ADAPTIVE HUMAN-MACHINE INTERFACES IN COGNITIVE PRODUCTION ENVIRONMENTS

F. Wallhoff, M. Ablaßmeier, A. Bannat, S. Buchta, A. Rauschert, G. Rigoll and M. Wiesbeck†

Technische Universität München
Human-Machine Communication, Theresienstr. 90, 80333 Munich
†Machine Tools and Industrial Management, Boltzmannstr. 15, 85747 Garching

ABSTRACT

This article presents an integrated framework for multi-modal adaptive cognitive technical systems to guide, assist and observe human workers in complex manual assembly environments. The demand for highly flexible construction facilities obviously contradicts longer training and preparation phases of human workers.

By giving context-aware building instructions over retina displays, text-to-speech commands or acoustical signals, a non-specialized industrial stand-by men in a production task can precisely be alloted to execute the next processing step without any previous knowledge. Using non-invasive gesture recognizers and object detectors the human worker can be observed in order to track the production line and initiate the subsequent step in the interaction loop.

Aiming at testing and evaluating the desired human-machine interfaces and its capabilities a virtual working place together with a concrete use case is introduced.

1. INTRODUCTION

Due to the fact that material goods on todays production lines are usually directly manufactured according to their customers' demands, such as the configuration of a car's interior with a certain radio model, a high demand for flexible and reliable workers immediately arises. Furthermore, the provision and storage of a huge number of parts and devices represents an additional and challenging aspect and increases the complexity of production processes [1].

As a consequence, the integration of cognitive machines in human control dominated manual assembly environments becomes desirable for tasks that can not be done by autonomous robots. However, with cognitive technical systems the work content can be allocated among human workers with situation adaptive assistance improving an ergonomic worker integration in a dynamical manner. Especially stand-by men can be scheduled into the daily personal belongings more efficiently.

By observation of the usual naturalistic human habits in a given production queue using innovative multimodal interfaces, such as emotion and gesture recognition, gaze tracking, motion tracking and others [2], the production process can be tracked and thus be accompanied with context dependent advices wherever necessary. The output can be presented over the acoustical and/or the visual channel. However, the introduced measurements must not interfere with the assembly task and have therefore to be non-invasive.

Altogether the above formulated constraints to a cognitive driven factory can be divided into four functional classes: the input or observation techniques, the output modalities, the task and knowledge representation and finally a user modelling. The main scope of the actual report focuses on these three topics that are derived from and evaluated on the basis of a well-defined virtual production process.

The paper is organized as follows: after the introduction of cognitive systems in the production domain the needed input and output modalities are listed in the next section. A concrete use case with a virtual production task is introduced. This process is represented as a finite state machine in section 5. After the implementation of the architecture, the treatise closes with a summary and additional modalities to be added in the future.

2. COGNITIVE TECHNICAL SYSTEMS

"Cognitive technical systems are equipped with artificial sensors and actuators, integrated and embedded into physical systems, and act in a physical world. They differ from other technical systems by performing cognitive control and have cognitive capabilities. Cognitive control orchestrates reflexive and habitual behavior in accord with longterm intentions. Cognitive capabilities such as perception, reasoning, learning, and planning turn technical systems into ones that *know what they are doing*. More specifically, a technical system becomes a cognitive system, if it can reason substantial amounts of appropriately represented knowledge, learn from its experience so that it performs better tomorrow than it did today, explain itself and be told what to do, be aware of its own capabilities and reflect on its own behavior, and respond robustly to surprise. Technical systems that are cognitive in this sense will be much easier to interact and cooperate with, be robust, flexible, and efficient." [3]

The presented work has been established in accordance with the *Cognitive Factory* within the CoTeSys cluster of excellence. An overview of possible applications in such a factory is given in Figure 1.

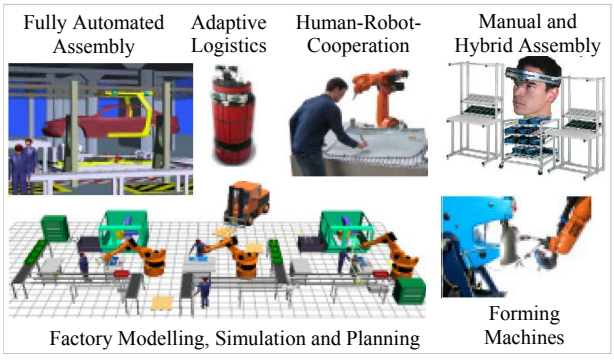

Fig. 1. Functional constellations in a cognitive factory.

3. HUMAN-MACHINE INTERFACES

For the cognitive technical system, several intelligent recognition systems are essential to reach a maximum of situation awareness. These recognizers must be able to detect the context and status of the human worker. Tracking systems have to be capable of locating the current position of the worker and record the trajectory. Emotion and workload recognizers (e.g. vision based) can classify the status of the worker and identify work-flow problems. The focus in this field is on multimodal integration and fusion of different recognizers to securely estimate the current mental status of the human worker and his current task according to her or his skills. After the multimodal fusion of multiple recognizers, sensors and the process status, methods using artificial intelligence are able to dynamically identify suitable support strategies for the human worker.

The solution strategies can be submitted after a so-called multimodal and context-adaptive fission via the visual, acoustic and/or haptic channel. The workload and capacity of each individual information channel has to be analyzed and the situation suitable channels can be selected. For the situation-adequate support of different groups of workers and efficient information flow methods of augmented reality are used. For example, a head mounted display (HMD) can be used for adaptive visual overlay of help and solution strategies, augmented acoustic can become important to direct the user's attention. The amount of presented information can be adapted to the worker and situation.

3.1. Input Modalities

In this section several input modalities are introduced that can be used for direct system interaction or for surveillance and sensing algorithms, which will be necessary for the use case presented in section 4. Due to unregulated noise in a factory, command oriented speech recognition has not further been investigated.

Hand gesture recognition By mounting a camera above the working area, the resulting images will show hands and gestures from above, such as grasping. The first step for a dynamic recognizer is based on skin color. Invariant moments can be computed on these hand shapes for all images of the sequence [4]. A dynamic sequence starts when its mean pixel difference is above a pre-defined threshold and stops when 3 consecutive frames are below this again. Unknown sequences are classified using Hidden Markov Models. The reclassification rate of the implemented system reaches the perfect score.

Photonic Mixer Device(PMD) Most vision algorithms use planar data which makes a segmentation task of an arbitrary scenery difficult. By using a PMD range sensor additional depth information can be gathered making the observation task faster and also more reliable [5]. This sensor can therefore be applied to track a person and to classify its behavior. A segmentation result of a gesture is depicted in Figure 2.

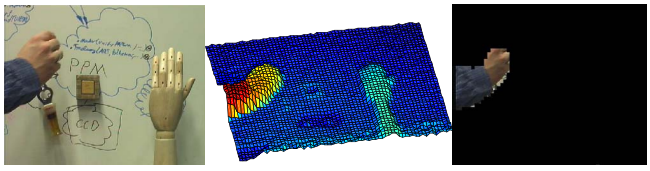

Fig. 2. Gesture recognition: raw color image (left), color coded depth map (middle) and segmented foreground object (right).

Target Tracking For implementation of augmented reality (AR) we use an commercial toolkit, i.e. the "Unifeyed SDK" by Metaio [6]. This toolkit offers basic AR functions, such as loading virtual objects, rendering and processing data of a tracking system (e.g. an infrared tracking system). "Unifeyed SDK" offers an COM (ActiveX [7]) interface. This interface defines a contract about all the possible functions, properties and events which may be used or which may arise in the ARDisplay (e.g. mouse click). In the later use case it is deployed to highlight an active disposal box.

Numerical Keypad For tactile command reception a regular USB number keypad has been foreseen.

Face Detection with Gaze Estimation In order to detect the position of a face while simultaneously recognizing the head, facial expressions and eye view direction we propose a dual camera setup as follows: one camera captures a scene using the regular visual wavelengths without additional illumination. The measurements based on the visual queue are mostly for face detection and rely on skin color models, face likelihood using neural networks, eye likelihood computation and estimation of the head orientation. The second camera records the infrared spectrum of the scene making use of self-emitted infrared illumination. The results gained from this infrared camera are an appearance based face detection, pupil detection and an eye view angle detection [8].

FASTrack Isotrack II This device offers real time six-degrees-of-freedom tracking and can be utilized also for localization tasks. The unit computes the position (X, Y, and Z Cartesian coordinates) and orientation (azimuth, elevation, and roll) of a small sensor as it moves through a electromagnetic space. The system's near zero latency makes it ideal for critical surveillance applications where real time response is critical.

3.2. Output Modalities

The implemented output modalities are:

Visual Screen Since vision is the most dominant sense, this modality is very important to transport complex matters over text or an image or video sequence. Therefore relevant process information are displayed at a static position.

Retina Display Besides the above displaying technique, AR represents a technology in which a user's view of the real world is enhanced or augmented with additional information generated from a computer model. According to Azuma [9] AR is defined as a combination of reality and virtual reality. Real objects and virtual objects are related in a three-dimensional manner. In order to have a working AR system, the see-through display system must be calibrated so that the graphics is properly rendered. By this such an optical see-through system represents an additional challenge, because one does not have direct access to the image data being needed for various calibration procedures [10]. In this use case a head-worn, see-through Nomad Display System (Microvision ND2000, distributed by Metaio) overlays computer based information over the real-world allowing hands free, head-up access to any digital information. It enables the worker with new capabilities to solve or improve existing processes. The HMD bridges the gap by delivering critical information directly to the work-space in an easy to use head-up system, i.e. the contact analog highlighting of destinations or important objects. With this system 32 different shades of red can be displayed.

Text-to-Speech System For generating speech output from textual input, a text-to-speech generation system developed at our institute is embedded into the system [11]. Through this modality it becomes possible to instruct the worker, even if his view is concentrated on something else.

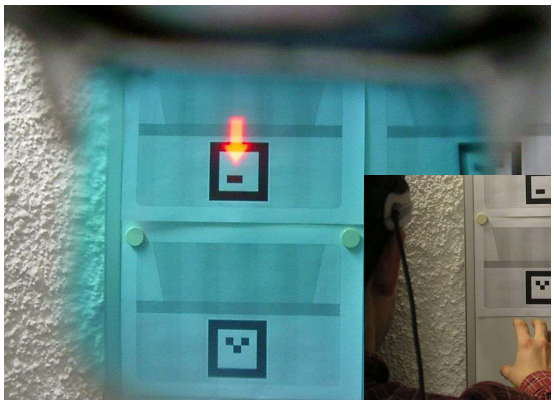

Fig. 3. Photographed augmented arrow indicating the target case together with external scene view (lower right).

4. VIRTUAL PRODUCTION TASK

As stated above, the adaptation of the system to the user is the main research topic in this paper. Therefore the cognitive machine has to observe the worker, recognize relevant actions and react to the situation accordingly. Since this is a high demanding and complex task consisting of numerous modules with several degrees of freedom, a fully controllable virtual production use case is established and studied. To lay the cognitive basis for situation and context aware information processing, in this phase a *virtual assorting task* is designed to incorporate and evaluate the above introduced input- and output modalities, see Figure 4.

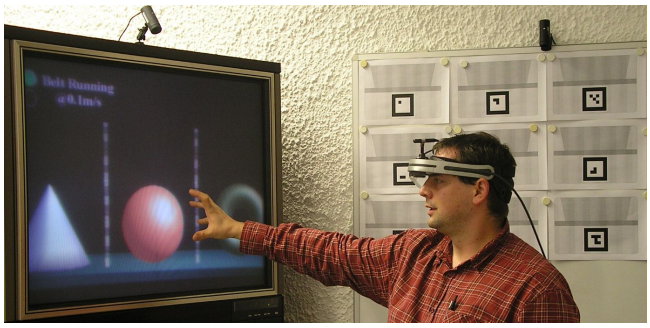

Fig. 4. Overview of virtual production task setup.

This virtual assorting task with given items running on a assembly line turns out to be an adequate environment, since the system has full access to the actual conditions in the work-flow. Herein a worker with some fundamental knowledge about the factory facilities is asked for the first time to work at this specific cite. The assembly line with its icons is screened on a rear-projection TV. Additional information such as the belt's speed can also be displayed here. The object destinations are represented by printed sorting boxes with markers. Acoustical feedback indicates a recognized grabbing and disposal of a virtual object. During all times the worker can hibernate the process by waving with his hands. Over a small numeric keypad tactile commands for controlling the belt speed can be received. Together with a see-through HMD highlighting the actual target this concept can be used for the exploration of chaotic allocation of storage space [1].

Except the HMD, this use case is non-invasive enabling a high degree of physical freedom. With this setup all necessary user surveillance problems are addressed. The work flow is tracked by observing the worker on a behavioral level, i.e gestures, arm movements, eye movements, error rate, response time. Depending on the situational context, the belt can be slowed down after an object is grabbed. Additionally, the level of instructions could be adapted according to the worker's experiences, background knowledge and skills. The following degrees of assistance can be distinguished:

Novice Assistance This assistance grade delivers the highest support for the worker, i.e. the item together with the action grabbing is displayed.

Advanced Assistance For experienced workers the number of support can be reduced.

Expert Assistance In this modus the worker gets only a minimum of instructions, for example when the location of a sorting box has changed.

The task, in the next section represented as a finite state machine, can be divided for a better understanding and structuring into the following steps:

Worker Introduction In this use case the worker is presented the sequence of his todos according to his background knowledge and experiences. Thus, the worker has to be identified and relevant parameters have to be stored before he is well-introduced in this task.

Identification of Objects At this step the worker is instructed about the relevant objects to be collected. This could be done through the visual channel by showing pictures or icons of the objects and complemented with text. For additional information the acoustical channel can be used and special exceptions or conditions can be communicated.

Selection of Objects For selection of the objects the worker has to be advised how to grab the object. Physical criteria of the worker as well as a efficient trajectory will be considered.

Selection of Tray To make sure that the worker drops the object in the right tray he can be supported by the identification of the right tray. Because the trays are ordered in a chaotic way a feedback on its position is necessary.

Dropping of Object Depending on the position and accessibility of the relevant tray the worker could be advised about how reaching it economically.

Pause To make sure that unpredictable situations or machine damages can be handled in a safe way, and to give the worker an ability to initiate a recreation phase, a pause mode has been foreseen.

5. FINITE STATE MACHINE

The above use case is formulated and implemented as a finite state machine (FSM). The graph of the FSM is depicted in Figure 5.

The first step in this graph is the initialization triggered by the arrival of the worker, who will be briefed by a textual and text-to-speech introduction. The user can configure the initial belt speed using a number pad here. By pressing return the sorting loop is started. The hand gesture grasp is accepted by the disposal state, which plays a sound when entered. The gesture dispose closes the loop with an acoustical sign. From both states the break can be reached by waving both hands. From there the setup can be entered again by pressing return or the shift ends by waving both hands again.

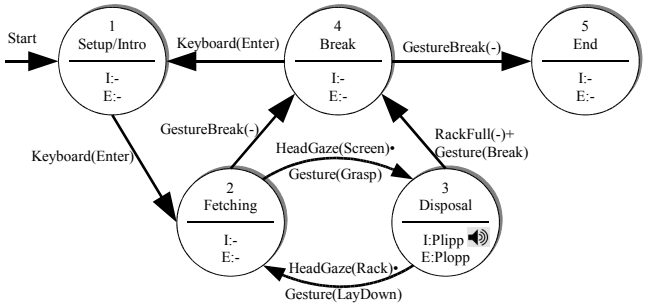

Fig. 5. Finite state machine representation of use case.

6. SYSTEM ARCHITECTURE

The framework for the cognitive system aiming at processing multimodal data consists of three separate functional units, see Figure 6. The entire architecture is driven by the FSM, which programs an event dispatcher according to the actual state context. Due to the distributed nature the communication between these units is chosen to be the Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). The event server itself communicates via UDP with the separate input and output modalities which may be spread over hybrid infrastructures (Linux, Windows, etc.).

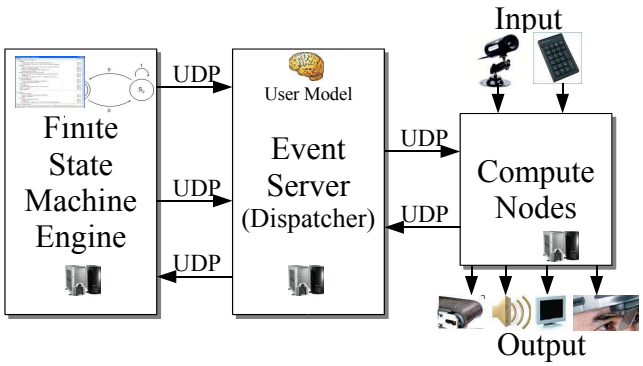

Fig. 6. Internal system architecture.

For our demands, the runtime critical algorithms are programmed in C, whereas the FSM is implemented as a JAVA program. The particular input and output modalities are those presented in section 3.

7. CONCLUSIONS AND OUTLOOK

This ongoing research report faces cognitive technical systems with modern human-machine interaction techniques in production environments. Besides a task dependent scalable system architecture, several input and output modalities for a virtual production task have been integrated into this.

In order to evaluate the facilities and possible drawbacks of this framework, a use case with a virtual sorting task has been introduced. Together with the deployed input and output modalities a FSM was able to track the work-flow steps and give user adapted instructions. Motivated by this positive experience, the next step will be to switch to a more complicated construction task incorporating more modalities.

Parallel to the expansion of the framework and the production task, the system will be expanded with mental models of the user to further improve the adaptation abilities [12]. Aiming at a better understanding of the user consciousness the interpretation of brain activities will also be investigated [13].

8. ACKNOWLEDGMENTS

This work has been funded within the Excellence Cluster CoTeSys by the German Research Foundation (DFG). Furthermore, the authors would like to thank their research partners Anna Schubö, Sonja Storck, Florian Engstler and Rolf Zöllner for their contributions and productive discussions within the ACIPE project.

9. REFERENCES

[1] M. F. Zäh, M. Wiesbeck, H. Rudolf, and W. Vogl, "Virtual and augmented reality," in *Proceedings of Virtual Concept*, 2006.

[2] B. Schuller, M. Ablaßmeier, R. Müller, S. Reifinger, T. Poitschke, and G. Rigoll, *Advanced Man Machine Interaction, K.-F. Kraiss (ed.)*, chapter Speech Communication and Multimodal Interfaces, pp. 141–190, Springer Verlag Berlin, Heidelberg, 2006.

[3] Technische Universität München, "Homepage of the excellence cluster cognitive technical systems," Internet Publication: *http://www.cotesys.org*, 12 2006.

[4] A. Chalechale, F. Safaei, F. Naghdy, and P. Premaratne, "Hand posture analysis for visual-based human-machine interface," in *WDIC 2005 APRS Workshop on Digital Image Computing*, In B. Lovell & A. Meader (Eds.), Ed. Queensland: The Australian Pattern Recognition Society, 2005, pp. (pp. CD Rom 91–96).

[5] T. Möller, H. Kraft, J. Frey, M. Albrecht, and R. Lange, "Robust 3d measurement with pmd sensors," in *In: Proceedings of the 1st Range Imaging Research Day at ETH Zurich, Zurich, Switzerland, pp. "Supplement to the Proceedings"*, 2005.

[6] Metaio, "http://www.metaio.com," .

[7] Microsoft COM: Component Object Model Technologies, "http://www.microsoft.com/com," .

[8] F. Wallhoff, M. Ablaßmeier, and G.Rigoll, "Multimodal face detection, head orientation and eye gaze tracking," in *Proceedings IEEE Conference on Multisensor Fusion and Integration (MFI), 03.-09.09.*, Heidelberg, 2006.

[9] Ronald Azuma, "A survey of augmented reality.," *Presence*, vol. 6, no. 4, pp. 355–385, 1997.

[10] Mihran Tuceryan and Nassir Navab, "Single point active alignment method (spaam) for opticalsee-through hmd calibration for ar," in *Teleoperators and Virtual Environments*, 2002, vol. Vol. 11, No. 3, pp. 259–276.

[11] G. Krost, G. Rigoll, and K. Salek, "Speech synthesis and recognition used for the operator's communication with expert systems supporting power system control," in *Engineering Intelligent Systems*, 2000, vol. 8 of *1*, pp. 11 – 18.

[12] N.A. Taatgen, C. Lebiere, and J.R. Anderson, *In R. Sun (ed.), Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation.*, chapter Modeling Paradigms in ACT-R, pp. 29–52, Cambridge University Press, 2006.

[13] A. Schubö, W. Prinz, and G. Aschersleben, "Perceiving while acting: action affects perception," in *Psychological Research*, 2004, vol. 68, pp. 208–215.