Integration of Gaze and Gesture Detection in Nature Language Instructing of Robot in an Assembly Scenario

Jianwei Zhang, Tim Baier and Markus Hueser
Department of Computer Science, University of Hamburg, 22527 Hamburg, Germany
E-mail zhang|tbaier|mhueser@informatik.uni-hamburg.de

Abstract

We present the development of and experiment with a robot system showing multimodal interaction capabilities. We focus on the understanding human instructions in natural language by integrating gaze and pointing hand gestures. A typical application of such a system is interactive assembly. A human communicator sharing a view of the assembly scenario with the robot instructs the latter by speaking to it in the same way that he would communicate with a child. His instructions can be under-specified, incomplete and/or context-dependent. After introducing the general purpose of our project, we present the hardware and software components of our system necessary for interactive assembly tasks. Finally, we outline a list of future research topics for extending our research.

1 Introduction

Human-beings interact with each other in a multimodal way. With the enhancement of robot intelligence and advance of human perception, human-robot interaction can be developed naturally and *inter-human like*. A user can instruct a robot by using natural language (NL), gesture and gaze information in the way he communicates with a human partner. Besides understanding the unconstrained natural language, tracking parts of a human body, especially hands, is one important part in human-robot interaction.

One reason for integrating natural communication in a robot control system is that fully automatic assembly under diverse uncertain conditions can rarely be realized without any failure. Several projects on communicative agents realized with real robots have been reported, e.g. [8]. In the projects described in [1] and [9], natural language interfaces were used as the "front-end" of an autonomous robot. In the SAIL project [9], level-based AA-learning combined with attention-selection and reinforcement signals was introduced to let a mobile robot learn to navigate and to recognize human faces and simple speech inputs. In [7], the main system architectures were compared, and an object-based approach was proposed to help manage the complexity of intelligent machine development. In the Cog project [3], the sensory and motor systems of a humanoid robot and the implemented active sensing and social behaviors were studied. To overcome the limitations of this approach, the concept of the "Artificial Communicator" was developed, [5] and [10].

2 The Situated Artificial Communicator

As a basic scenario, the assembly procedure of a toy aircraft (constructed with "Baufix" parts) is selected. We have been developing a two-arm robotic system to model and realize human sensorimotor skills for performing assembly tasks and to facilitate human interaction with language and gestures. This robotic system serves as the major test-bed of the on-going interdisciplinary research program of the project SFB[1] 360 "Situated Artificial Communicators" at the University of Bielefeld. A number of parts must be recognized, manipulated and built together to construct the model aircraft. Within the framework of the SFB, in each of these steps, a human communicator instructs the robot, which implies that the interaction between them plays an important role in the whole process. One new function of a stereo vision system is to trace the hand and gaze direction of the human instructor. It is non-contact, passive, robust, accurate, low-cost consumer hardware and no need of vision precalculations.

The Situated Artificial Communicator and the human instructor interact through natural speech and with hand gestures. First, an instruction is spoken to the robot system and recognized with the *ViaVoice* speech engine. In the current system, *ViaVoice* recognizes only sentences, which the grammar we developed allows. In practice, hundreds of grammar rules can be

[1] Collaborative research unit funded by the Deutsche Forschungsgemeinschaft (DFG).

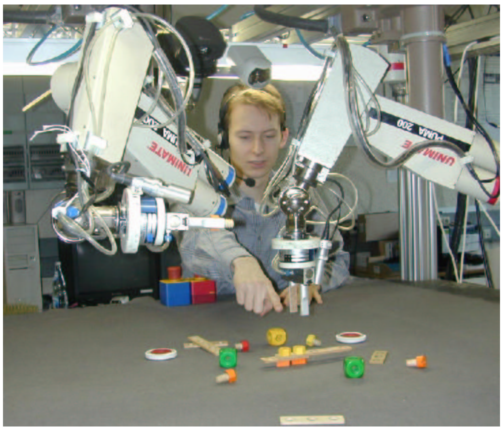


Figure 1: The two-arm multisensor robot system for dialogue-guided assembly.

used. If the recognition succeeds, the results are forwarded to the speech recognition/understanding module.

By their very nature, human instructions are situated, ambiguous, and frequently incomplete. In most cases, however, the semantic analysis of such utterances will result in sensible operations. An example is the command "*Grasp that left screw*". The system has to identify the operation (*grasp*), the object for this operation (*screw*), and the situated specification of the objects (*left*). With the help of a hand gesture and gaze the human instructor can further disambiguate the object. The system may then use the geometric knowledge of the world to identify the right object. Other situated examples are: "*Insert in the hole above*", "*Screw the bar on the downside in the same way as on the upside*", "*Put that there*", "*Rotate slightly further to the right*", "*Do it again*", etc.

3 Kalman-Templates Based Tracking of Head-Motion and Gaze

The gaze detecting system provides real-time determination of head motion and viewing direction of the human instructor (Fig. 3).

The algorithm depends on a 3D facial feature model. The facial features are represented by templates (Fig. 3). The feature tracking of these features is based on an improved template tracking algorithm. First of all a Kalman filter is assigned to each template. Secondly we developed an algorithm to classify wrongly tracked templates. Then we use a Multivariate Least Squares approach for ultra fast model fitting. This gives the up-to-date position and rotation parameters of a tracked human head at each time. Finally we propose an algorithm to recover/reconstruct the correct position of lost or wrongly tracked templates. Based on this part the computation of gaze is reduced to a simple model based region-growing algorithm. Fig. 5 gives an overview of the gaze detection software system.

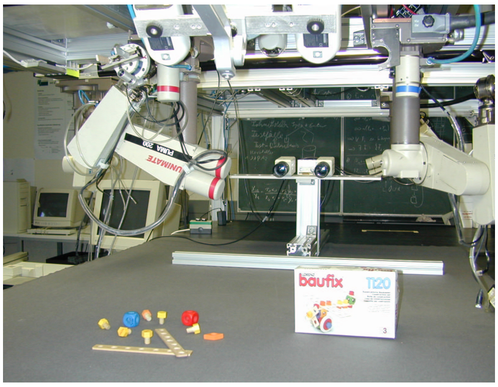


Figure 2: The stereo vision system installed in the assembly cell.

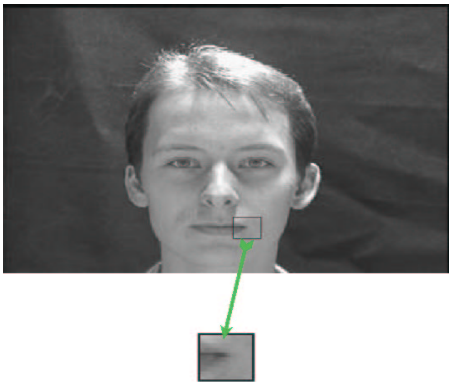


Figure 3: Example of selecting a template.

3.1 Kalman-Templates

The template-based tracking approach was first proposed by [6]. By this approach the tracking is performed without predicting the template's position. Therefore its algorithm needs long time for matching templates with the entire image. Another problem is that it possesses no classifier for determining wrongly tracked templates, but a model fitting using a weighting factor for the quality of each measurement. To reduce the amount of matching operations and to improve the accuracy of the template matching process we attach a Kalman filter to each template. The input for each Kalman filter is the position of the best match of the corresponding template at time k. The output is a prediction of template's position at the next time

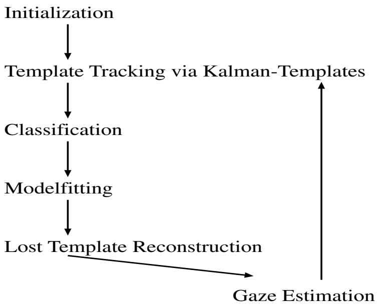

Initialization
Template Tracking via Kalman-Templates
Classification
Modelfitting
Lost Template Reconstruction
Gaze Estimation

Figure 4: The processing steps for gaze detection.

step $k+1$. This new position determines the center of a small search window, which limits the possible positions of the template's best match at the next step $k+1$. Therefore, a large part of wrong positions is excluded from all possibilities and thus the accuracy of template matching is improved. Especially in situations where the original feature searched for is distorted because of rotation, translation or scaling, the improvement in accuracy is significant. Our approach brings the following two advantages: a). no image preprocessing with a gauss like filter is necessary because of this high accuracy; b). the speed is enormously increased because of the much more smaller size of the search window in contrast to the entire image.

3.2 Classification of Wrongly Tracked Templates

To classify templates as wrongly tracked a constraint satisfaction problem (CSP) is defined. At initialization time a model of the observed $3D$ head is constructed, which measures all geometric distances of the observed facial features which are different in pairs, Fig. 6. Every distance defines one constraint.

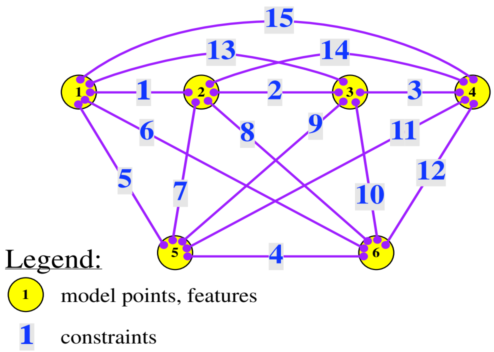


Figure 5: The observed constraints of facial features.

If one or more templates are not tracked correctly, the distances between some templates change in certain way. Based on this observation a classifying algorithm is proposed to determine the wrongly tracked templates. This information is very important because it helps avoiding incorrect data for the model fitting algorithm and identifying which templates have to be reconstructed after model fitting.

3.3 3D Model Fitting by Multivariate Least Squares

Obtaining the best estimate of head position and rotation can be defined as a problem of determine the rotation matrix R and the translation vector t, which minimize the squared error e:

$$e(R,\vec{t}) = \sum_{i=1}^{n} \left(\left\| v_i - (Rm_i + \vec{t}) \right\|_2 \right)^2 \qquad (1)$$

Because only correctly tracked templates are used no weighting factor is needed in this equation. To solve this problem we use a modified version of the scalar Least Squares method. Compared to several gradient descent methods this analytical solution provides a straightforward method to get rotation and translation parameters which minimize e.

Using the previously obtained rotation matrix R and the translation vector t the initial model coordinates of the lost templates can be transformed into the correct up-to-date positions of the head in the coordinate system of the cameras. The perspective projections of both cameras give the correct position of the lost templates in both stereo images.

3.4 Gaze Estimation

Based on the position of the eye corner provided by the 3D facial feature model the gaze can be easily determined by considering the eye as an sphere. Assuming the focus is meant to be in the center of this sphere the horizontal gaze is given by angle α between the line from the center of the sphere to the pupil and the radius of the eye. Fig. 7 illustrates this theme for the horizontal gaze angle as plenary view. Similarly the vertical gaze angle can be determined.

Computation of the eye's radius is hardly possible because it cannot be guarantee that no part of the eye is occluded. Therefore the radius is presupposed as known. A mean radius of 13 mm has proven to produce good results, but results depend on user's distance to camera and the concrete anatomy of the user itself.

3.5 Experimental Results

Tab. 1 gives a glance at the amount of needed matching operations per second for a stereo camera

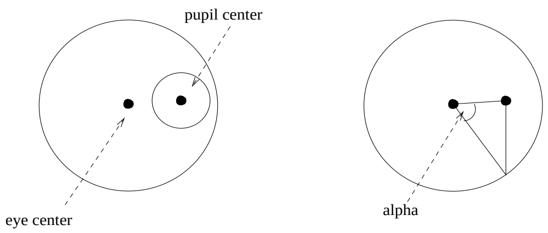


Figure 6: Estimation of horizontal gaze angle with spherical eye model.

system at PAL-resolution of 768x576 Pixel and different search windows. It shows the enormous amount of saved matching operations by using Kalman-Templates.

Size of search window	Matching operations / s
768x576 (full PAL)	135.895.449.600
384x288 (half PAL)	33.973.862.400
100 x 100	3.072.000.000
50 x 50	768.000.000
20 x 20	122.880.000
10 x 10	30.720.000

Table 1: Matching operations per second for different sizes of search windows.

Our experiments show that the accuracy of head motion tracking depends mainly on the used camera resolution. At a distance of 1,36m from head to camera with a resolution of 384x288 Pixel the mean error lays between 0,02 and 0,43 cm with a variance of 0,001 to 0,212cm when error equation (1) is used. Using higher resolutions produces better results and vice versa.

4 Perceiving Pointing Hand

4.1 B-Spline-Snakes

One of the problems in tracking body parts is caused by the possibility of deformation of the parts during motion and tracking. We propose a special formulation of B-Spline-Snakes with an underlying shape-space (see [2] for more details). The "shape" of the Snake is only defined by the *deBoor*-points of the spline function. By using this, a shape-space $\mathcal{S} = \mathcal{L}(\mathcal{W}\vec{\mathcal{Q}}_{\prime})$ for a contour could be define by :

$$\vec{Q} = W\vec{X} + \vec{Q}_0.$$

Thereby the shape-vector $\vec{X} \in \mathbb{R}^{\mathbb{N}_{\mathbb{X}}}$ is a linear mapping to a spline-vector $\vec{Q} \in \mathbb{R}^{\mathbb{N}_{\mathbb{Q}}}$ (the deBoor-points), with W as an $N_Q \times N_X$ shape-matrix and Q_0 an offset curve against which the shape variations are measured ($N_Q = 2N_B$ and N_B number of basis functions needed for the spline, N_X the dimension of the shape-space). The shape-matrix W is build depending on the dimension of the shape-space. For a space of euclidian similarities with four dimensions W is:

$$W = \begin{pmatrix} \vec{1} & \vec{0} & \vec{Q}_0^x & -\vec{Q}_0^y \\ \vec{0} & \vec{1} & \vec{Q}_0^y & \vec{Q}_0^x \end{pmatrix}$$

with the N_B-vectors $\vec{0}$ and $\vec{1}$

$$\vec{0} = (0, 0, \ldots, 0)^T, \quad \vec{1} = (1, 1, \ldots, 1)^T$$

Now the shape could be transformed affine by $\vec{X}$, which acts a weight for the columns of W. The first two columns of W cover translation and the third and fourth rotation and scaling.
The fitting of a template to a give feature-curve $\vec{Q}_f$ is given by:

$$\min_{\vec{X}} ||W\vec{X} + \vec{Q}_0 - \vec{Q}_f||^2$$

If the image data are used as feature map for a quadratic approximation of the external energy of a Snake with:

$$E_{ext} \approx \int (\vec{r}(s) - \vec{r}_f)^2 ds$$

with $\vec{r}(s)$ the initial curve and $\vec{r}_f$ as curve of the image features.

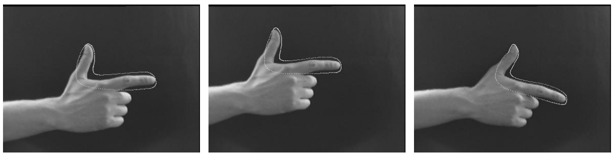
4.2 Feature Detection

The detection of the "correct" features in natural images is one of the biggest problems. We use image-filters applied along the normals of a given contour to determine the new feature location. This solution has the advantage that the image processing has not to be done over the entire image, but only in a small range depending on the length of the normals, which leads to computational efficiency.

With the described snake model a good tracking performance with high frame rates can be developed. Fig. 7 shows some examples of tracking a finger pair.

4.3 Hand Tracking and Direction Determination

The detection of pointer indicated by a hand can be solved as follows. The orientation of the hand can be directly recovered form the shape-vector $\vec{X}$, because it contains the rotation and translation parameters. The intersection of the indicating gesture with the object layer can be recovered from a pair of stereo-images [4]. In the image plane the pointer can be elongated



(a) Initial (b) Translated (c) Rotated

Figure 7: B-Spline-Snakes for hand tracking.

through the whole image (li1). The referenced line in the object plane is given by a camera-world calibration with $lp1 = T(li1)$. The fusion of the images with the two elongation of the pointers leads to an intersection of the two pointer-lines. The projection of this point onto the world plane is the, by the pointer referenced, point. A pair of images is needed because from one image only the line lp1 on the object plane can be determined.

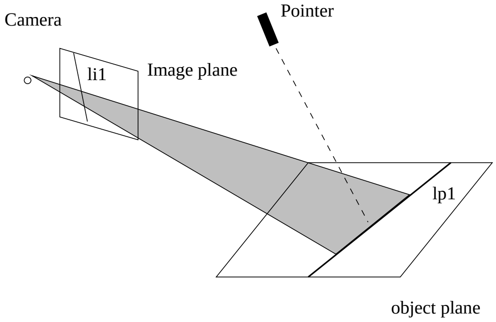


Figure 8: Determining the projection of the pointer on the object plane [4].

5 Integration and Application in the Baufix Assembly Scenario

One typical action of the toy airplane assembly is that one robot picks up the slat following the human instructor. Before this may happen, however, it has to be cleared up, which slat to take. This involves the incorporation of the gesture and gaze recognizer. By tracking of the users head motion, gaze and pointing hand, the users fields of interest can be focused. In ambiguous situations this information is used to dissolve the uncertainties (Fig. 10). The information from gaze and indicating gesture is used to determine which screw was to be meant.

Then the screwing is triggered, involving the peg-in-hole module mentioned above followed by the screwing module. The screwing is shown in Fig. 11.

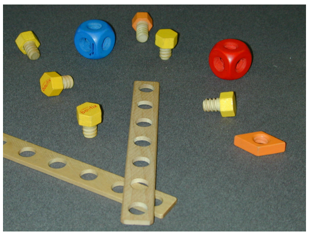


Figure 9: Uncertainties in the scenario of Baufix assembly.

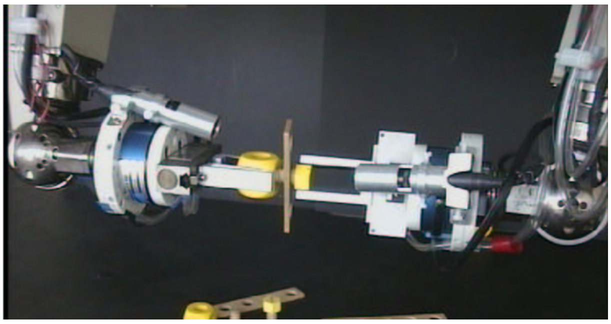


Figure 10: Screwing scenario with two robot hands.

Fig. 12 shows two typical objects that can be built with the setup as developed up to now.

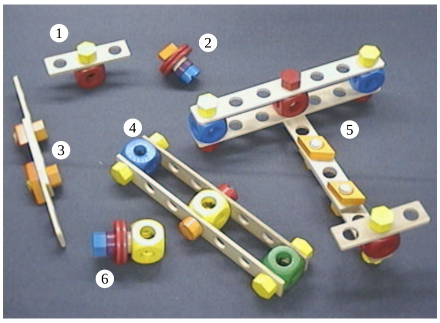


Figure 11: Sample aggregates made by our interactive assembly system.

6 Future Work

Among many topics to be explored, some important ones can be listed as follows:

Seamless communicator. Interfaces will be closely coupled with planning and monitoring. If the nature of tasks cannot be fully predicted, they are automatically decomposed into more elementary actions. Ideal action needs to be inferred based on motion and action planning while considering the context and the human preference.

Active intention detection based on multiple cues. Speech, gesture, motion sequences (human demonstrations) will be integrated and combined with contexts, knowledges and personal preference. The cross-modal interplay will be investigated. Since the system resources are limited, sensory input needs to be selected by using factor analysis, signal synthesis and tracking focus of interests.

General human perception. General human perception. Human motions are captured without using artificial markers. Wide-range, active camera configurations are applied in human recognition and precise gaze perception, also by low-quality input and occlusions. The robustness of the voice input in real environments should be significantly improved. This task is even more challenging if non close speaking microphones are used.

Grounded learning of human activities. The long-term-memory is learned from the short-term-memory so that symbols, sequences, names and attributes are anchored in the real sensor/actuator world. To enable the arbitrary transition between digital measurements and concepts, symbolic sparse coding, granular computing, fuzzy sets and rough sets will be investigated and integrated. The sensor capability can be extended by using linguistic modeling of human perception and sensor fusion so that information which is difficult to measure, incomplete or noisy can be perceived. Learning on the higher level should be conducted to select action strategies and to generate intelligent dialogs. This will need the tight integration of more components and more knowledge. The combination of grounded learning and communication will make the human-robot interaction work like interaction with a growing child which will be really entertaining.

References

[1] R. Bischoff and V. Graefe. Integrating vision, touch and natural language in the control of a situation-oriented behavior-based humanoid robot. In *IEEE International Conference on Systems, Man, Cybernetics, Tokyo*, 1999.

[2] A. Blake and M. Isard. *Active Contours.* Springer, 1998.

[3] R. A. Brooks, C. Breazeal, M. Marjanovic, and B. Scassellati. The Cog project: Building a humanoid robot. In C. L. Nehaniv, editor, *Computation for Metaphores, Analogy and Agents*, volume 1562 of *Lecture Notes in Computer Science*, pages 52–87. Springer, 1999.

[4] Roberto Cipolla and Nick Hollinghurst. Human-robot interface by pointing with uncalibrated stereo. *Image and Vision Computing*, 14:171–178, 1996.

[5] A. Green and K. Severinson-Eklundh. Task-oriented dialogue for cero: a user-centered approach. In *Proceedings of Ro-Man'01 (10th IEEE International Workshop on Robot and Human Communication)*, pages 146–151, Bordeaux-Paris, September 2001.

[6] Y. Matsumoto and A. Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proceedings of IEEE Fourth International Conference on Face and Gesture Recognition (FG'2000)*, pages 499–505, Grenoble, France, March 2000.

[7] R. T. Pack, M. Wilkes, G. Biswas, and K. Kawamura. Intelligent machine architecture for object-based system integration. In *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, June 1997.

[8] K. R. Thorissen. *Communicative Humanoids - A Computational Model of Psychosocial Dialogue Skills.* PhD thesis, MIT Media Lab., 1997.

[9] J. Weng, C. H. Evans, W. S. Hwang, and Y.-B. Lee. The developemental approach to artificial intelligence: Concepts, developmental algorithms and experimental results. In *In Proc. NSF Design & Manufacturing Grantees Conference*, 1999.

[10] J. Zhang and A. Knoll. A two-arm situated artificial communicator for interactive assembly. In *Proceedings of 10th IEEE International Workshop on Robot and Human Communication*, pages 292–299, Bordeaux-Paris, September 2001.