MULTIMODAL DATA COMMUNICATION FOR HUMAN-ROBOT INTERACTIONS

Frank Wallhoff, Tobias Rehrl, Jürgen Gast, Alexander Bannat and Gerhard Rigoll

Human-Machine Communication
Department of Electrical Engineering and Information Technologies
Technische Universität München
Munich, Germany

ABSTRACT

In this paper, the development of a framework based on the Real-time Database (RTDB) for processing multimodal data is presented. This framework allows readily integration of input and output modules. Furthermore the asynchronous data streams from different sources can be approximately processed in a synchronous manner. Depending on the included modules, online as well as offline data processing is possible. The idea is to establish a real multimodal interaction system that is able to recognize and react to those situations that are relevant for human-robot interaction.[1]

***Index Terms*—** multimodal data communication, human-robot interaction, real-time data processing, augmented reality

1. INTRODUCTION

The excellence research cluster *Cognition for Technical Systems CoTeSys* attempts to establish a more intelligent and useful behavior of technical systems, especially in unseen situations. Key goals of a cognitive system are the abilities to perceive and react on actions occurring in its environment and to develop its capabilities using learning strategies. One important aspect is to improve the interaction and communication between humans and technical systems in a way to become more intuitive. Hence, today's technical systems are becoming more and more complex, the requirements on a framework for coordinating different technical modules increase.
The general idea is to engineer a common framework for harmonizing the different software modules to establish a natural interaction between a human and a robot. Although in the literature several middleware architectures have already been introduced to deal with multimodal data streams, most of them are suffering from transparency or from real-time capabilities, e.g. [1].
Especially distributed systems with a high overhead of network data exchange are not suited to act as a multimodal processing-tool appropriately. Therefore, we propose to use the RTDB as a sensory buffer with a high data bandwidth applicable for online as well as offline applications or experiments. The underlying RTDB architecture has established itself as a reliable platform in conjunction with Cognitive Vehicles [2, 3, 4]. Furthermore, it has served as a convincing integration platform, where several groups of researchers simultaneously work on the same framework.

The rest of this paper is organized as follows: In Section 2, a short application scenario for the proposed common framework is described. In Section 3, we introduce the integrated system framework basing on the RTDB with its interface modules in greater detail. In Section 4 the use-case, a human-robot-interaction, is delineated more precisely. The paper closes with a summary and an outlook over the next planned steps.

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

2. APPLICATION SCENARIO

A human-robot interaction scenario similar to the setup of the CoTeSys Project *Joint Action of Humans and Industrial Robots (JAHIR)* presented in [5] serves as field of application for the proposed framework basing on the RTDB. The framework has to be adapted in order to view the outputs of different sensors and in addition to control these outputs and react to the worker's actions. This implies that the software modules are expanded in the way that for example several different computer vision algorithms can access the same video source in parallel (shared-memory).

A typical joint action scene with a human worker and an industrial robot arm is depicted in Figure 1. This setup has been presented

Fig. 1. Human worker and industrial robot in hybrid assembly (picture taken at AUTOMATICA trade fair 2008 in Munich, Germany).

during the AUTOMATICA trade fair 2008 in Munich with great success. The area of application for the AUTOMATICA-Demonstrator is situated in a hybrid assembly task in the so called Cognitive Factory [6]. In this case, the robot is used as an assistant for the human worker. Besides, the robot helps fulfilling the worker's task. The hardware as well as the software modules will be considered in the following sections.

2.1. Hardware Setup

In Figure 2 the hardware setup of the AUTOMATICA demonstrator is shown. 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

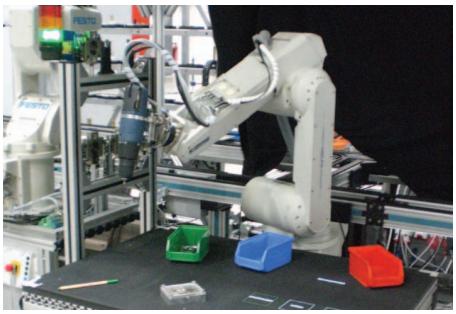


Fig. 2. Hybrid assembly station: tool station, robot arm (with electric drill) and assembly-table.

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. The tools stored in the station are:

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

Thus, the robot has the capabilities 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 Photonic Mixer Device (PMD) range sensor is mounted above the workbench (see section 3.3).
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. This allows an ergonomic information presentation, because the worker is not disturbed in his workflow while performing his task to perceive relevant instructions.
Moreover, flexible interaction fields can be displayed to communicate with the system (see following section).

3. SYSTEM ARCHITECTURE

In this section an overview of the system architecture and the software modules used in this application-scenario will be given. The system architecture consists of the Real-time Database as a sensory buffer and communication backbone between the input and output modules. The used architecture is depicted in Figure 3.

3.1. Real-time Database

The RTDB presented in [7] is able to deal with large amounts of sensor data and can provide data exchange and recording in real-time on a Linux PC equipped with an AMD Phenom 2.2GHz quad-core and four gigabyte RAM. In cognitive autonomous vehicles the database is used to manage all sensor inputs to keep the vehicle on track. The RTDB manages objects that can be created and updated by input modules, also called writers. These writers also have to submit a timestamp for their committed data. Thereby, it is possible for the RTDB to synchronize the data coming in asynchronously from multiple sources and at different sampling rates. Output modules (called readers) wait for new objects to process them. For example, a module can write the image of a camera and multiple other modules can analyze this image in parallel to generate information on a higher level and write this output back for other modules without blocking effects. These data-objects can be recorded in real-time, bringing up two major advantages:

- The recorded sensor-input can be taken for replay or simulation of certain situations. In addition, the gathered material can be analyzed by humans offline, e.g. to reveal important gestures in the co-operation process.
- The data can also be used for benchmarking purpose. Different reader implementations can be tested on the same data under the same conditions. After the new reader proves to work better than the old one, they can be used on the real set-up online without any modification to the code. The recorded database of sensor data can also be used by other projects to evaluate their system or algorithms in an unknown environment.

3.2. Video Processing

The video processing modules deliver raw RGB data from different sources like firewire cams or USB cams in a common representation based on OpenCV. The OpenCV library has been chosen, because it is widespread and numerous working output modules exist, e.g. to localize faces or hands. In addition to these modules it is planned to train and implement gesture recognition like [8]. To obtain better results in training and recognition, it has been decided to record the video data uncompressed. In our setup we use one webcam (for top-down observation) and one firewire cam (mounted on the robot arm). In order to compensate different lighting conditions at our setup during the experiments the gain of the cameras can be controlled online.

3.3. Range Maps

For a "deeper" view of the scene, input from a camera providing range maps has been recorded. Based on the novel PMD technology, the camera collects depth-information in real-time by emitting infrared light and measuring the time-of-flight. Thereby, the distances from the camera can be calculated. It has a resolution of 64 x 48 pixels at 25 frames per second. This additional depth information can be used to improve segmentation tasks for image processing or detection of human activities like handovers.
More information regarding this sensor and calibration techniques can be found in [9]. However, because the camera is sensitive for infrared light, it can also be used to provide intensity based gray scale images.

3.4. Software Modules

3.4.1. Gaze-Control

The system is capable of extracting gaze information. Therefore the worker has to wear eye-tracking glasses [10]. The perceived information consists of the intersection of the gaze trajectory and the workbench surface. Using this result, important information can be displayed in the worker's field of view. Furthermore, the later introduced Soft-Buttons (see Section 3.4.6) can also be activated via gaze.

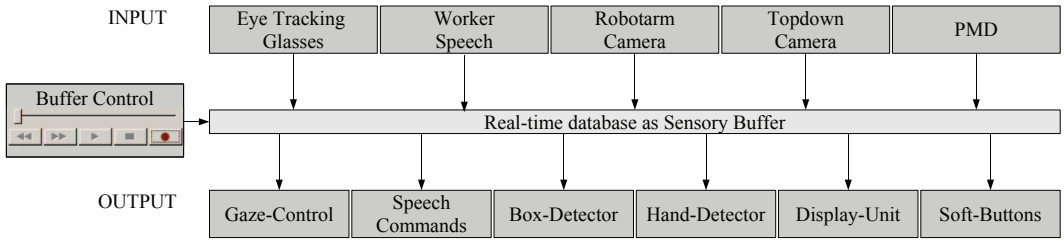


Fig. 3. Overview of the multi-modal framework.

3.4.2. Speech-Recognition

Another input modality is constituted via speech. Therefore, a commercial speech recognizer software is integrated into the framework. The recognition is based on phonemes to detect command words only, which can be defined in a Backus-Naur Form (BNF). This grammar can be adapted while runtime with regard to the task and the task-related command words.

3.4.3. Box-Detector

The robot has to grasp storage boxes located on the workbench. This requires to find these boxes first, which is done by analyzing the acquired image from the top-down view camera. The used boxes can vary in color and size (nearly square and rectangular). The detection of their position in pixel coordinates is done with a color-based image segmentation using thresholding filters in the HSV-color-space. The relevant areas are extracted from the image plane and background information is purged [11]. The filter operation is followed by a binary conversion for each box color and additional size features are included, resulting into a so called ”blob”. The system will identify blobs as boxes, if their size is in a given interval of the image-area. By these constraints it can be determined, if an area is a box or not. The center of gravity for each box and additional rotation information are calculated. This gained information constitutes the input for the computation of the box-position in robot-coordinates.

3.4.4. Hand-Detector

Detecting the worker's hands is done by applying a skin color filter operation, similar to the color-segmentation filter used for the box detection. It is possible to find areas in the image, containing only human skin-color. The parameters of this filter operation are adjusted according to the skin color model of [12] in the rg-Chroma color space. Because hands have a certain height-to-width value, regions with other values can be purged. Thus, the resulting mask has only hand related blobs in it. Again, the center of gravity of these hand-blobs are calculated, representing the actual position of the hand. Combined with the restriction of the area of interest, the value for the extracted location of the hand in image coordinates can be improved.

3.4.5. Display-Unit

The table projector is mounted above of the workbench. Thus, the whole area of the workbench can be used for displaying information. One important function is to display the Soft-Buttons creating the only human-machine interface (HMI) in the visible domain. Besides, textual and/or image based information can be presented on the workbench to assist the human worker by accomplishing his/her task. For introducing more flexibility in the information presentation, the desired display content as well as its position, can be configured dynamically according to the worker's preferences. As known from a standard PC, additional system notifications can also be reported with the Display-Unit.

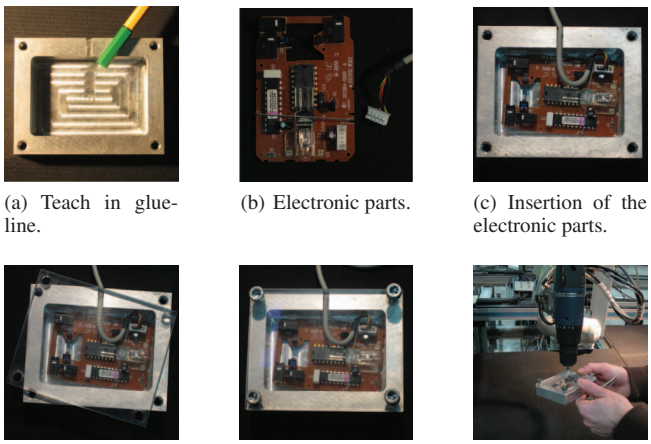
3.4.6. Soft-Buttons

Our concept of the so called Soft-Buttons is the following: In a certain area reserved for human-system interactions, fields are projected onto the workbench. Using the above introduced hand-detector, the system is capable of detecting whether the worker's hand is within a field or not. Yet, the function associated with the Soft-Button only becomes active, if the worker's hand remains hovering above the sensitive field for at least five frames.
Furthermore, the Soft-Buttons can be selected via the above described Gaze-Control. The gaze control is especially suited for situations where hands-free interaction is needed and the environment does not allow reliable speech recognition.
The location of the sensitive field can be freely distributed on the workbench to allow an ergonomic interface design and an optimal workspace usage. The position of the buttons is communicated via the RTDB (see Section 3.1). The area of the sensitive fields can be defined in the display unit.

4. USE-CASE

The product constructed in the hybrid assembly scenario is a high frequency transmitter consisting of a base plate, two electronic parts, a plastic cover and four screws, shown 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, similar to the box-detection (see Section 3.4.3). While the line is perceived, its trajectory is online projected back onto the work piece as a direct feedback for the worker. On completion of this step the robot changes 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. As per assembly instructions, the robot reaches out for the electronic parts.
Fine motor skills are required for assembling the electronic parts (see Figure 4(b)). 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 present-



(a) Teach in glue-line. (b) Electronic parts. (c) Insertion of the electronic parts.

(d) Sealing in the electronic parts. (e) Pre-fitting of the screws. (f) Screwing operation.

Fig. 4. Assembly steps of Use-Case product.

ing 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 or a speech based command, 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.

5. CONCLUSIONS AND FUTURE WORK

We have presented the implementation of a unified software framework based on the RTDB that efficiently allows for on- and off-line processing of multimodal, asynchronous data streams. The presented framework serves as an effective middleware in a human-robot joint action scenario with real-time constraints, demonstrated at the trade fair AUTOMATICA 2008 in Munich with great success. Having this proof of concept, it is now imaginable to apply this framework to more complex human-robot interaction scenarios, like multi-robot human scenarios.
The integration of the presented multimodal data communication into a cognitive kitchen scenario is a next planned working step. In this scenario, several autonomous mobile robot platforms are involved in the entire process from taking orders over serving food and drinks to human customers to tidy up the kitchen.

6. ACKNOWLEDGEMENT

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.

7. REFERENCES

[1] D. S. Touretzky and E. J. Tira-Thompson, "Tekkotsu: A framework for aibo cognitive robotics," in *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, Pittsburgh, PA., July 2005.

[2] M. Göbl and G. Färber, "Interfaces for integrating cognitive functions into intelligent vehicles.," in *In Proc. IEEE Intelligent Vehicles Symposium*, June 2008, pp. 1093–1100.

[3] M. Thuy, M. Göbl, F. Rattei, M. Althoff, F. Obermeier, S. Hawe, R. Nagel, S. Kraus, C. Wang, F. Hecker, M. Russ, M. Schweitzer, F. Puente León, K. Diepold, J. Eberspächer, B. Heißing, and H.-J. Wünsche, "Kognitive automobile - neue konzepte und ideen des sonderforschungsbereiches/tr-28," in *Aktive Sicherheit durch Fahrerassistenz*, Garching bei München, 7-8 April 2008.

[4] C. Stiller, G. Färber, and S. Kammel, "Cooperative cognitive automobiles," in *Intelligent Vehicles Symposium, 2007 IEEE*, June 2007, pp. 215–220.

[5] C. Lenz, N. Suraj, M. Rickert, A. Knoll, W. Rösel, A. Bannat, J. Gast, and F. Wallhoff, "Joint actions for humans and industrial robots: A hybrid assembly concept.," in *Proc. 17th IEEE International Symposium on Robot and Human Interactive Communication*, August 2008.

[6] M. Zäh, C. Lau, M. Wiesbeck, M. Ostgathe, and W. Vogl, "Towards the Cognitive Factory," in *International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV)*, Toronto, Canada, July 2007.

[7] M. Goebl and G. Färber, "A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles," in *Intelligent Vehicles Symposium*. June 2007, pp. 737–740, IEEE Press.

[8] S. Reifinger, F. Wallhoff, M. Ablaßmeier, T. Poitschke, and G. Rigoll, "Static and dynamic hand-gesture recognition for augmented reality applications," in *Proceedings of the International Conference on Human-Computer Interaction*, C. Stephanidis, Ed., Beijing, July 2007, Springer.

[9] F. Wallhoff, M. Ruß, G. Rigoll, J. Göbel, and H. Diehl, "Surveillance and activity recognition with depth information," in *IEEE International Conference on Image Processing (ICIP)*, San Antonio, Texas, USA, September, 16-19 2007.

[10] S. Bardins, T. Poitschke, and S. Kohlbecher, "Gaze-based Interaction in various Environments.," in *Proceedings of 1st ACM International Workshop on Vision Networks for Behaviour Analysis, VNBA 2008, Vancouver, Canada*, October 31 2008.

[11] A. Bannat, J. Gast, G. Rigoll, and F. Wallhoff, "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.

[12] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen, "Skin Detection in Video under Changing Illumination Conditions," in *Proc. 15th International Conference on Pattern Recognition*, Barcelona, Spain, 2000, pp. 839–842.