are possible candidates, the robot can react by asking for more information or simply grasp that object which is the only one known to him.

## 8 Conclusion and Outlook

We have presented a basic concept for a vision based interactive robot control as a part of multi-modal man-machine interaction system. Implementing a straight forward technique for tracking a human hand and extracting a pointing gesture we could demonstrate how man and machine become a coupled system in a very natural way.

The object recognition system can use those pointing gestures to minimize the searching area or, in the common case, can search the whole scene for a special object. Thereby the object's position and orientation is extracted very robustly.

By using multiple channels of information, the whole system is able to overcome ambiguities and the robot can react in an adequate matter.

For the future we plan to extend the multi-modal man-machine interaction system by integrating recognition of the human's direction of view.

## Acknowledgment

This work is supported by the BMBF grant MORPHA (925 52 12).

## References

- [1] T. Bergener and P. Dahm. A framework for dynamic man-machine interaction implemented on an

autonomous mobile robot. In *Proceedings of the IEEE International Symposium on Industrial Electronics, ISIE'97*, 1997.

- [2] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen. Walking Pedestrian Recognition. In *ITSC*, pages 292–297, Tokyo, Japan, 1999. IEEE.
- [3] Percy Dahm. *Beiträge zu Anthropomorphen Robotorarmen (Ph.D. thesis)*. Ruhr Universität Bochum, Institut für Neuroinformatik, Lehrstuhl für Theoretische Biologie, 1999.
- [4] U. Handmann, T. Kalinke, C. Tzomakas, M. Werner, and W. von Seelen. An Image Processing System for Driver Assistance. *Image and Vision Computing (Elsevier)*, 18(5):367 – 376, 2000.
- [5] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans. on PAMI*, PAMI-15-9:850–863, 1993.
- [6] I. Iossifidis and A. Steinhage. Control of an 8 dof manipulator by means of neural fields. In *FSR2001, International Conference on Field and Service Robotics*, Helsinki, Finland, 2001.
- [7] Rainer Menzner. *A Unified Architecture for Speech-Controlled Robot Behavior Based on Nonlinear Dynamics*. ibidem-Verlag, 2001.
- [8] Michael Schwarzinger and Detlev Noll. Object recognition with constrained elastic models. *Mathematical and Computer Modelling*, pages 163–184, 1995.
- [9] Michael Schwarzinger, Detlev Noll, and Werner von Seelen. *Object Recognition with Deformable Models using Constrained Elastic Nets*, pages 96–104. Springer Verlag, 1992.