A Multimodal Human-Robot-Interaction Scenario: Working Together with an Industrial Robot

Alexander Bannat[1], Jürgen Gast[1], Tobias Rehrl[1], Wolfgang Rösel[2], Gerhard Rigoll[1], and Frank Wallhoff[1]

[1] Department of Electrical Engineering and Information Technology, Institute for Human-Machine Communication
[2] Faculty of Mechanical Engineering, Institute for Machine Tools and Industrial Management
Technische Universität München
80290 Munich, Germany

Abstract. In this paper, we present a novel approach for multimodal interactions between humans and industrial robots.

The application scenario is situated in a factory, where a human worker is supported by a robot to accomplish a given hybrid assembly scenario, that covers manual and automated assembly steps. The robot is acting as an assistant as well as a fully autonomous assembly unit.

For interacting with the presented system, the human is able to give his commands via three different input modalities (speech, gaze and the so-called soft-buttons).[1]

1 Introduction

In general, the cooperation between humans and industrial robots is emerging more and more in ongoing research fields, concerning human machine interaction as well as industrial safety requirements.

The here presented work is part of such a research group and aims to integrate industrial robots in human-dominated working areas using multiple input modalities to allow a true peer-to-peer level of collaboration. Thus, a smart working environment for joint-action between a human and an industrial robot is to be designed as an experimental setup consisting of various sensors monitoring the environment and the human worker, an industrial robot, an assembly line supplying the manufacturing process with material and tools, and a working table. The long term goal of this project is to form an umbrella for so-called "Cognitive Factories", where the new kind of cognitive systems abolish the currently used static and inflexible technical systems. A first approach for our realization of this long-term goal is presented in [1].

[1] All authors contributed equally to the work presented in this paper.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, pp. 303–311, 2009.

This paper is organized as follows: In the next section, a motivation for our presented multimodal approach towards intuitive and robust human-robot-interaction is given, followed by a description of the demonstrator set-up. A detailed explanation of our system architecture is lined out, succeeded by some first results. Finally, the next planned steps are shortly sketched and close this paper.

2 Motivation

As mentioned above, our long-term goal is to establish the so-called "Cognitive Factory" as the manufacturing concept for the 21st century. Here we present a small slice of our concept, constituting a foundation for a multimodal communication between a human and a robot (see Figure 1), which is flexible, robust and most appropriate for hybrid assembly. Furthermore, by utilizing more than two input modalities, we introduce redundancy in the communication channel to compensate failures in general or the total loss of an interaction modality. For the maximization of this redundancy, our interaction framework is constituted on three different channels (eyes, voice and hands). This approach does not only ameliorate the redundancy but also the diversity of possible communication ways. Therefore, consider the case, two input modalities basing on the same human interaction method (e.g. the human hand) can be blocked by a complex task requiring this interaction method (e.g. the work piece must be held with both hands). Having three input modalities basing on the three channels, the interaction between the human and the robot can be significantly shaped into a more natural, robust and intuitive way.

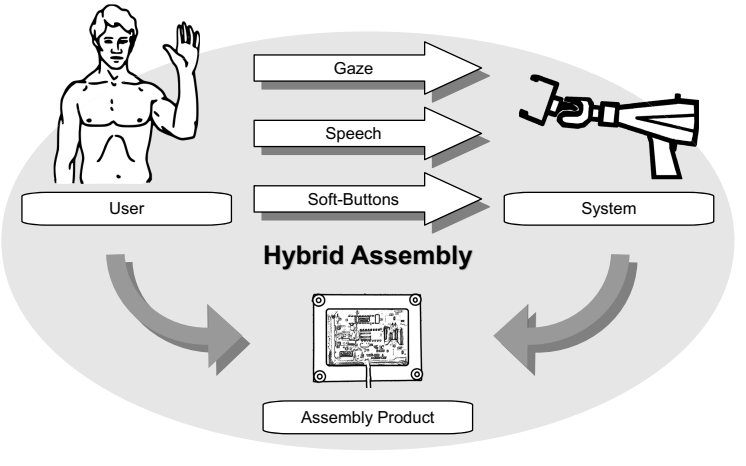

Fig. 1. Hybrid Assembly: Exploiting three independent communication channels for accomplishing a given construction task for a human worker and an assisting industrial robot

3 Set-Up Description

In the following two sections the hardware set-up as well as the assembly product will be described in detail. However, the main focus in this two sections will not be laid on the assembly process, which will be given in the Section 5.

3.1 Hardware Set-Up

In Figure 2 the demonstrator set-up is depicted. As it can be seen, the worker is wearing a head-mounted microphone as well as eye-tracking glasses. The industrial robot manipulator arm used in the scenario is a Mitsubishi robot RV-6SL. It has six degrees of freedom and can lift objects with a maximum weight of six kilograms. Its workspace lies within a radius of 0.902 m around its body. Its tool point is equipped with a force-torque-sensor and a tool-change unit. Furthermore, the robot is able to change the currently installed gripper by itself at a station, depicted on the left side of the table in Figure 2. The station features four distinct kinds of manipulators performing specific operations:

- two finger parallel gripper
- electronic drill
- camera unit for automatic observations
- gluer

These manipulators give the robot the capabilities of being able to solve entirely different tasks, like screwing and lifting.

The workbench has an overall workspace of approximately 0.70 square meters. A global top-down view camera is mounted above the workbench. This device has the overview over the entire work-area and makes it possible to watch the actions on the workbench and locate objects on the surface. Additionally, a PMD

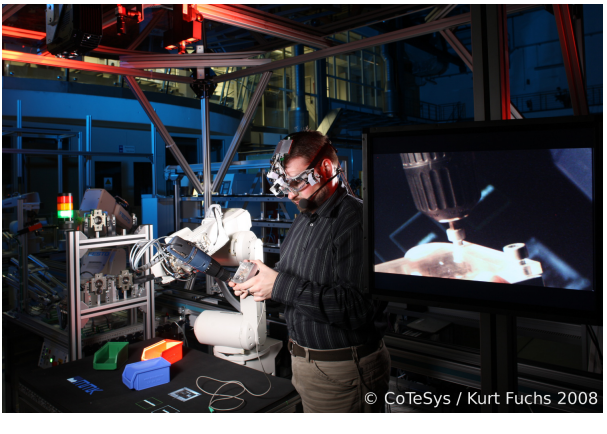

Fig. 2. Hybrid assembly station: tool station, robot arm (with electric drill) and assembly-table

range sensor is mounted above the workbench for providing depth information of the scenery.

For bringing information into the worker's field of view, a table projector is also installed above the workbench. This device projects information directly onto the surface of the workbench. With this modality, it is possible to show contact analog assembly instructions and system feedbacks.

Moreover, flexible interaction fields can be displayed to communicate with the system (see Section 4.3).

3.2 Assembly Product Description

The product constructed in the hybrid assembly scenario is a high frequency transmitter, as it can be seen in Figure 3. The following parts are required for the construction of this high frequency transmitter:

- *One base plate:* The base plate delivers the foundation for the hybrid assembly on which all remaining construction parts will be mounted on.
- *One electronic part:* The electronic part is an inlay, which has to be mounted into the base plate.
- *One wiring cable:* The wiring cable has to be plugged into the electronic part. It needs to be put into the designated cable line on the base plate.
- *One plastic cover and four screws:* With the four screws, the plastic cover is mounted on top of the base plate to protect the inlay against influences from the environment, like dust and dirt.

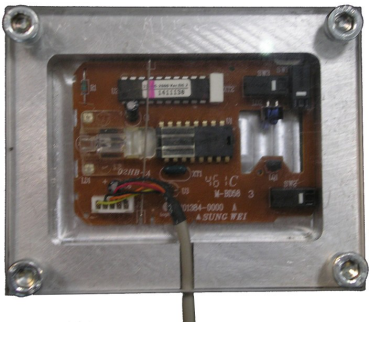

Fig. 3. Assembly product: a base plate, one electronic part, a wiring cable, a plastic cover and four screws

4 System Overview

As a backbone for the integration and fusion of the three input modalities, a short term memory is used in this scenario. A real-time capable framework processes all kinds of data streams (Real-Time Data Base [2,3,4]). Here, the data streams constituted by the three input modalities have very diverse features.

These differences appear especially in the sampling rates of sensory sources and losses of data packets. Another reason, why the Real-Time Data Base (RTDB) is used, is the following: The RTDB can process asynchronous data streams in real-time and allows a best match access to these streams, so that a high-degree of data coherence is achieved for further processing.

Three different modules capture the required data for the processing: First, the gaze-module captures the current human gaze vector via the EyeSee Camera [5] and the result is stored into the RTDB. Second, the speech module transmits the raw audio data into the RTDB and last, the soft-buttons module displays different buttons onto the workbench and returns the button status to the RTDB (selected/not selected).

4.1 Human Gaze

The first channel in this scenario is human gaze. Therefore, the worker has to wear eye-tracking glasses, enabling the system to extract the gaze information. The perceived information consists of the intersection of the gaze trajectory and the workbench surface. As a result of this sensor device, the actual 2D-coordinates in the table plain are transmitted into the RTDB. The update rate of this position data is approximately about 120 frames per second.

Having the actual position of the worker's gaze on the workbench, the system is now capable to provide the worker with important information directly in his current field of view.

Besides with having the gaze point, a new kind of human-machine interaction is now available: control by gaze. The gaze point can be used to select certain sensitive fields either by fixating these over a certain period of time or choose the desired function with gaze and confirm the selection with a further modality. This input can then be used to initiate several system functionalities, e.g. to browse through an assistant assembly instruction system, like presented in [6]. This form of interaction is especially suited for situations where hands-free interaction is needed and the environment does not allow reliable speech recognition.

4.2 Speech Input

As known from many human-human interaction scenarios, speech is a common and favored communication channel. Thus, speech constitutes the main interaction method in our scenario (as our second channel). Therefore, a speech recognition software is integrated into the framework.

The raw audio data is captured with a head-mounted microphone in PCM-standard format. This data stream is delivered via the RTDB towards the speech recognition engine, which in turn writes the recognition results with their corresponding confidence back into the RTDB-buffer.

So far, the recognition is based on phonemes. This enables the system to detect predefined command words, which can be defined in a Backus-Naur Form [7].

This grammar can be adapted during runtime according to the current task. Thus, to the actual state-of-the-art of speech recognition software, it is situated to use command language instead of natural language. Because the complexity of the speech recognition is drastically reduced by applying a command-word structure and, therefore, a high acceptable recognition rate is accomplished. Furthermore, command-word-based recognition software is also able to cope with different users more efficiently. The complex training process of natural speech recognition software is not necessary in the used set-up.

4.3 Button Interface

The third channel is our concept of the so-called Soft-Buttons. With a Soft-Button it is possible to control a function associated with a certain sensitive field. The content and appearance of this field, as well as its position, can be adjusted at run-time according to the desired user-preferences with respect to the current available space on the workbench for displaying. Therefore, in a certain area reserved for human-system interactions, fields are projected onto the workbench with a table-projection unit mounted above the workspace.

The system is capable of detecting whether the worker's hand is within a sensitive area or not, therefore a vision based hand-detector is applied. A skin color model, like presented in [8], delivers the foundation for the hand-detector module. Furthermore, the Soft-Buttons can be selected via exploiting the above introduced human gaze information.

The free distribution of the sensitive fields on the workbench allows an ergonomic interface design and an optimal workspace usage. Due to the fact that the information is directly displayed into the worker's field of view, there is no need for him to shift his attention from his current task towards a monitor. This modality decreases disturbances in the workflow.

5 Preliminary Assembly Process

The assembly process of the product described in Section 3.2 is depicted in Figure 4. The first interaction – initiated via Soft-Button/Speech Command – between the human and the robot is the supply of the worker with the base plate. Having the required work pieces available at hand, the worker starts to teach in a glue-line (see Figure 4.a). The track of the glue-line on the base plate is taught with *Programming by Demonstration* (PbD). This is done by tracking a colored pointer. Therefore, a color-based image segmentation is performed on the output frame of a top-down-view camera, mounted above the workbench.

The result is then used to locate the green tip of the pen, now visible as a binary-coded plane in the filtermask. The motion of this object in the image plane is analyzed and transformed into the world-coordinate-system. While the line is perceived, its trajectory is on-line projected back onto the work piece as a direct feedback for the worker. On completion of this step the robot changes

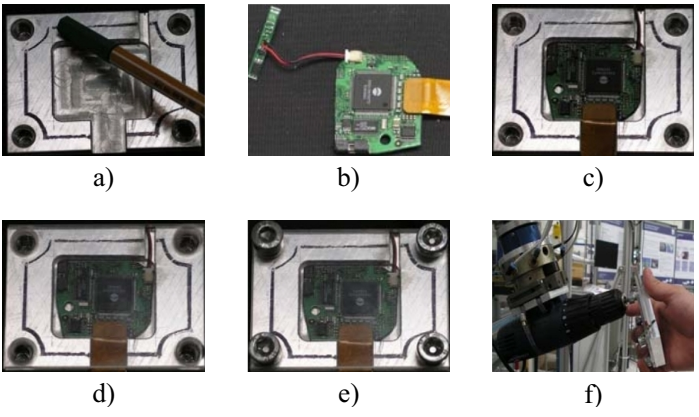

Fig. 4. Assembly steps of Use-Case product

its tool device from the gripper to the gluer according to the next step in the work plan. After the PbD, the robot protracts the glue on the work piece. Via transforming these gained world-coordinates into the robot-coordinate-system, the industrial robot is now able to exactly repeat this motion trajectory and perform the gluing-operation autonomously.

As per assembly instructions the robot reaches out for the electronic parts. Therefore, a fully automated tool change operation is performed by the robot, thus, exchanging the currently mounted gluer towards gripper. The following assembly of the electronic parts (see Figure 4.b) require fine motor skills. Therefore, the next step is solely done by the human. In spite of the fact that the robot does not give any active assistance in this assembly step, the system supports the worker via presenting the manufacturing instructions for the insertion of the electronic parts into the base plate (see Figure 4.c and 4.d).

After the worker has acknowledged the completion of the current step via Soft-Button (third channel) or a speech based command (second channel), the robot fetches the four screws for the final assembly step. While the worker is pre-fitting the screws in the designated mount ports (see Figure 4.e), the robot retrieves the automatic drill device from the tool changer station. The velocity of the drill is adjusted to the contact pressure of the work piece against the drill (see Figure 4.f). The more pressure is applied, the faster the drill goes. As soon as the human recognizes that the screw is fixed – the rattling noise of the slipping clutch – he will loosen the former conducted pressure. This modality allows for an intuitive screwing behavior of the system.

The final step requires the worker to acknowledge the completion of the screwing operation via speech command or Soft-Buttons. The system will go into the init state and the whole assembly process can be again started for the next production turn.

6 Conclusion and Outlook

A first technical implementation of a human-robot interaction scenario is introduced and was exhibited at the trade fair "AUTOMATICA 2008" in Munich, Germany. The system has not reached its final stage and it shows still room for improvement. Nonetheless, the presented approach offers high potential for a new multimodal human-robot interaction system and is currently evaluated by human experiments within the project JAHIR [9].

Furthermore, it is possible to produce the above described high frequency transmitter – this product is of course non functional. At this stage, we showed a working hybrid assembly process – human working together with an industrial robot. In spite of the fact of the functionality of this new form of cooperation, there is still a lack for security and safety features. These features have to be integrated into our system to protect the worker from injuries, what will be the next task in our research.

As a further field of improvement, the EyeSee Camera mentioned in Section 4.1 does not allow a noninvasive gaze recognition because of the eyetracking glasses and the required wiring. Therefore, a remote gaze tracking system is currently developed at our institute for enabling an undisturbed and naturalistic workflow.

Acknowledgment

This ongoing work is supported by the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see `www.cotesys.org` for further details. The authors further acknowledge the great support of Matthias Göbl for his explanations and granting access to the RTDB repository.

References

1. Zäh, M.F., Lau, C., Wiesbeck, M., Ostgathe, M., Vogl, W.: Towards the Cognitive Factory. In: Proceedings of the 2nd International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV), Toronto, Canada (July 2007)
2. Goebl, M., Färber, G.: A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles. In: Intelligent Vehicles Symposium, pp. 737–740 (June 2007)
3. Stiller, C., Färber, G., Kammel, S.: Cooperative cognitive automobiles. In: Intelligent Vehicles Symposium, pp. 215–220. IEEE, Los Alamitos (2007)
4. Thuy, M., Göbl, M., Rattei, F., Althoff, M., Obermeier, F., Hawe, S., Nagel, R., Kraus, S., Wang, C., Hecker, F., Russ, M., Schweitzer, M., León, F.P., Diepold, K., Eberspächer, J., Heißing, B., Wünsche, H.J.: Kognitive automobile - neue konzepte und ideen des sonderforschungsbereiches/tr-28. In: Aktive Sicherheit durch Fahrerassistenz, Garching bei München, Garching bei München, April 7-8 (2008)
5. Bardins, S., Poitschke, T., Kohlbecher, S.: Gaze-based Interaction in various Environments. In: Proceedings of 1st ACM International Workshop on Vision Networks for Behaviour Analysis VNBA 2008, Vancouver, Canada (2008)

6. Bannat, A., Gast, J., Rigoll, G., Wallhoff, F.: Event Analysis and Interpretation of Human Activity for Augmented Reality-based Assistant Systems. In: IEEE Proceeding ICCP 2008, Cluj-Napoca, Romania, August 28-30 (2008)
7. Naur, P.: Revised Report on the Algorithmic Language ALGOL 60. Communications of the ACM 3(5), 299–314 (1960)
8. Soriano, M., Huovinen, S., Martinkauppi, B., Laaksonen, M.: Skin Detection in Video under Changing Illumination Conditions. In: Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain, pp. 839–842 (2000)
9. Lenz, C., Nair, S., Rickert, M., Knoll, A., Rösel, W., Bannat, A., Gast, J., Wallhoff, F.: Joint Actions for Humans and Industrial Robots: A Hybrid Assembly Concept. In: Proc. 17th IEEE International Symposium on Robot and Human Interactive Communication, Munich, Germany (2008)