An Integrated System for Cooperative Man-Machine Interaction

C. Bauckhage, G.A. Fink, J. Fritsch, F. Kummert, F. Lömker, G. Sagerer, S. Wachsmuth
Technical Faculty, Bielefeld University
P.O. Box 33594, Bielefeld, Germany
`http://www.techfak.uni-bielefeld.de/techfak/ags/ai/`

Abstract To establish robotic applications in human environments as e.g. offices or private homes the robotic systems must be instructable by ordinary users in a natural way. In interpersonal communication humans usually apply different sensory information and are capable of integrating all perceptual cues fast and consistently. Additionally, knowledge acquired during the communication process is directly used to resolve ambiguities. As a step towards realizing similar capabilities in automatic devices this paper presents an integrated system combining automatic speech processing and image understanding. The system is intended to be an intelligent interface of a robot which manipulates objects in its surroundings according to the instructions of a human. The enhanced capabilities necessary for carrying out a multimodal man-machine dialog are realized by combining statistical and declarative methods for inference and knowledge representation. The effectiveness of this approach is demonstrated using an examplary dialog from our construction task domain.

1 Introduction

Perceptual front-ends for robotic systems are becoming increasingly important as robotics research is gradually moving from factory floor applications to domestic environments. In this paper we present a system integrating state-of-the-art modules for digital image and spoken language processing which is intended to be the intelligent user interface of an *artificial communicator*. This term characterizes a machine that performs visual and acoustic recognition tasks and is able to combine the resulting symbolic information to communicate and act appropriately. Users may talk to an artificial communicator in an everyday fashion. Additionally, while communicating the machine is capable to extend predefined knowledge of its surroundings, i.e. it can learn. Learning systems with sophisticated perceptual capabilities enable intuitive interaction even for an unexperienced user and they easily adapt to new tasks. Therefore, artificial communicators seem to be the natural choice in applications like for instance service robotics (cf. e.g.[1, 17]).

The work reported here focuses on the problem of integrating different modalities for automatic perception which has gained attention nowadays (see e.g. [16]). A recent workshop has documented the state-of-the-art in this field [21].

Due to the complexity of human cognition the realization of a communicating system must be restricted to a particular domain. Our system aims at intelligent support for flexible assembly. It is designed to enable the cooperative construction of objects like toy-airplanes or toy-trucks using parts from a wooden construction-kit for children. This scenario allows to study all the principles underlying the integration of different types of perception. The techniques proposed in the following point out practical ways towards intelligent man-machine interfaces.

2 System Architecture

Architecture is one of the key issues in designing intelligent systems. Different types of signals being generated at different time scales have to be integrated to guarantee a fast and consistent system reaction. Well known approaches to this problem are for instance behavior-based realizations, agent-based concepts, or blackboard architectures (cf. e.g. [2, 6, 10] respectively). Our system is based on both hierarchical and distributed representations of knowledge sources and therefore cannot be categorized into one of these paradigms.

2.1 Module Interaction

According to the perceptive capabilities the system consists of different components performing the necessary computations. As shown in Fig. 1 the individual modules are arranged as two tracks of speech processing and image understanding tasks converging in an integrating component.

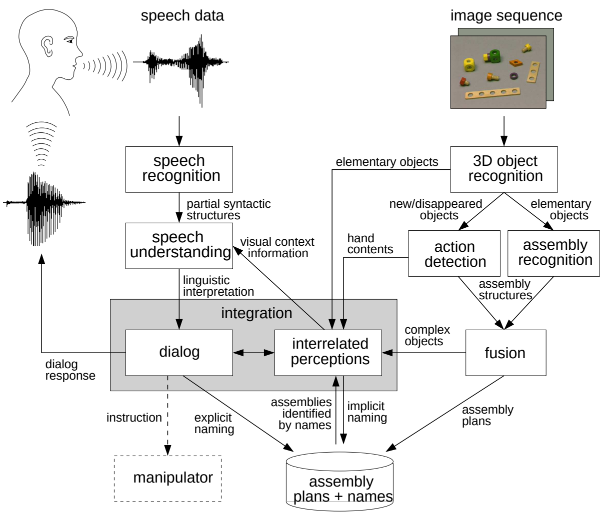

Figure 1: Module interaction in the proposed system architecture.

Given an acoustic signal the left track starts with a speech recognition procedure. The corresponding results are transferred to an interpretation module whose results are used in a component that controls the dialog with a user.

In IEEE International Symposium on Computational Intelligence in Robotics and Automation, pages 328-333, Banff, Canada, 2001.

In the image understanding track an object recognition procedure is used to recognize elementary objects of our scenario. Stereo image data allows to estimate the spatial position and orientation of detected elements. This information is monitored over time so that the system can determine if and how objects in its surroundings are moved. The memorized object data is passed on threefold: first, it is scanned by a process which detects and describes assembly structures. Secondly, it is used to detect basic construction actions. Thirdly, the memorized objects are sent to the module interrelating the perceptive modalities. In the component labeled *fusion* the assembly structures extracted from action detection and assembly recognition are fused yielding assembly plans that are stored in a database.

Integration of different perceptive modalities is the main aspect of our system. Accordingly, the module that interrelates results from speech and vision processing plays a central role giving the system access to related acoustic and visual information. The interrelation module uses the database to retrieve an intended assembly given an abstract name. Explicit and implicit naming of assemblies is realized via connections from the dialog module and the interrelation module to the database, respectively.

2.2 Distributed Development Environment

As an integrated architecture comprising many computationally demanding sub-tasks can never be realized in a monolithic system a distributed processing approach has to be adopted. Although there exists a large variety of approaches for building distributed systems they all suffer from inherent limitations when applied to problems of multi-modal signal analysis. Therefore, a new approach for a simple but powerful framework to integrate large distributed pattern analysis systems was proposed: the Distributed Applications Communication System (DACS) [13].

A communication demon on each participating machine acts as router for all internal traffic and establishes connections to demons on remote machines. Communication is based on simple asynchronous message passing with some extensions to handle dynamic reconfigurations of the system during runtime. On top of this elementary layer DACS provides some more advanced communication semantics like synchronous and asynchronous remote procedure calls and *demand streams*. The latter allow to set up the data-driven computations in the lower level vision and speech processing modules easily in just the same way as a producer-consumer interaction. Resources generated by some module can be accessed via a symbolic name transparently by any number of consuming modules which in turn can offer their processing results to other components of the overall distributed system.

The modules described in this paper are running concurrently on a cluster of several COMPAQ Personal Workstations 500 (19.5 SPECfp95) and one SGI workstation for the visualization. Images are taken by means of Sony XC-007P cameras and AV321 grabbing devices. For speech recording we use a Sennheiser wireless microphone system 1083-U and a Gradient Desklab 216 for digitalization. Speech synthesis is done via a commercial text-to-speech system which is based on an OKI-Chip.

3 Speech And Language

To allow a fluent communication between an instructor and an artificial communicator our system is capable of understanding speaker independent speech input. The instructor neither needs to know a special command syntax nor the exact terms or identifiers of the objects. Consequently, the system has to face a high degree of *referential uncertainty* from vague meanings, speech recognition errors, un-modeled language structures, and, on the visual input channel, erroneous object classification results.

Our approach to robust spoken language understanding uses a vertical organization of knowledge representation and an integrated processing scheme to overcome the drawbacks of the traditional horizontal architecture [7]. As baseline module [12] it makes use of an enhanced statistical speech recognizer. The recognition process is directly influenced by a partial parser which provides linguistic and domain-specific restrictions on word sequences [24]. Therefore, partial syntactic structures instead of simple word sequences are generated, like e.g. object descriptions ("the yellow bolt") or spatial relations ("...in front of..."). These can easily be combined by a subsequent speech understanding module to form linguistic interpretations. Although there has been some progress recently, the detection of *out-of-vocabulary* words can still not be performed robustly on the level of acoustic recognition. Therefore, we employ a recognition lexicon which exceeds the one used by the understanding component but covers all lexical items frequently found in our corpus of human-human and human-machine dialogs. The syntactic modeling then allows to use these additional words to be filled in for open lexical categories as e.g. nouns.

Since spontaneous speech input largely deviates from speech read from e.g. text prompts acoustic models for spontaneous speech phenomena, namely for so called *human noises* (e.g. breathing or lip smacks) and hesitations (like e.g. `uhm`), are part of the recognition lexicon.

Syntactic phenomena frequently found in spontaneous German speech are repairs and extrapositions in which information given in one constituent is specified or corrected by subsequent material. As mentioned above only domain specific constituents are derived without any restriction on their order. Consequently, it is necessary to resolve repairs and extrapositions by inferring the possible references between both participating constituents. The resolution of repairs and extrapositions is done by detecting them, finding a suitable reference constituent and integrating them into the reference constituent. For this purpose a modified LR(1)-parser and a unification grammar are used [14]. The approach is robust to out-of-vocabulary words, to speech recognition errors due to morphologically similar words, and to deletions of function words.

An example of a correction is given by the first utterance of the example (Fig. 5) where the correcting constituent *"the orange one"* is integrated into the antecedent utterance *"Take the red bolt"* to yield the correct interpretation *"Take the orange bolt"*. To avoid the integration of contradictory information visual context information is integrated into the parsing process by using the techniques to interrelate different perceptions described in section 5.

4 Objects, Assemblies and Actions

It is widely agreed that sight is the most important of the human senses. Correspondingly, computer vision research has a long tradition (cf. e.g. [3, 11]). Using computer vision to monitor assembly, however, is scarcely explored by now [18] and remains an interesting topic of research, especially if the system is to cooperate with a human instead of operating in a restricted industrial environment.

4.1 Object Recognition in Image Sequences

Since in our domain only a fixed set of different fundamental colors occurs the recognition of elementary Baufix objects is based on regions of homogeneous color.

The recognition process employs explicit object models which are represented using the semantic network language Ernest [20]. Every object is modeled by its parts and is embedded in a specialization hierarchy, e.g. bar is a generalization of 3-holed-bar. A classifier calculates up to three competing object hypotheses for every region and depending on the object type an appropriate specialization is selected which is verified according to the structural knowledge represented in the semantic network. During this instantiation process attributes like position, color, and shape are calculated and propagated in a model-driven way [15]. The structure of the semantic network is retained for the next time step to enable a fast processing. By this technique only the evaluation of the structural knowledge has to be carried out in every time step.

In order to enable a robot to approach an object of interest we realized a coarse but efficient 3D-reconstruction method. The object recognition process outlined above is applied independently on a stereo image pair. For the detected objects a stereo matching is computed. On the basis of these matches and of the calibration data a depth value for the center of mass of every object is calculated. The orientation of an object is estimated by aligning the main axes of the detected object and the object model.

4.2 Recognition of Assembled Objects

For the recognition of assemblies we developed a syntactic method to detect and describe structures made from elementary objects. It is based on the idea that every bolted assembly necessarily consists of a bolt and a nut with some miscellaneous parts in between. All three components are either given by elementary objects or by subassemblies leading to a cyclic model (see Fig. 2).

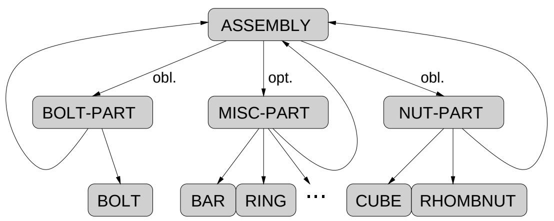

Figure 2: Recursive model of bolted assemblies: bolt- and nut-part are obligatory components, the misc-part is optional.

Implementing this model as a semantic network using Ernest and sequentially analyzing clusters of detected elementary objects allows to derive structural descriptions of assembled objects as shown for the object called *"cockpit"* in the lower-most image in the fourth column of Fig. 5. It consists of a bolt, a bar and a complex nut which is represented by the *"engine block"*.

Moreover, examining the topological appearance of a cluster of objects yields detailed information about the mating relations among the objects. These relations specify which holes of a bar or a cube are occupied by bolts. Therefore, structural descriptions supplemented by mating relations enable the generation of assembly sequence plans [5]. The system employs these plans to virtually reconstruct assemblies detected in an image (see Fig. 5).

4.3 Detection of Assembly Operations

The previously presented image-based recognition of assemblies reaches best performance if all single objects are visible and recognized correctly. Due to perspective occlusions these prerequisites can not be guaranteed for all objects within a complex assembly. Therefore, we developed a second method to extract assembly structures by observing the construction process leading to an assembly. As the visual observation of actions is computationally expensive and difficult, we use symbolic information about the temporal changes of objects, i.e. new objects being put into the scene and disappeared objects being taken out of the scene to infer actions [4]. Therefore our monitoring approach is independent from whether a human or a robot acts in the scene and can reuse results from object recognition.

A two-hand model is used to reflect the actual contents of the hands which are usually outside the field of view. Based on rules with prerequisites for the hand states we hypothesize actions if object changes occur. Currently we have rules to infer *Pick*, *Place* and *Connect* actions. An action is hypothesized and the hand states are changed accordingly if an object change is detected and the preconditions of the appropriate rule hold, e.g. one hand is empty.

After a sequence of construction actions has been detected the hand models contain the structure of the assembly. While this information is quite reliable regarding the object types and their order in the constructed assembly, the connection details are unknown due to the missing visual observation. After placing the assembly back into the scene, the visual assembly recognition obtains a second assembly structure including connection details, but here some parts may be missing due to visual occlusion. By fusing the two assembly structures tiny objects which are occluded in the image of the assembly can be inserted into the final assembly structure as they are contained in the assembly structure from action detection.

5 Interrelated Perceptions And Dialog

In the two preceding sections we have presented the speech and vision modules of our system. These modules extract interpretations of utterances and structural descriptions of objects from the input signal. To perform construction tasks a dialog with an instructor must be enabled and his verbal instructions must be executed. Therefore, links between the analyzed speech and objects in the scene must be established. In this section we describe the two modules responsible for this.

5.1 Interrelating Speech and Vision

If a naive user describes an object in the scene by using attributes he will typically use another vocabulary than the fixed vocabulary appropriate for processing of visual data. Therefore, several kinds of uncertainties have to be considered when correlating a verbal object description and object recognition results, such as vague attributes (e.g. *"the long, thin stick"*), vague spatial and structural descriptions (e.g. *"the object left to the cube", "the cube with the bolt"*), or speech and object recognition errors.

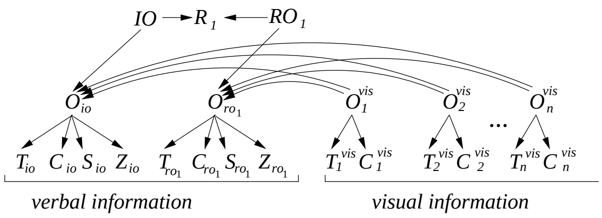

Figure 3: Bayesian network connecting speech and vision data for one related reference object.

In order to cope with uncertainty, we have developed a Bayesian network approach that robustly combines verbal and visual information through different abstraction levels [23]. On the first level, basic features from vision (C^{vis} *:color,* T^{vis} *:elemental type*) and speech (T *:type,* C *:color,* S *:shape,* Z *:size*) are modeled as evidential nodes of the Bayesian network for each of the n visual objects and N verbally referenced objects. On the second level, these are fused to the *visual object class* $O^{vis}_{k\in\{1,\ldots,n\}}$ and *verbal object class* $O_{io/ro_j, j\in\{1,\ldots,N-1\}}$ which are connected by the intended object and reference object variables $IO, RO_j \in \{1\ldots n\}$. The verbally mentioned spatial or structural relations between them are established by introducing additional evidential nodes R_j (Fig. 3). The different kinds of uncertainties are modeled by conditional probability tables that have been estimated from experimental data [23]. The objects which are denoted in the utterance are those explaining the observed visual and verbal evidences e^{vis}, e^{verb} in the Bayesian network with the maximum a posteriori probability, e.g. for the sentence *"the red ring in front of the bar"*:

$$(io^*, ro_1^*) = \underset{io,ro_1,io\neq ro_1}{argmax} \; P(io, ro_1|e^{vis}, e^{verbal}) \quad \text{where}$$

$$e^{verbal} = \{C_{io} = \text{red}, T_{io} = \text{ring}, T_{ro_1} = \text{bar}, R_{io,ro_1} = \text{infront}\}$$

$$e^{vis} = \{C_1^{vis} = \text{red}, T_1^{vis} = \text{felly}, C_2^{vis} = \text{wooden}\ldots\}$$

The intended object io^* is used by the dialog component for system response and manipulator instruction.

5.2 Dialog

In the last years, many spoken dialog systems have been developed (e.g. [9, 19]). However, most of these systems lack the integration with other modalities. In contrast to these approaches our dialog module integrates utterances of the instructor, information of the visible scene, and feedback from the robot to realize a natural, flexible and robust dialog strategy.

The dialog module is realized within the semantic network language ERNEST using the dialog model shown in Fig. 4. The model is based on an investigation of a corpus of human-human and simulated human-machine dialogs. Every path through the model reflects a course of a possible human-machine dialog. The admissible sequence of intermediate states is nearly unrestricted leading to a very natural and robust dialog behavior [7].

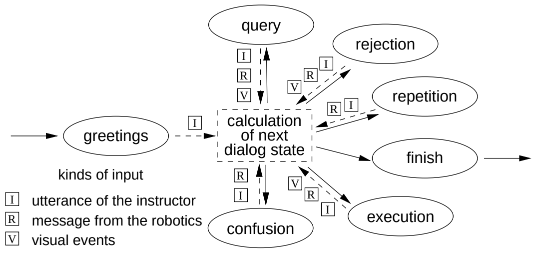

Figure 4: Model of the man-machine communication.

State transitions are initiated if new information from the instructor or, if connected to the system, the robotics is available. The state transition function analyzes the new information and combines it with the current dialog context and information gathered from the interrelation module to select the next state.

Using the dialog context references between objects can be resolved, e.g. *"Take the red bolt. Put* it *into the bar."* and information accumulated in the dialog can be combined. The dialog module can react on new information from different modalities to e.g. inform the instructor about errors during the execution of an action and can actively control the dialog to e.g. query for missing or unprecise information. The overall goal of this module is to continue the dialog in every situation. Actions which can not be executed are immediately rejected. For verbal instructions which could not be analyzed a repetition is requested up to two times. If the dialog has gathered completely contradictory information it expresses its confusion and asks for a completely new instruction.

6 A Performance Example

In order to emphasize the potential of intuitive man-machine interaction we present an experiment carried out with our system. Besides pointing out common problems in integrated speech and image understanding the experiment illustrated in Fig. 5 characterizes the current performance of the system. The example reflects the systems competence in speech understanding, scene interpretation and learning of new objects.

In Fig. 5 every row of pictures shows: the view of the scene together with the object labels, a virtual representation of the internal state of the system before the instruction, the user instruction and the system state after the instruction. Whenever the instructions require the manipulation of a certain object by the robot the corresponding virtual object is highlighted in the fourth column. In the virtual representations the objects are arranged according to rough estimates of the object positions from the 3D-reconstruction.

The interaction consists of five stages:

1) Initially, a couple of objects and an assembly are spread arbitrarily within the visible area of the system. The first instruction contains a verbal correction resulting in the orange bolt being intended by the instructor which is highlighted in the virtual representation in the fourth column.

2) After removing the bolt from the scene the situation

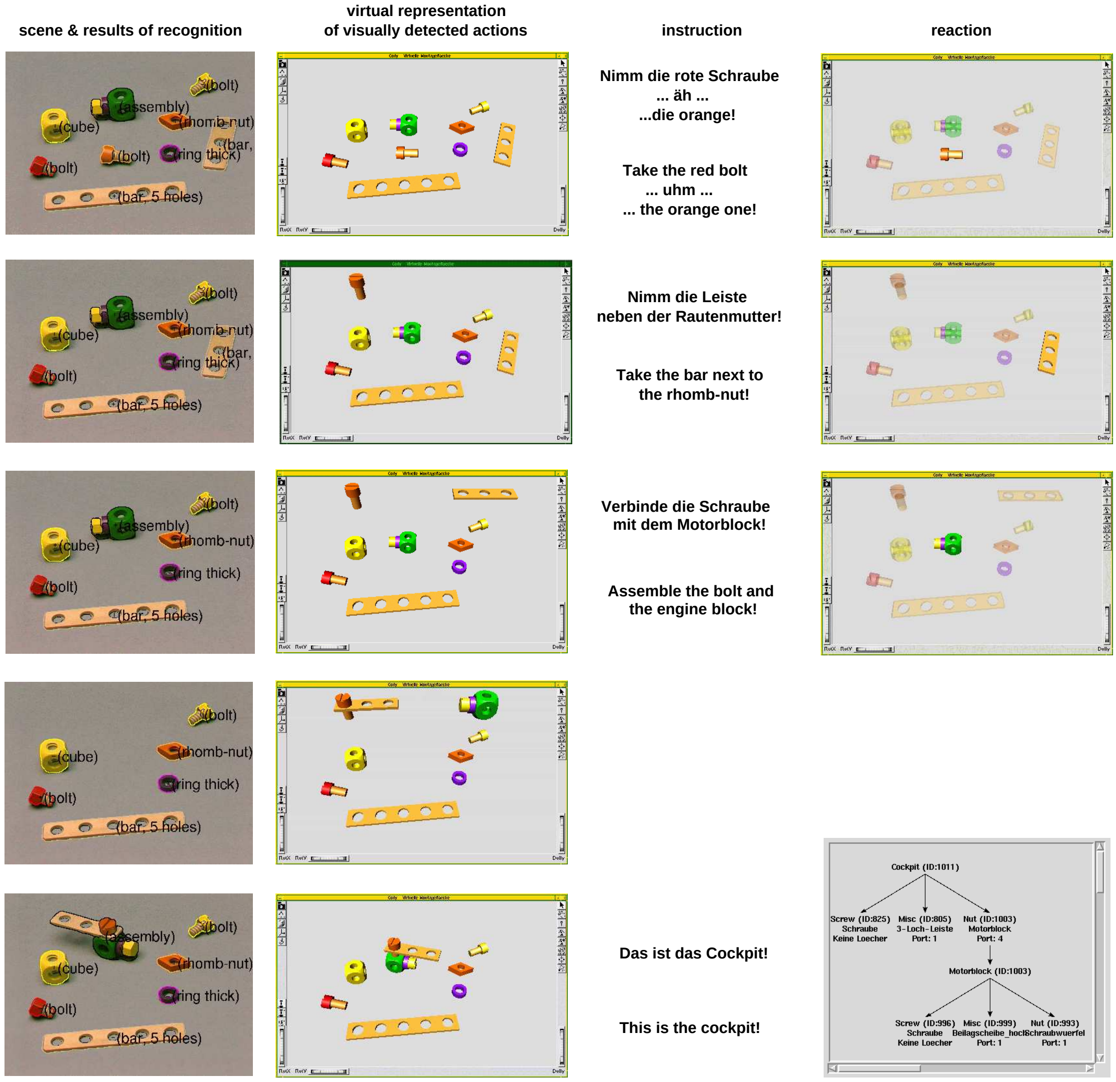

Figure 5: Visualization of the course of an experiment demonstrating integrated speech and image understanding. First column: the observed scene together with results of the visual object recognition procedure. Second column: a virtual representation of the internal state of the system. (see [21] for details of the visualization tool). Third column: spoken German instruction together with English translation. Fourth column: the system state with the object intended by the user instruction highlighted.

shown in the first image of the second row is obtained. In order to indicate that the system has noticed the disappearance of the bolt and believes it to be grasped by a robotic manipulator the bolt is virtually lifted up. In this instruction the bar is referenced using a spatial relation.

3) After the bar was removed from the scene its virtual counterpart is lifted up assuming it is grasped by the second manipulator of a two-handed robot. In this instruction no prior knowledge of *"engine block"* is modeled in the system except for the acoustic model for word recognition and the syntactic category. But since there is a complex object in the scene which could have an abstract meaning, the system concludes that the user is referring to this assembly. Due to this implicit name designation the structural description of the assembly in the database is labeled *"engine block"*.

4) Then, taking the course of the construction into account and using its knowledge of the dialog history, the system infers that the bar in the one hand of the manipulator has to be connected with the bolt in the other hand to have a free hand for grasping the engine block. The systems conclusion is illustrated in the second image of the fourth row where the bolt and the bar are virtually connected and the assembly is lifted up. Note that the depicted mating relation of the bar and the bolt do not necessarily correspond to reality since the action detection has no visual observation of the action details.

5) In the final step of this experiment all three objects are assembled and the resulting assembly is put back into the scene. Again the virtual representation is generated based on the results from action detection which lacks visual observations, therefore the 3D position and orientation of the assembly elements is not correct. Using the fused assembly structure the system rebuilds the assembly virtually. Finally, the explicit naming causes

the system to label the new assembly structure shown in the lower-most image in the fourth column as *"cockpit"*.

This example demonstrates the possibility to instruct our system in a way that resembles interpersonal conversation. Generally, a user can refer to objects using their visual features or their relative position. Corrections may be used, natural pronunciation and hesitations do not disturb the speech understanding processes. If an instruction can not be disambiguated automatically the system is able to ask for further information or more detailed instructions. Thus, intuitive operating of the system is possible; if necessary one has to reformulate an instruction to achieve the desired action. This, however, can also be done in a conversational manner and does not require special experience.

We have measured the system performance with 5 cooperative users which were told to construct an assembly similar to the "cockpit" from Fig. 5. For each construction step an average time of 45 seconds was used with approximately 1.4 interactions to specify the complete instruction.

7 Conclusion

We presented an artificial communicator that is designed to behave cooperatively in order to support flexible assembly. We demonstrated the 'look and feel' of our system with an example to show that our system is capable of dealing with many problems occurring in man-machine-interaction. Additionally, we provided preliminary results about the overall system performance given cooperative users.

The individual modules described in this paper were tested separately and their performance has been evaluated (see [7, 12, 24, 14, 15]). However, there is no widely accepted method to asses the performance of an integrated system interacting with a human. Some authors argue that experiments demonstrating complex systems running are sufficient to verify the quality of the engaged techniques [22]. This point of view corresponds to what is known as *end-to-end evaluation* [8]. Viewing at system performance in this way allows for individual recognition results to be incorrect while the complete system is capable of resolving such situations by e.g. a dialog with the instructor. By integrating different perceptive modalities, learning of name designations, and applying flexible dialog strategies our system is capable to reach a user's intended target. Thus, it meets the requirements we expect from cooperative man-machine interaction.

References

[1] *Image and Vision Computing*, volume 16, no. 4. 1998. Special Issue on Vision-Based Aids for the Disabled.

[2] R. Arkin. *Behaviour based robotics*. MIT Press, Cambridge, MA, 1998.

[3] D. H. Ballard and C. Brown. *Computer Vision*. Prentice Hall, New York, 1982.

[4] C. Bauckhage, J. Fritsch, F. Kummert, and G. Sagerer. Towards a Vision System for Supervising Assembly Processes. In *Proc. Symposium on Intelligent Robotic Systems (SIRS'99)*, pages 89–98, 1999.

[5] C. Bauckhage, F. Kummert, and G. Sagerer. Learning Assembly Sequence Plans Using Functional Models. In *Proc. IEEE Int. Symposium on Assembly and Task Planning (ISATP'99)*, pages 1–7, 1999.

[6] J.M. Bradshaw, editor. *Software Agents*. AAAI Press/MIT Press, Menlo Park, CA, 1997.

[7] H. Brandt-Pook, G. A. Fink, S. Wachsmuth, and G. Sagerer. Integrated recognition and interpretation of speech for a construction task domain. In *Proc. Int. Conf. on Human-Computer Interaction*, volume 1, pages 550–554, 1999.

[8] T. Bub and J. Schwinn. VERBMOBIL: The Evolution of a Complex Large Speech – to – Speech Translation System. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP'96)*, volume 4, pages 2371 – 2374, 1996.

[9] E. den Os, L. Boves, L. Lamel, and P. Baggia. Overview of the ARISE project. In *Proc. European Conf. on Speech Technology*, pages 1527–1530, 1999.

[10] R. Engelmore and T. Morgan, editors. *Blackboard Systems*. Addison-Wesley, Reading, MA, 1988.

[11] O. Faugeras. *Three-Dimensional Computer Vision*. MIT Press, Cambridge, MA, 1993.

[12] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin, 1999. Springer.

[13] G. A. Fink, N. Jungclaus, H. Ritter, and G. Sagerer. A communication framework for heterogeneous distributed pattern analysis. In *Int. Conf. on Algorithms And Architectures for Parallel Processing*, pages 881–890, 1995.

[14] S. Kronenberg and F. Kummert. Soft unification: Towards robust parsing of spontanous speech. In *IASTED Int. Conf. on Artificial Intelligence and Soft Computing*, pages 381–385, 1999.

[15] F. Kummert, G.A. Fink, G. Sagerer, and E. Braun. Hybrid Object Recognition in Image Sequences. In *Proc. Int. Conf. on Pattern Recognition (ICPR'98)*, volume II, pages 1165–1170, 1998.

[16] P. McKevitt, editor. *Integration of natural language and vision processing*. Kluver Academic Press, Dordrecht, 1994.

[17] V. Mittal, H. Yanco, J. Aronis, and R. Simpson, editors. *Assistive Technology and Artificial Intelligence: Applications in Robotics, User Interfaces and Natural Language Processing*, volume 1458 of *Lecture Notes in Computer Science*. Springer Verlag, New York, 1998.

[18] J. Miura and K. Ikeuchi. Task Planning of Assembly of Flexible Objects and Vision-Based Verification. *Robotica*, 16:297–307, 1998.

[19] A.I. Rudnicky, E. Thayer, C. Constantinides, et al. Creating natural dialogs in the Carnegie Mellon communication system. In *Proc. European Conf. on Speech Technology*, pages 1531–1534, 1999.

[20] G. Sagerer and H. Niemann. *Semantic Networks for Understanding Scenes*. Plenum Publishing Corporation, New York, 1997.

[21] G. Sagerer and S. Wachsmuth, editors. *Proc. of ICCV Workshop: Integration of Speech and Image Understanding*, Corfu, Greece, September 1999. IEEE Computer Society.

[22] C.-P. Tung and A.C. Kak. Integrating Sensing, Task Planning and Execution for Robotic Assembly. *IEEE Transactions on Robotics and Automation*, 12(2):187–201, 1996.

[23] S. Wachsmuth, H. Brandt-Pook, G. Socher, F. Kummert, and G. Sagerer. Multilevel integration of vision and speech understanding using bayesian networks. In H. I. Christensen, editor, *Computer Vision Systems: First Int. Conf.*, volume 1542 of *Lecture Notes in Computer Science*, pages 231–254. Springer-Verlag, 1999.

[24] S. Wachsmuth, G. A. Fink, and G. Sagerer. Integration of parsing and incremental speech recognition. In *Proceedings of the European Signal Processing Conference*, volume 1, pages 371–375, Rhodes, 1998.