Selective Visual Perception Driven by Cues from Speech Processing

Reinhard Moratz, Hans-Jürgen Eikmeyer, Bernd Hildebrandt,
Alois Knoll, Franz Kummert, Gert Rickheit, Gerhard Sagerer

SFB 360, Universität Bielefeld
Postfach 100131, 33501 Bielefeld
email: reinhard@techfak.uni-bielefeld.de

Keywords: Computer Vision, Speech Processing, Cognitive Modelling, Distributed AI

1 Introduction

At the University of Bielefeld within the framework of the colloborative research center "Situated Artificial Communicators" (SFB 360) an integrative system of vision and speech understanding is developed. By the term "artificial communicators" we mean formal systems which reconstruct the behaviour of natural communicators in relevant aspects. Since most cognitive skills are situation-dependent, the SFB's research has concentrated on a specific basis scenario. The subject of this scenario is a task-orientated discourse about construction acts. The cooperative assembly of a model aeroplane from construction-kit parts (see Figure 1) is the long term goal.

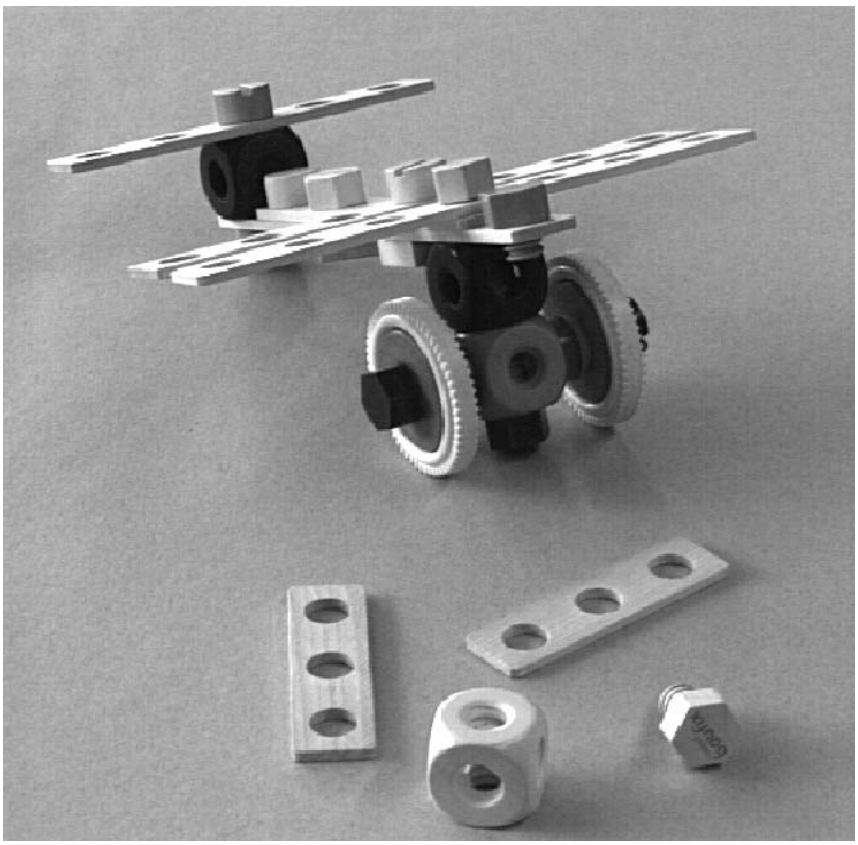

Fig. 1. The model airplane

The system operates on the basis of semantic networks and facilitates the symbolic interpretation of sensory data whilst providing a uniform representation for visual and linguistic knowledge by means of which an interaction between modules is simplified.

As the formalism for knowledge representation the semantic network formalism ERNEST has been used [NSSK90, KNPS93]. Besides the general characteristics of semantic networks such as the representation of concepts and the relationships between them, the formalism also has a variety of additional features which are highly interesting from a cognitive perspective [HMRS95].

2 Knowledge representation in a semantic network

The main influence on the development of ERNEST was the theory of semantic networks as described for example by Sowa [Sow84, Sow91]. Important elements of a semantic network are concepts, their attributes and the relations between concepts. These are usually represented as nodes, their internal structures and links between nodes. The main task of the formalism is to provide the symbolic interpretion of sensory data; in this context these data are primarily visual and acoustic signals. In ERNEST there are three types of nodes:

- A *concept* can represent a class of objects, events or abstract conceptions.
- An *instance* is understood as the concrete realisation of a concept in the sensory data; i.e. an instance is the copy of a concept by which the general description is replaced by concrete values.
- In addition, there are also *modified concepts*. A modified concept represents knowledge which is adapted to a concrete situation of analysis.

Features of a concept, such as the size of an object or the syntactical gender of a noun phrase, can be represented by means of attributes. In this way, concepts in ERNEST are given an internal structure. Since the attributes of a concept are sometimes dependent on each other, ERNEST also makes it possible to represent relationships between attributes. In ERNEST, there are the following link types:

- Through the link type *part*, two concepts are connected with each other if one concept is understood as a part of the other concept.
- Another well-known link type is the *specialisation*, with a related inheritance mechanism by which a special concept inherits all properties of the general one.
- The link type *concretisation* connects two concepts to each other if a concept is represented on different levels of abstraction. Thus the visual perception of an ellipse may be a hole or a tyre on a higher level of abstraction.

The creation of modified concepts and instances constitutes the knowledge utilization in the semantic network. For the creation of instances, this process is based on the fact that the recognition of a complex object has the detection of all its parts as a prerequisite. For concepts which model terms only defined within a certain context the instantiation process must proceed in the opposite direction. In this case the context must exist before an instance of the context-dependent concept can be created. In the network language, these ideas are expressed by problem-independent inference rules. Context-independent parts, contexts, and concretes are the prerequisites for the creation of instances and modified concepts in a data-driven strategy. The opposite link directions are used for model driven inferences. Since the results of an initial segmentation are hypotheses, the definition of a concept is completed by a judgement function estimating the degree of correspondence of a part of the signal to the term defined by the related concept. On the basis of these estimates and the inference rules an A*-like control algorithm is applied.

3 Architecture of the situated artificial communicator

3.1 Speech understanding

The speech-processing component in ERNEST is based on a speech-understanding dialogue system for railway information [MKE+94]. The system's purpose is to automatically understand spontaneous spoken language and to answer the questions put. A speech recognition system delivers word hypotheses at the interface with the linguistic knowledge base.

The structuring of the knowledge base is orientated towards Winograd's cognitive speech-processing model [Win83], which advocates stratified processing. In this model, the essential processing levels are respectively a syntactic, a semantic and a pragmatic level, all based on a uniform representation.

Since the order of syntactic constituents in spoken German language is relatively free, no attempt was made to model a complete sentence grammar. The semantic level follows Fillmore's deep case theory [Fil68], according to which syntactic-semantic roles are associated with verbs. The processing strategy in the dialogue system alternates between data-driven and model-driven processes. This facilitates the efficient use of relevant information from both the acoustic data and the linguistic knowledge base.

Figure 2 shows the individual levels of the integrative architecture's speech understanding component, illustrated by examples of some concepts. The first letter of a concept name indicates the level to which it belongs: P_CUBE is a concept on the pragmatic level, S_OBJECT is a concept on the semantic level, and SY_NOUN is a concept on the syntactic level. Since the speech recognition system can now deliver the full forms of words as hypotheses on the hypothesis level H, there is a so-called word level W, which includes the collective concepts for individual word forms. This structuring follows psycholinguistic criteria.

3.2 Image understanding

Model-driven processes are particularly important for cognitively motivated vision processing where an interaction with speech is being aimed at. The latter demands the consideration of special purposes which are derived from the system's situated global requirements [Bro92].

For this reason, the exclusion of model-driven influences, as often practised in more traditional approaches (e.g. [Mar82]), is no longer appropriate for a comprehensive architecture. In our approach, objects are modelled by means of individual entities which can be detected robustly and which specify redundantly the object. In this, lighting conditions and perspective are taken into consideration on the perceptive level. The decomposition approach has a cognitive foundation found in the work of Biedermann [Bie87]. In Figure 2, these mechanisms correspond to level *PE* of the knowledge base. An object corresponds to only a few percepts, since only a small number of topologically differing views can be derived from the contour structure of an object (see [RM94, CE94]). For example, a rhomb-nut is modelled by its upper side, the hole and the two visible sides in front, in a particular spatial arrangement. In ERNEST, a spatial arrangement of this kind can be represented by a relation between attributes within concepts. A three-dimensional reconstruction of the scene is created on subsequent processing levels. In Figure 2, this level is shown by concretisation level *R*.

The interface with the segmentation processes on the signal level (level *I*) is given by regions of homogeneous colour and their model-based approximation. The example image shows a typical scene and a detail containing a rhomb nut which is a typical construction element of our szenario (see Figure 3). Figure 4 shows the results of the segmentation and approximation steps. These processes result in an affine invariant detection of elements of a small set of image primitives (for example ellipses, parallelograms).

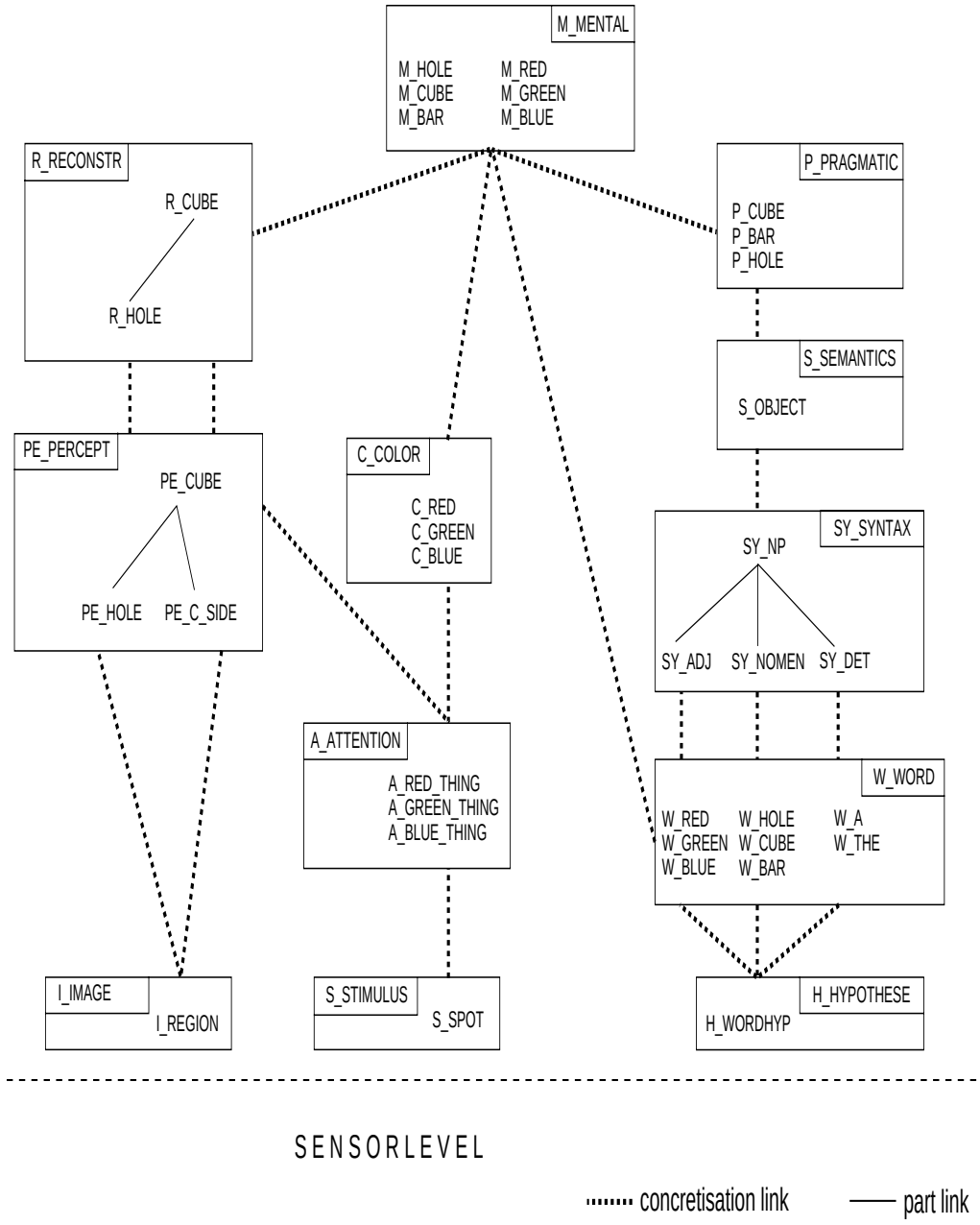

Fig. 2. integrative architecture (simplified)

In selective perception, spatially-oriented attention with low resolution (level S) provides an important initial indication for subsequent focussing mechanisms (level A). Object-related attention is realised in the PE module that has been described above. This architecture allows an interaction of low resolution and colour with subsequent focusing and shape processing. As such, it represents an abstraction and coarsening of biologically motivated architectures which model saccadic eye movements [vSBE+94].

3.3 The conception of the robotics component

It is obviously desirable to control robot manipulators directly by natural language or even speech input. Given the elaborate representation models and integration modules described above, one of the main objectives of future research is the control of a bi-arm robot system modelled on the geometry of the human arms.

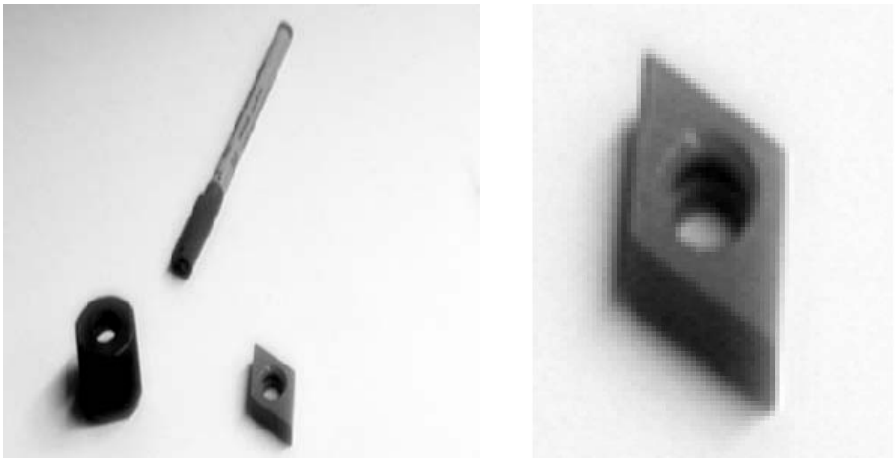

Fig. 3. An example 512 × 512 image and detail

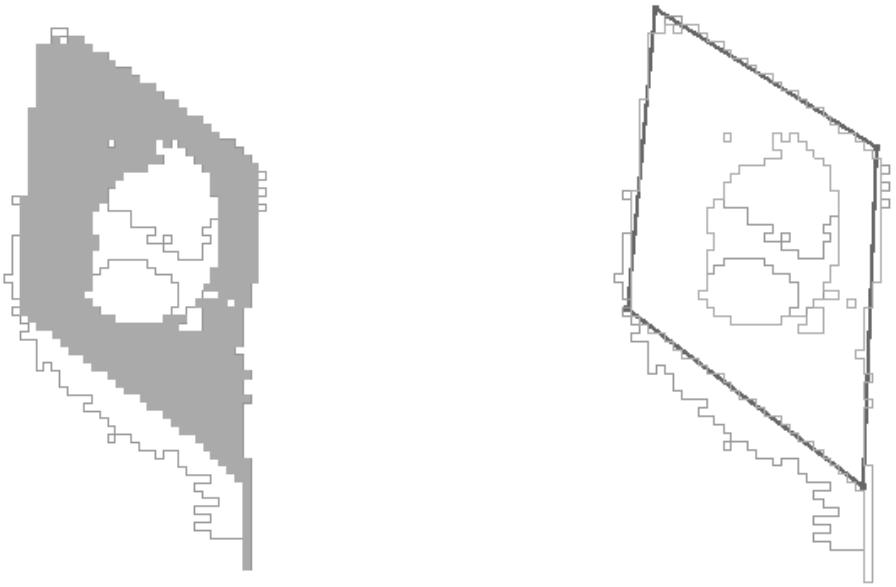

Fig. 4. Segmentation and approximation results

The general goal of our approach is to realise human sensorimotor skills for performing real-world manipulation and assembly tasks, i.e. to carry out a complex assembly task rather than to simulate it. This requires a comprehensive set of actuators and sensors, which perform functions similar to those of the human arms, hands and perception channels (i.e. vision, touch, acoustics). To organise the interaction of the complex sensor and control subsystems, sensor data cannot be acquired and processed independently of the movements of the actuators. It is mandatory that both must be performed simultaneously and in view of the task the actuators will work on. Therefore major components of an intelligent robot system are considered synergetically instead of separately. It is only this kind of synergy that makes the realisation of complex, cooperative behaviours involving a large number of sensors and actuators feasible.

The aim of our work is to fully automate the process of multi-sensor supported assembly by gradually enabling the robot/sensor system to carry out the individual steps in a more and more autonomous fashion. A fully automatic assembly, however, presupposes a precise task description and a sophisticated error handling.

3.4 The integration of the components

The integration of the modalities speech and vision is based on Johnson-Laird's theory of mental models [JL83]. Important in this regard is the integrative and coherent representation of objects and facts as well as the cognitive processes based upon such a representation. By means of direct access to various aspects of a concept, it is possible to provide adequate modelling of temporal sequences of cognitive processes. For instance, on the evidence of current psycholinguistic experiments, it seems likely that word recognition may have a direct influence on saccadic eye movements, i.e. on visual processing, even when a word which is heard has not yet been processed through all linguistic levels [SKSET94].

The integration of the individual components of the knowledge base takes place on a common level of abstraction, which we take as a represention of mental models (level M). In Figure 2, the entire conceptual hierarchy can be seen. A significant characteristic of this hierarchy is that concretisation relations exist not only between adjacent processing levels, but that there are also direct connections between the mental models level and conceptual levels near the signal. By this means, the modelling of an immediate interaction between the visual and speech components should be possible.

In addition to the intended influence which the instructor's directions have on the conscious actions of the constructor, unconscious processes are also set in motion. The implication for our modelling is that visual processes already begin during the incremental processing of the verbal instructions. For example, the visual search for red objects in the scene can begin as soon as the word *red* from the first instruction "now you take the red cube" has been understood, before a complete linguistic interpretation of the entire utterance has been completed.

The assembly of the model-aircraft implies numerous requirements for sensor and actuator cooperation. It can be used as an ideal test-bed for investigating the interaction between human communicators and machine systems in the real-world. Furthermore it will be used for validating the complete integration concept by subsuming different linguistic and cognitive components: A number of separate parts must first be recognised, then they must be manipulated and finally built together to construct the aircraft. Within the framework of the SFB, in each of these steps, a human communicator instructs the robot (which implies that the interaction between them plays an important role in the whole process). Due to many degrees of freedom in our model, more detailed empirical data are however still required for a more differentiated modelling.

4 Conclusion

The integrative architecture presented here is based on the requirements arising from the situated integration of speech, vision and robotics. This integration requires a homogeneous representation for image and speech understanding. Accordingly, a close interaction of early processes provides the vision module with cues for colour focusing and selective shape processing. We are adapting our present system for a distributed workstation environment based on message passing [FJRS95].

The architecture which has been briefly outlined here offers great potential for further research. Psycholinguistic experiments are needed however, since the process of speech and vision interaction is at present underspecified from the empirical point of view.

Acknowledgement

This work has been supported by the German Research Foundation (DFG) in the SFB 360 project.

References

[Bie87] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[Bro92] C.M. Brown. Issues in Selectiv Perception. In *11th International Conference on Pattern Recognition*, volume I, pages 21–30, The Hague, 1992.

[CE94] F. Cutzu and S. Edelman. Canonical views in object representation and recognition. *Vision Research*, 34:3037–3056, 1994.

[Fil68] C. Fillmore. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguisitic Theory*, pages 1–90. Holt, Rinehardt, and Winston, New York, 1968.

[FJRS95] G.A. Fink, N. Jungclaus, H. Ritter, and G. Sagerer. A Communication Framework for Heterogeneous Distributed Pattern Analysis. In *International Conference on Algorithms And Architectures for Parallel Processing*, pages 881–890, Brisbane, 1995.

[HMRS95] B. Hildebrandt, R. Moratz, G. Rickheit, and G. Sagerer. Integration von Bild- und Sprachverstehen in einer kognitiven Architektur. In *Kognitionswissenschaft*, volume 4, pages 118–128, Berlin, 1995. Springer-Verlag.

[JL83] P. Johnson-Laird. *Mental Models.* Cambridge University Press, Cambridge, 1983.

[KNPS93] F. Kummert, H. Niemann, R. Prechtel, and G. Sagerer. Control and Explanation in a Signal Understanding Environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145, 1993.

[Mar82] D. Marr. *Vision.* Freeman, San Francisco, 1982.

[MKE+94] M. Mast, F. Kummert, U. Ehrlich, G. Fink, T. Kuhn, H. Niemann, and G. Sagerer. A Speech Understanding and Dialog System with a Homogeneous Linguistic Knowledge Base. *IEEE Trans. on Patt. Analysis and Machine Intelligence*, 16(2):179–194, 1994.

[NSSK90] H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(9):883–905, 1990.

[RM94] H. Rom and G. Medioni. Part Decomposition and Description of 3D Shapes. In *1994 ARPA Image Understanding Workshop*, volume II, November 1994.

[SKSET94] M. Spivey-Knowlton, J. Sedivy, K. Eberhard, and M. Tanenhaus. Psycholinguistic study of the interaction between language and vision. In *AAAI-94 Workshop on Integration of Natural Language and Vision Processing*, pages 189–192, Seattle, WA:, 1994. American Association for Artificial Intelligence.

[Sow84] J.F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Reading, MA, 1984.

[Sow91] J.F. Sowa, editor. *Principles of Semantic Networks.* Morgan Kaufmann, San Mateo, Calif., 1991.

[vSBE+94] W. von Seelen, S. Bohrer, C. Engels, W. Gillner, H. Janßen, H. Neven, G. Schöner, W.M. Theimer, and B. Völpel. Visual information processing in neural architecture. *Proceedings 16. DAGM-Symposium*, pages 36–57, 1994.

[Win83] T. Winograd. *Language as a Cognitive Process. Vol.I: Syntax.* Addison Wesley, Reading Mass., 1983.

This article was processed using the LaTeX macro package with TTP style