Multimodal Interaction: Methods and Applications for Joint Cooperation between Humans and Cognitive Systems

Gerhard RIGOLL
Institute for Human-Machine Communication
TU München, Munich, Germany

Abstract. In this paper, we present an overview on recent advances in our work on multimodal human-machine communication between humans and technical cognitive systems in order to enable the solution of complex problems that require the joint cooperation of humans and intelligent systems to accomplish challenging tasks that can be only successfully handled if both interact in a cooperative manner. The technical system has cognitive capabilities that support an intelligent interaction with the human operator. An example for such cooperation is the joint assembly of a heavy part by a human assisted by a cognitive robot during the manufacturing process of a complex product. As will be explained, such challenges in joint cooperation occur often in intelligent manufacturing environments which provide a very rich application scenario for the successful combination of advanced multimodal human-machine communication techniques together with cognitive components for improved, cooperative interaction between technical systems and human operators.

Keywords. Human-Machine Communication, Human-Robot-Interaction, Cognitive Systems, Pattern Recognition.

Introduction

Multimodal Human-Machine Communication requires the efficient use of pattern recognition methods in areas such as speech recognition, face and gesture recognition, tracking, emotion recognition and other research fields, mainly in the area of visual communication techniques. In case, that the communication partner of the human operator is a machine with a high degree of autonomous intelligence, such as e.g. a humanoid robot or an intelligent car, such a system can be characterized by the term "Cognitive System". By definition, technical cognitive systems are systems that exploit acquired knowledge to reason and learn from experience about their environment, that adapt to novel situations, explain themselves and are aware about their own capabilities, including the ability to perform skilled actions [1]. Typically, such systems are not only able to communicate efficiently with their environment, but their capability to learn and to act partly autonomously in unrestricted environments make them an interesting partner of humans who can communicate with such systems in order to solve complex tasks, where the cognitive technical system can act as intelligent, cooperative assistant for a human operator. This is illustrated in Fig. 1 that shows the so-called "Perception-Cognition-Action Loop" of Cognitive Technical Systems, where perception is

accomplished by appropriate sensors capturing mainly visual and acoustic input from the user and the environment, cognition is employed for content abstraction and understanding of the information gained from the perception component and the system responds by an action that is often the result of intelligent reasoning utilizing the knowledge-base of the system.

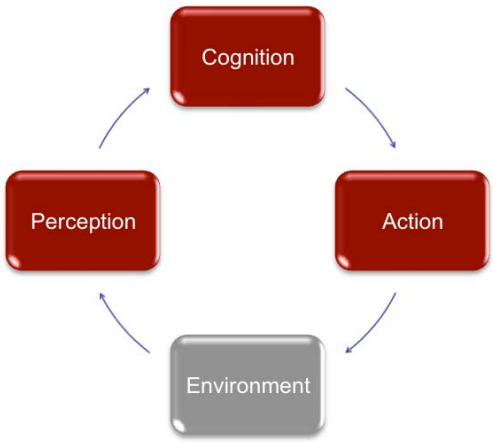

Fig. 1. Perception-Cognition-Action Loop

From the viewpoint of Human-Machine Interaction, obviously the perception module in Fig. 1 is the major topic of interest addressed in this paper, involving mainly recognition techniques from statistical pattern processing. Although it is not possible to present in this paper a complete overview of the major algorithms that are employed for such recognition methods, it is worthwhile to mention the rising popularity of statistical pattern recognition algorithms for such tasks, which have been very much influenced by the use of Hidden Markov Models (HMMs, see [6]) in the last decades and are now more and more dominated by the unifying framework of Graphical Models (GMs, see also [7]). GMs have the advantage that they can incorporate almost all important statistical method in one single representation, including even rule-based approaches. As an example for this modeling power, Fig. 2 shows the implementation of different popular HMM-types by GMs. In this case, the recognition process is typically carried out by computing the joint probability between the time sequence of visited nodes x_t and the corresponding observed labels y_t (represented by the feature vectors derived from the observed audio or video stream that shall be recognized) as follows:

$$p(x, y) = p(x_0)\prod_{t=0}^{T} p(x_{t+1} \mid x_t)\prod_{t=0}^{T} p(y_t \mid x_t) \tag{1}$$

where the desired observation probability p(y) can then be computed by marginalization over all possible state variables. Training is typically performed by the more general Eq. (2)

$$p(x_t \mid Y_1^t) = p(y_t \mid x_t) \cdot \int p(x_t \mid x_{t-1}) \cdot p(x_{t-1} \mid Y_1^{t-1})\ dx \tag{2}$$

and one can observe that this is a more generalized case of the well-known Forward-Backward algorithm employed for HMMs, as expressed in the following equation:

$$\alpha_t(j) = p(o_t \mid x_t) \cdot \sum_i a_{ij} \cdot \alpha_{t-1}(i) \tag{3}$$

In this case, the forward probabilities α_t in Eq. (3) correspond to the probabilities $p(x_t|Y_1^t)$ which represent the state probabilities given the observed feature label stream $y_1, y_2, \ldots y_t$ at time t. The state transition probabilities $p(x_t \mid x_{t-1})$ in Eq. (2) correspond to the HMM transition probabilities a_{ij} in Eq. (3).

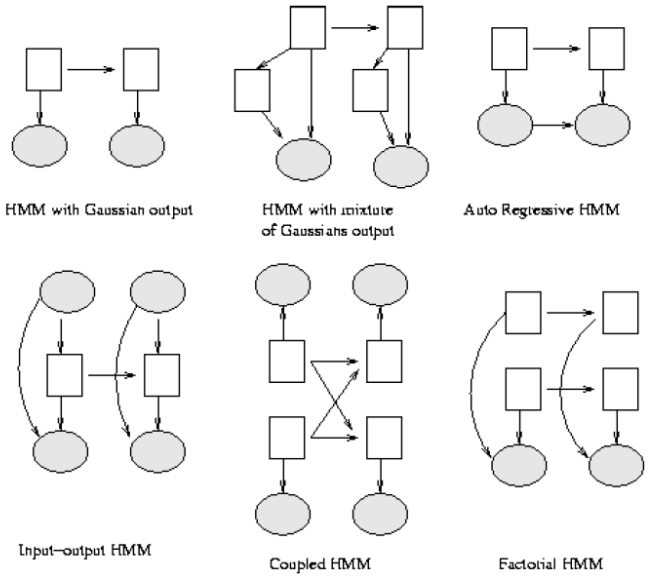

Fig. 2. Popular HMMs implemented as Graphical Models

Let us now move to a more concrete example for such a "Perception-Cognition-Action Loop" which would be the joint cooperation between a human operator and a cognitive system, in order to accomplish a specific task. The technical system performs perception via its sensors in order to take commands from the user and observe his behavior, it makes use of its cognitive capabilities to figure out how it can support the user and performs an appropriate action in order to assist the human in performing the joint task.

Fig. 3 shows an example for such a cooperation between a user and a cognitive system, which is in this case represented by an intelligent robot with cognitive

capabilities. Although our approach considers generally the cooperation between humans and any kind of technical cognitive system, it is in fact true that cognitive robots are especially suitable for our studies since they represent intelligent systems which have not only sophisticated perception and recognition capabilities, but additionally can perform human-like actions, such as e.g. grasping, moving or walking and therefore are indeed suitable technical partners for solving complex tasks in cooperation with humans [2].

The following sections describe the various aspects which have to be considered in order to design cooperative cognitive systems that are capable of solving complex tasks within specifically defined environments and highlight the challenges that come along while pursuing the related research goals.

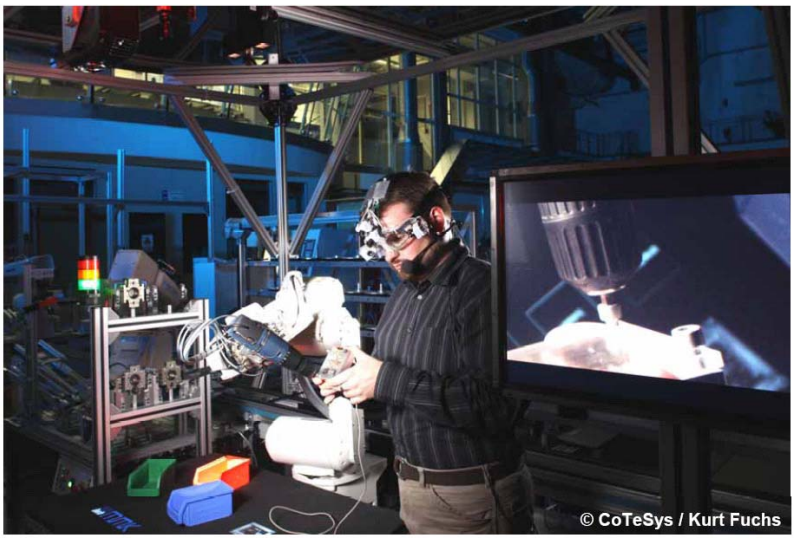

Fig. 3. Joint cooperation between human and cognitive robot

1. Cognitive Systems in Intelligent Manufacturing Environments

As previously mentioned, complex joint cooperation between humans and cognitive systems can be currently achieved only within quite specifically predefined environments. A proper selection of the application domain that automatically defines such a specific environment is therefore the first important step for the realization of such systems. We decided to choose an intelligent manufacturing environment (which we also call “Cognitive Factory” [3]) as most suitable application scenario for the following reasons:

- In manufacturing environments, humans as well as machines play an important role and need to share a common workspace which requires in many cases intensive interaction between users and machines.
- Especially in manufacturing environments with robots, human-machine collaboration typically bases on a master-slave level where the human worker

operates the robot or programs it off-line allowing it to execute only static tasks. To ensure safety, the workspaces of humans and robots are strictly separated. In many cases, human workers are completely excluded from the production lines, where automated robots are executing assembly steps. On the other hand robots are not integrated in assembly line manufacturing along with human workers.

- Such approaches do not take advantage of the potential for humans and robots to work together as a team, where each member has the possibility to actively assume control and contribute towards solving a given task based on his capabilities. Such a mixed-initiative system would allow the human and robot to support each other in different ways, as needs and capabilities change throughout a task. With the subsequent flexibility and adaptability of a human-robot collaboration team, production scenarios in permanently changing environments as well as the manufacturing of highly customized products would become possible.
- A "Cognitive Factory" can therefore provide a smart working environment in which a human and an industrial robot can work together on a peer-to-peer level. Potential applications lie in an efficient collaboration of human and robot in production processes e.g. in order to semi-automate complex construction tasks, or enable different order in part assembly to support human workers and provide them with a higher degree of flexibility.
- This scenario has furthermore high practical importance and relevance for a large number of potential users since it represents a typical realistic working environment for large parts of the labor force, along with a huge economical impact.

After having chosen the "Cognitive Factory" as suitable scenario for joint human-machine cooperation for complex task solving, the research goals resulting from this scenario can be defined as follows:

- Exploration of multiple sensor input, including cameras, infrared cameras, and laser-scanners to determine the current state of the user, to infer the next action step (anticipatory behavior) and to ensure safety for the human co-worker (reactive behavior).
- Exploitation of multiple modes, such as haptics, hand gestures, body motion or gaze, for robust intention and action recognition in adverse environments with a high degree of noise, changing illumination conditions and disturbations.
- Employment of adaptation and machine learning principles in order to cope with the above mentioned flexible environmental and logistic conditions.
- Implementation of intelligent actions for the involved machines in order to support the current assembly step invoked by the user in the most possible flexible manner.
- Incorporation of cognitive processes in the planning and realization of such actions, by employment of appropriate knowledge sources and implementation of intelligent behavior.

This leads to the functional system overview as shown in Fig. 4, where the major function blocks of the entire system are displayed. They can be roughly subdivided into components for multimodal input, interpretation, multimodal output, and all modules that are responsible for machine intelligence of the cognitive system, e.g. the reasoning component, the dialogue management and knowledge representation.

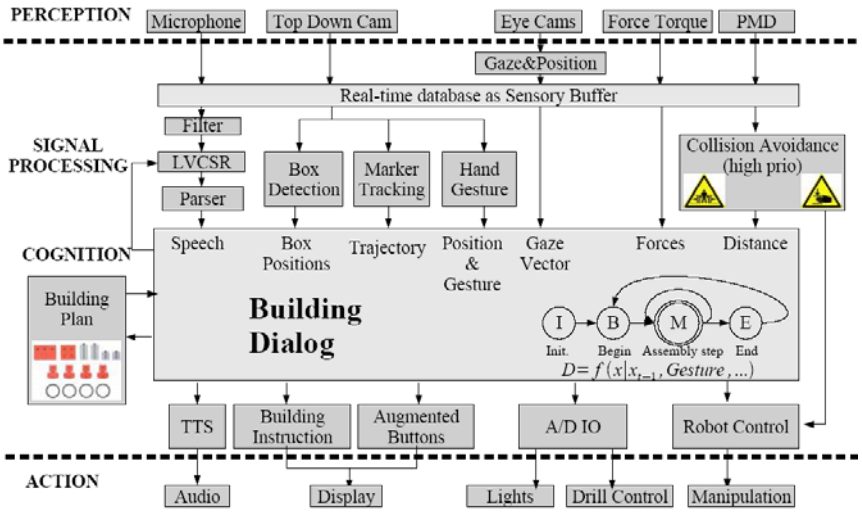

Fig. 4. Functional system overview with modules for perception, cognition and output generation

2. Multimodal Recognition of Human Behavior and Activity Detection

As outlined in the previous section, one of the basic crucial capabilities of cooperative technical systems should be the ability to perceive the presence of one or several human operators within their surrounding, in order to observe their actions, to anticipate their intentions and to communicate with the humans. Therefore, it is obvious that accurate detection of humans and tracking their motion is a mandatory basis for the perception of human presence for assistive cognitive systems. The technical requirements are in this case very high, since illumination conditions can change any time, the background can be very complex and could contain additionally moving objects, since the cooperation typically takes place within a complex assembly scenario, as e.g. shown in Fig. 5.

There, it becomes obvious, that the detection of a human hand within an assembly environment is substantially complicated, for instance through the presence of a variety of colored boxes containing diverse assembly parts. Hand detection is carried out in this case by a top-down view camera using color- and motion-based segmentation. Since it is not only required to detect the human hand, but also to determine which parts have been taken out of the boxes during an assembly cycle, the complete scene has to be captured, interpreted and the grasping action has to be identified together with the position of the box where a particular assembly part has been taken out.

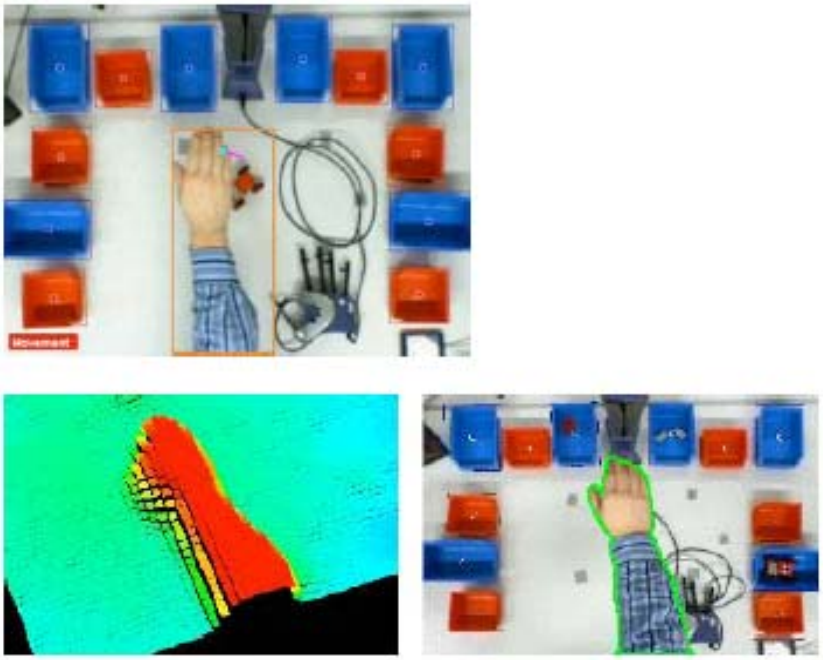

Fig. 5. Color- and motion-based segmentation to detect arm and assembly part box positions

In order to make this observation task more reliable and faster, 3D vision techniques are additionally deployed, supported by a Photonic Mixer Device (PMD [4]). The PMD can be considered as infrared sensor which delivers a depth map. Each pixel value yields the distance between the camera and the associated area of the object. In this way, the acquired 3D-information can be useful to identify objects with pre-known significant heights, such as e.g. boxes and can help to distinguish these objects from human parts, such as hands or arms.

Another important input modality in this scenario is the use of eye tracking in order to detect the gaze of the human operator. Fig. 6 shows the detection of the pupil position with the special device developed and used in our project [5]. With this method, it is possible to detect the focus-of-attention of the human operator, which is an important information for the cooperative cognitive system, e.g. to detect which tool or assembly part a worker wants to get handed over from an assistive humanoid robot. Within our project, we have developed several gaze tracking modules, starting from a head-mounted gaze tracking system that exploits the so-called cornea reflex resulting from light reflection, up to a remote gaze tracker using the principle of conic reconstruction based on stereo image capture of the eyes. The latter has a remote operating range of 0.4 up to 1.3 meter, which is sufficiently large in order to enable an interaction via gaze with a natural distance between user and technical system.

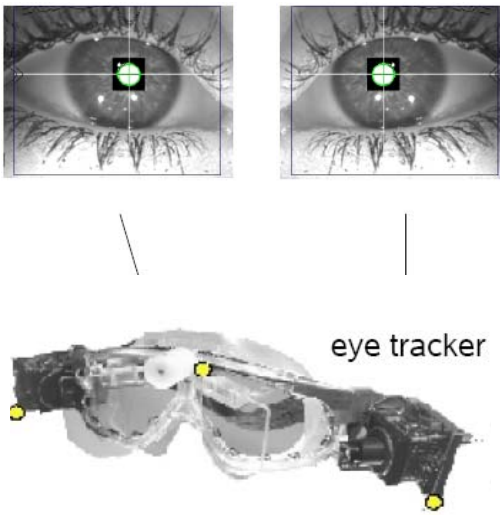

Fig. 6. Eyetracking device and result of image processing to detect the position of the pupil

3. Cognitive Information Processing and Knowledge Representation

A cognitive cooperative system should not only have perceptive capabilities as described in the previous section, but additionally requires a huge amount of knowledge that represents the basis of its cognitive task solving capabilities. Cognition is then mainly performed by exploiting the available knowledge-base and performing appropriate reasoning. Suitable knowledge structures are required in the following areas:

- general domain knowledge of the targeted scenario
- structure of the assembly environment
- geometrical data of the entire scene, assembly parts and tools
- performance of possible tasks and workflows
- possible behavior of the human partners
- possible actions according to specific input

as well as control of the dialogue for joint human-machine cooperation. An example for such a knowledge source is shown in Fig. 7, where a tree-structure is displayed that shows alternatives about how single assembly parts can be grouped into intermediate modules that eventually form the final product in the root of the tree.

Knowledge representation is not the only required basis in order to enable the system to behave in an intelligent manner and to incorporate cognitive capabilities. One of the important goals in assembly is e.g. to design this task as flexible as possible so that a human can decide himself about the assembly order and the cooperative

cognitive system is still able to assist him and thus to adapt itself to possible sudden changes in the assembly operation. For that purpose, the system needs knowledge about the assembly order and all possible deviations from the default order. It needs to know the final configuration of the end product in order to determine if the final product is finished or parts are still missing, and also the parts that can be picked to start the assembly task. Furthermore, the system needs also the appropriate reasoning methods for decision making to decide about the next action step in a specific state of the assembly process.

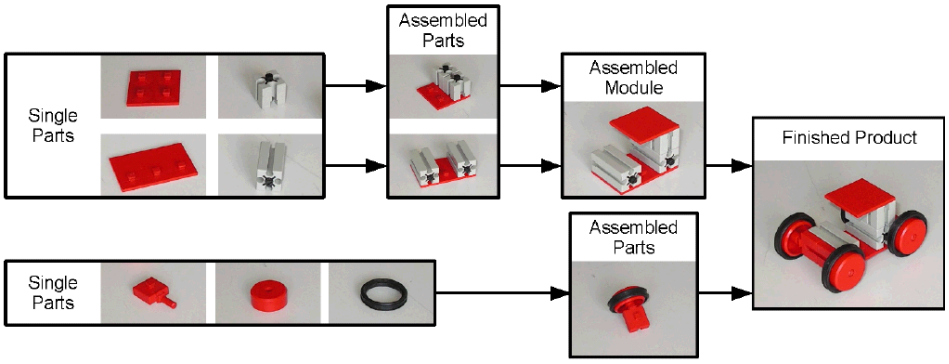

Fig. 7. Tree-structured assembly plan for a toy product

4. Dialogue Management and User Guidance

Finally, as in the case of most interactive system architectures, the management of the dialogue plays an important role for the usability and user friendliness of the cooperative system. Without such component, the human operator would not be able to cooperate with the system in an efficient manner. Fig. 8 shows the basic architecture of the dialogue manager, functionally embedded into the Perception-Cognition-Action Loop. As can be seen, the dialogue manager is involved in all function blocks of this cognitive loop. It can be in fact considered as the high level control component of this loop. Perception is accomplished by the modalities described in the previous sections, mainly using different camera sensors and gaze information. One of the major components of the dialogue manager is concerned with cognition, since cognition here is not only the reasoning about the best possible assistance of the user but also is responsible for the interpretation of the user's input command. Action is controlled and generated by the dialogue manager in different ways. One is the action for assisting the user in his task, which is to be generated by actuators, e.g. robot arms in case of a handing-over task or drilling devices in case of the assembly task. Furthermore, possible system dialogue actions consist also of typical multimodal information output, such as spoken messages for confirmation or user guidance and graphical information display.

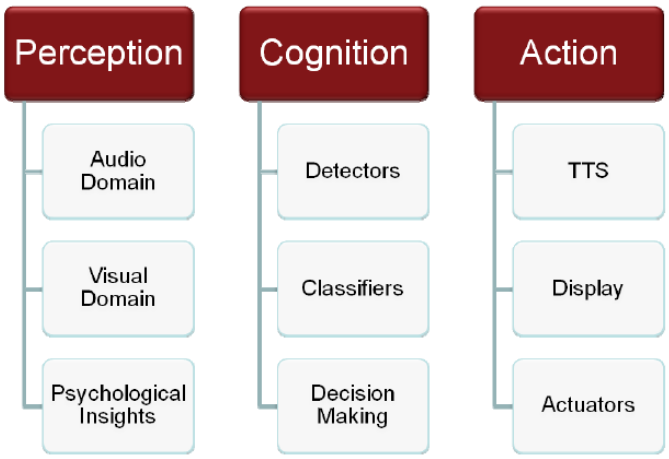

Fig. 8. Basic modules of the dialogue manager for cognitive assistive systems

In case of joint human-machine cooperation, output generation plays are more important role than in many other classical human-machine dialogues, since the human partner has to be precisely informed about the recent action of the cognitive system, about its current state and – especially in complex assembly scenarios - he might also need to have some guidance about his next possible actions and the current progress of the assembly task.

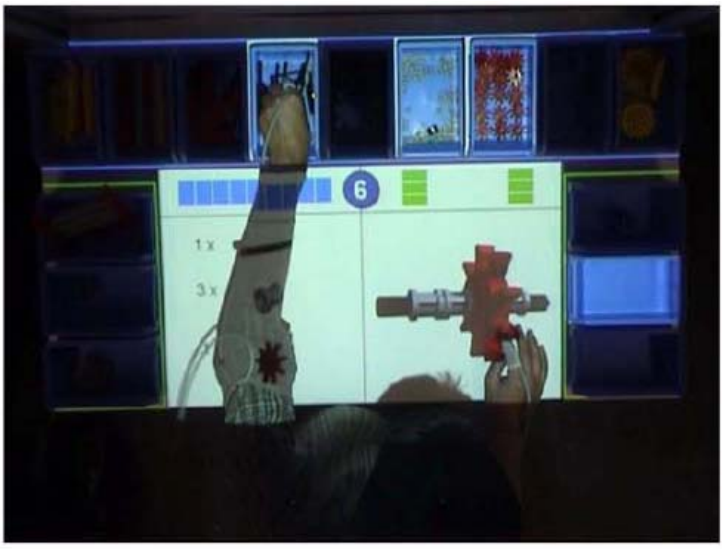

Fig. 9. Table projection for user feedback and guidance

One interesting output modality is in this case the table projection, as shown in Fig. 9. In this case, the required information can be elegantly projected onto the surface of

the workbench with a video projector mounted above the work area. This opens up an entire spectrum of interesting options for information display to the user which resembles quite closely the well-known Augmented Reality technology, where e.g. contact analog instructions like arrows or highlighted zones and assembly parts can be merged with the real scenario on the workbench. Other output modalities, such as e.g. classical flat screen display or even HMD (Head Mounted Display) have been investigated, but the above mentioned table projection has been by far the most effective method for user feedback and user guidance in this case.

5. Conclusion

We presented an overview on our activities to develop cognitive technical systems that are capable of cooperating jointly with humans in order to solve demanding complex tasks in a selected manufacturing environment, called the Cognitive Factory. It has been outlined in this paper that such systems require the complete repertoire of multimodal recognition techniques, advanced output generation, as well as intelligent knowledge processing in order to be able to assist human workers in assembly tasks with and without additional support of robots. Obviously, such systems are still in their preliminary application phase and many further details and challenges that require advanced multimodal interaction methods will have to be solved in the future.

References

[1] CoTeSys: Excellence Cluster "Cognition for Technical Systems", *www.cotesys.de*

[2] C. Lenz, N. Suraj, M. Rickert, A. Knoll, W. Rösel, J. Gast, A. Bannat, and F. Wallhoff. Joint-Action for Humans and Industrial Robots for Assembly Tasks. *Proc. 17th IEEE Intern. Symposium on Robot and Human Interactive Communication, ROMAN 2008*, Munich, Germany, pp. 130–135.

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

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

[5] S. Kohlbecher, S. Bardins, K. Bartl, E. Schneider, T. Poitschke, M. Ablaßmeier.Calibration-free eye tracking by reconstruction of the pupil ellipse in 3D space. *Proc. of the 2008 Symposium on Eye Tracking Research & Application, ETRA-08*, Savannah, Georgia, USA, pp. 135–138. ACM Press, NY, 2008.

[6] L. R. Rabiner: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE, Vol. 77, No. 2, pp. 257–286, 1989

[7] Koller; Friedman. *Probabilistic Graphical Models*. Massachusetts: MIT Press, 2009