

On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER

Elisabeth André, Gerd Herzog, Thomas Rist

SFB 314, Project VITRA, Universität des Saarlandes
D-66041 Saarbrücken, Germany

Abstract

The aim of previous attempts at connecting vision systems and natural language systems has been to provide a retrospective description of the analysed image sequence. The step from such an *a posteriori* approach towards simultaneous natural language description reveals a problem which has not yet been dealt with in generation systems. Automatic generation of simultaneous descriptions calls for the application of an incremental event recognition strategy and for the adequate coordination of event recognition and language production. In order to enable free interaction between these processes, it is useful to implement them in parallel. In this paper¹ the system *Soccer* will be presented, which is based upon such a conception. Short sections of soccer games have been chosen as the domain of discourse. In analogy to radio reports, the system generates a description of the game which it is watching and which the listener cannot see.

This paper appeared in: Proc. of the 8th ECAI, pp. 449–454, Munich, 1988.

¹The work described here was partly supported by the SFB 314 Research Program of the German Science Foundation (DFG) on AI and Knowledge-Based Systems, project N2: VISual TRANslator.

1 Introduction

Although image understanding and natural language understanding constitute two major areas of AI, they have been studied rather independently of each other. Only a few works are concerned with the integration of computer vision and the generation of natural language expressions for the description of image sequences.

In his pioneering work, Badler dealt with the interpretation of object motions in terms of natural-language-oriented concepts such as '*swing*' and '*bounce*' as well as associated directional adverbials (c.f. Badler [1975]). His approach has been improved by Tsotsos, who proposed a largely domain-independent hierarchy of conceptual motion frames (cf. Tsotsos et al. [1980]) which is specialized further within the system *Alven* (Tsotsos [1985]) to treat left ventricular heartwall motion. Both Badler and Tsotsos based their motion concepts on categories developed in Miller [1972], a linguistic study on English motion verbs. Although Tsotsos is not concerned with natural language processing, most concepts of motion recognized by the *Alven* system correspond to certain natural-language notions of change.

A more linguistic approach is due to Okada. Based on his study of almost 5000 Japanese verbs, Okada developed a set of 20 semantic features which is used within the system *Supp* to select, out of the 1200 verb concepts implemented, those concepts which are applicable to a sequence of simple line drawings Okada [1979]. His primitives do not form the basis for a taxonomy but constitute features for pattern recognition.

Traffic scenes have been studied in various systems. For instance, they constitute one of the domains of the dialog system *Ham-Ans* (Hoepfner et al. [1983]), an access system to very diverse back-end systems. Based on a procedural reference semantics for certain verbs of locomotion, questions concerning the motions of vehicles and pedestrians at an intersection can be answered (cf. Wahlster et al. [1983]). Besides question-answering, the system *Naos* (Neumann [1984]) allows for free verbalization of recognized events (cf. Novak [1987]). In *Naos*, event recognition is based on a hierarchy of event models, i.e. declarative descriptions of classes of events organized around verbs of locomotion. Another system for event recognition called *Epex*, which has many similarities to *Naos* but tries to use *KL-ONE*²-like representation schemes, is proposed in Walter [1987].

Summing up, it may be said that the aim of these different approaches has been to provide a retrospective description of the analyzed image sequence. A totally new task results if events are to be recognized simultaneously as they occur in the scene. The following question has to be asked thereby: How can partly recognized events be represented in order to make them available for further processing? With respect to the generation of simultaneous scene descriptions in natural language, this problem becomes obvious: If the description is to be focused on what is currently happening, it is very often necessary to verbalize events even while they are currently happening

²cf. Brachman and Schmolze [1985]

and not yet completed. On the one hand, this task leads to specific requirements for the modelling of events. Simultaneous recognition of events can only be carried out by means of an incremental recognition strategy in which events are recognized stepwise as they progress and in which event instances are explicitly represented in the knowledge base of the system right from their first detection. On the other hand, automatic generation of simultaneous descriptions for image sequences has consequences for the planning and realization of natural language utterances (cf. André et al. [1987b]). As a scene is not described *a posteriori*, but instead simultaneously as it progresses, the complete course of the scene is unknown at the moment of text generation. Thus, planning is restricted to a limited section of the scene. Since the description should concentrate on what is currently happening, it is necessary to start talking about events even while they are still progressing and not yet completely recognized. In this case encoding has to start before the contents of an utterance have been planned in full detail.

In this light, the aim of our work was to conceive and to implement a system which, in analogy to simultaneous reporting, should analyse a successively perceived image sequence and describe the immediately recognized occurrences in natural language. Since radio reports of soccer games are a linguistically well-studied example of simultaneous descriptions (cf. Rosenbaum [1969] and Brandt [1983]), short sections of video recordings of soccer games have been chosen as the domain of discourse. We are primarily interested in those aspects of soccer reports which are relevant for simultaneous descriptions in general. Our goal is the generation of objective descriptions mainly in written, but also in spoken, German. Phenomena known from real radio reports such as metaphor, exaggeration etc. are neglected. The conversational setting can be described as follows: The system generates a simultaneous report of the game it is watching for a listener who cannot see the game her/himself, but who is assumed to have prototypical knowledge about the static background.

2 The System SOCCER

A geometrical description of the scene, initially represented by an image sequence, forms the input for the system *Soccer*. The stationary part of the so-called *Geometrical Scene Description* (cf. Neumann [1984]), an instantiated model of the static background, is fed into the system manually. The trajectories of the dynamic objects are to be supplied by a vision system (cf. Sung and Zimmermann [1986]) which is under development at the Fraunhofer-Institute for Information and Data Processing (IITB) in Karlsruhe. The segmentation and cueing of moving objects is done by computing and analyzing displacement vector fields. First results obtained in the intersection domain have been used within our system *Citytour* (André et al. [1986], André et al. [1987a]) to answer questions about simple events (cf. Schirra et al. [1987]). Sequences of up to 1000 images (40 seconds), recorded with a stationary TV-camera during a game of the German professional soccer league, are currently being investigated. In this domain,

segmentation becomes more difficult because the moving objects cannot be regarded as solid bodies and occlusions occur more frequently. The algorithms developed so far are currently being improved for the new domain. At present, the trajectory data can still only be generated manually by means of a special trajectory editor (Herzog [1986]).

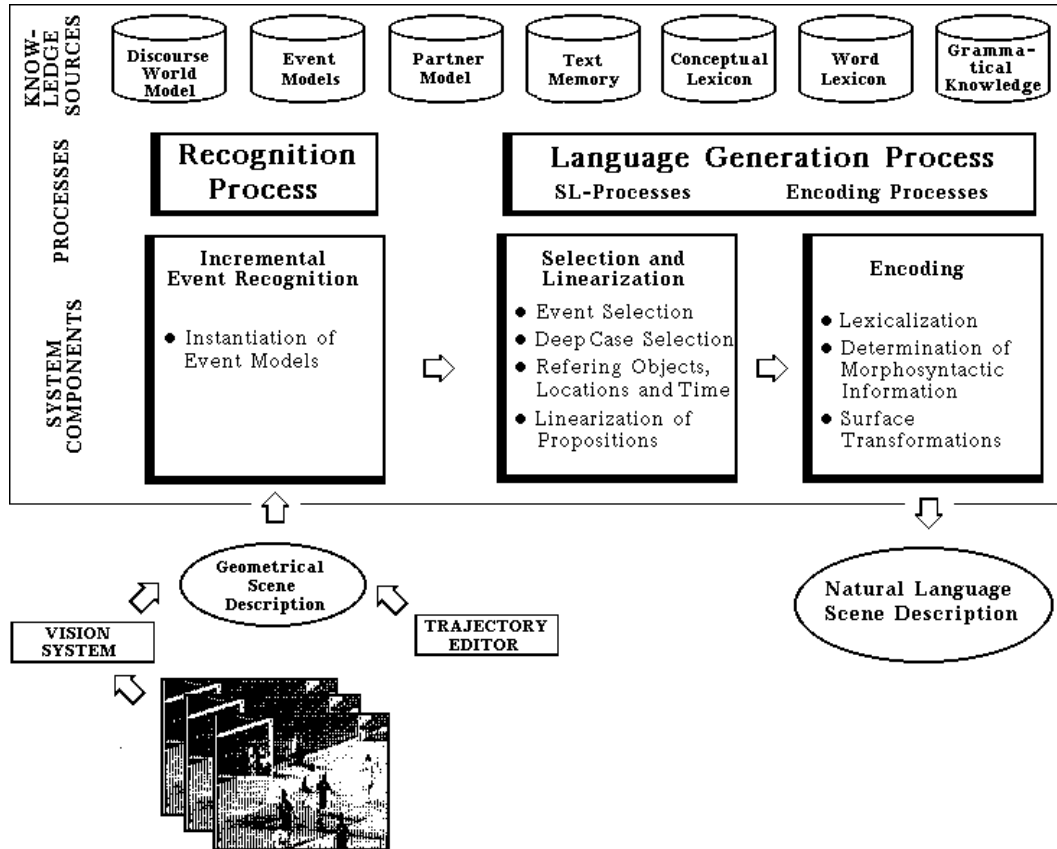


Figure 1: The architecture of the *Soccer* system

Fig. 1 gives an overview of the system: Based on the successively incoming geometrical data for the dynamic objects, the incremental event recognition mechanism provides information in propositional form concerning events occurring at the moment. The selection/linearization component selects relevant propositions, orders them and passes them on to the encoding component. Encoding processes transform non-verbal information into an ordered sequence of words. The output is either written or spoken German. For the latter case a speech synthesis module is used. As the scene is to be described continuously, temporal aspects such as the duration of speech generation, or the time the reader needs for decoding in the case of a written text, have to be considered for the coordination of event recognition and language generation. In order to guarantee that current events can immediately influence language generation,

event recognition has to continue during the course of language generation. In the *Soccer* system, the multitasking capability of the Lisp-machine is utilized to implement event recognition and event selection in parallel with the encoding processes.

3 Incremental Event Recognition

The major goal of the *Soccer* system is the description of events as they are occurring in the scene. For this purpose, we have developed a formalism for the representation of events as well as a mechanism for incremental event recognition.

3.1 Event Models

In analogy to object models, events are described conceptually using so-called *event models* (cf. Neumann [1984]). Event models represent *a priori* knowledge about typical occurrences in a scene. The recognition of an event occurring in a particular scene then corresponds to an instantiation of the respective generic event model.

Besides a specification of *roles* denoting participating objects, the core of an event model is its *course diagram*. It specifies the sub-concepts and situational context which characterize the instances of the particular event model. The recognition of an occurrence can be thought of as traversing the course diagram. We have defined course diagrams as labeled directed graphs (cf. Rist et al. [1987]). Each edge is characterized by a tuple (*source*, *goal*, *condition*, *type*). *source* and *goal* are nodes in the graph and determine the direction of the edge. If the condition of an edge (*condition*) can be satisfied at a given instant, then the edge can be traversed from *source* to *goal*. Finally, each edge has a type (*type*) in order to define special predicates for the description of events occurring at the moment.³ These predicates are:

- $TRIGGER(t_i \text{ event})$ for the beginning,
- $PROCEED(t_i \text{ event})$ for the progression and
- $STOP(t_i \text{ event})$ for the termination of an event.

For durative events, for example '*bewegen*' (to move), a further predicate has been introduced to express that an event is still going on:

- $SUCCEED(t_i \text{ event})$.

One of these predicates is satisfied if an edge of the respective type in the event's course diagram is traversed at the specified instant. By means of these basic predicates, further predicates can be defined: the predicate *ACTIVE* for example, which is the disjunction of *TRIGGER*, *PROCEED* and *SUCCEED*.

³Following Allen [1984] and McDermott [1982], events occur as objects within a temporal logic.

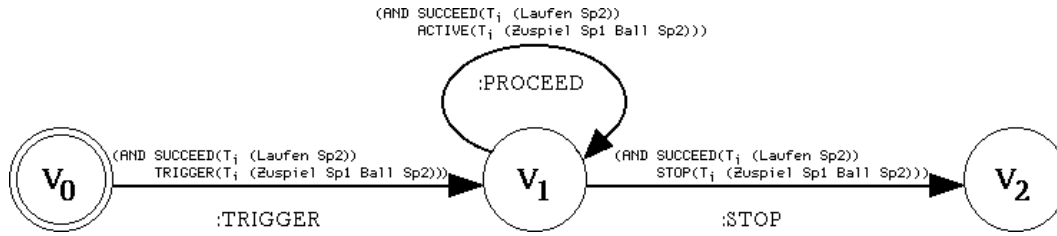


Figure 2: Course diagram

Fig. 2 shows the course diagram of the concept 'Pass in den Lauf' (running pass). It describes a situation in which a player passes the ball to another member of his team, who must be running at the same time. The event is triggered if a pass is triggered and the receiving player is running. The event proceeds as long as its sub-events are active and stops when the pass is completed. Using course diagrams guarantees:

- *Uniformity*
Primitive motion concepts as well as complex activities can be defined by means of course diagrams.
- *Declarative event descriptions*
Knowledge concerning the occurrence of an event is represented using a declarative formalism.
- *Applicability of an incremental recognition process*
Course diagrams can be traversed incrementally by traversing exactly one edge per unit of time.

3.2 Instantiation of Event Models

The task of event recognition can be seen as the instantiation of event models based on the scene data. As soon as new input data are provided by the vision system, the recognition component continues traversing course diagrams already activated and tries to trigger new ones. For each event model, there is one special instance, the so-called *demon event*, which waits for its *TRIGGER*-condition to become true. When this is the case, the demon event becomes an activated event instance and a new demon will automatically be generated. All event instances of an event model are managed by a so-called *event handler*. Fig. 3 shows the handlers for the event models 'Laufen' (run) and 'Pass in den Lauf' (running pass) with some event instances as well as the demon events.

Event recognition is based on a data-driven bottom-up-strategy. First, the traversal of course diagrams corresponding to basic events is attempted. More complex event instances then communicate with the handlers of their sub-events. To continue traversing the course diagram for the event (*Pass in den Lauf Sp#4 Ball#1 Sp#7*) in fig. 3,

the handler for the concept 'Laufen' is asked if it knows an instance (*Laufen Sp#7*) such that the predicate *ACTIVE* is satisfied at time T_{now} .

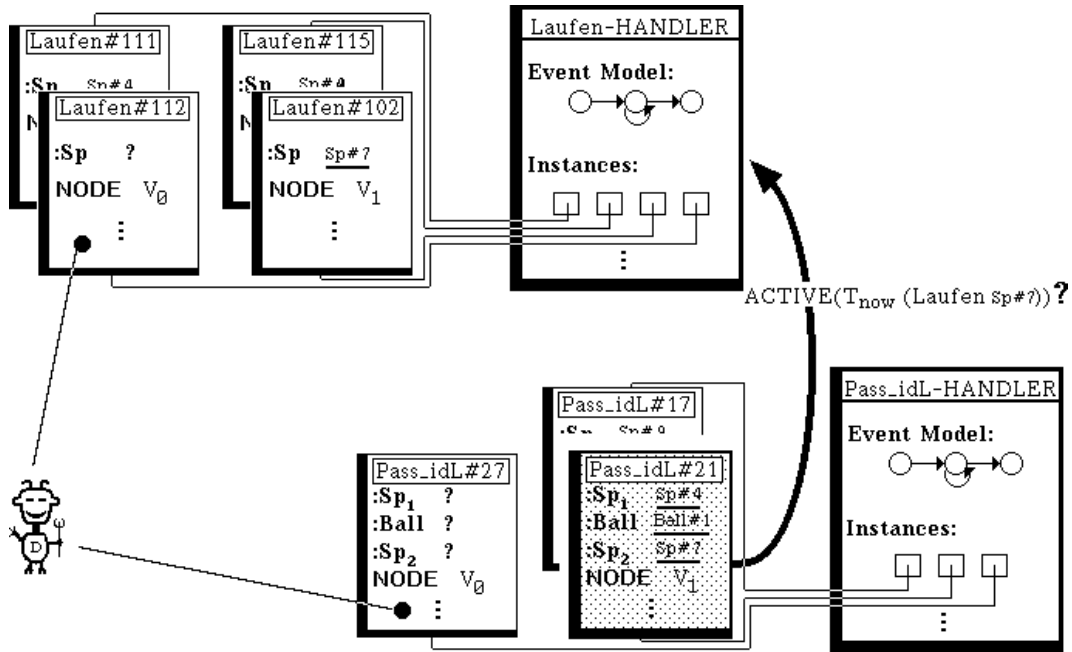


Figure 3: Management of event instances in *Soccer*

4 Language Generation

Information about recognized and partly recognized events is stored in the event proposition buffer and updated continuously as the scene progresses. The contents of this buffer form the input data for the language generation component. In *Soccer*, language generation includes processes that handle the selection, linearization and verbalization of event propositions.

4.1 Selection and Linearization of Events

As the system has to describe a scene continuously, language generation underlies strong temporal restrictions. Hence, the system cannot talk about all events which have been recognized. At each point in time the system has to decide which events should be verbalized in order to enable the listener to follow the scene. According to the conversational maxims of Grice (cf. Grice [1975]), the listener should be informed about all relevant events and redundancy should be avoided. In *Soccer*, the relevance of an event depends on its state, its salience and its topicality. Events with state *SUCCEED* or *STOP* are preferred to events with state *TRIGGER* or *PROCEED*. To determine

the salience of an event, factors such as its frequency of occurrence or the complexity of its generic event model are taken into account. Whereas values denoting the salience of events are stored *a priori* in the database, the topicality of events which have already occurred decreases as the scene progresses, i.e. relevance continually changes. To avoid redundancy, the occurrence of an event will not be mentioned if it is implied by the occurrence of some other event already verbalized.⁴

If more than one event proposition has been selected, the system has to determine the order in which the propositions should be mentioned in the text. The linearization process first considers the temporal ordering of the corresponding events, i.e. if an event *A* ends before an event *B* starts then *A* should be verbalized before *B*. Furthermore, focusing criteria are used to maintain discourse coherency. A preliminary text plan is constructed according to these criteria. The need for changing this text plan arises when an outstanding event (e.g. a goal kick) occurs which has to be verbalized as soon as possible, or when the topicality of events already selected decreases to such an extent that they no longer need to be verbalized.

4.2 Verbalization of Event Propositions

The transformation of symbolic event descriptions into natural language utterances is performed by the encoding component. A verb is selected by accessing the concept lexicon, which links non-linguistic with linguistic concepts, and the case-roles associated with the verb are instantiated. Control passes back to the selection component in order to decide which information concerning the case-role fillers should be conveyed. The selected information is transformed into natural-language expressions referring to time, space or objects. Time is referred to by the verb tense and by temporal adverbs such as '*währenddessen*' (in the meantime) or '*jetzt*' (now). In order to refer to space, spatial prepositions and appropriate objects of reference, for example '*am Mittelkreis*' (at the center circle), are selected. The computational semantics for the spatial prepositions is based on the proposals in André et al. [1986] and André et al. [1987a]. For referring to objects, their internal identifiers (e.g. *player#1*) are transformed into nominal phrases. To this purpose, the system selects attributes enabling the listener to uniquely identify the intended referent whereby it must access the partner model and the text memory. If an object cannot be characterized by attributes stored *a priori* in the partner model, it will be described by means of spatial relations, for example '*der linke Elfmeterpunkt*' (the left penalty spot), or by means of events already mentioned in which it was (is) involved, for example '*der Spieler, der angegriffen wurde*' (the player who was attacked). In order to increase text coherency, anaphoric expressions are generated if the referent is in focus and confusion can be excluded.

To meet the requirements of simultaneous scene description, the event recognition component also provides information concerning only partly-recognized events. The

⁴In SOCCER, as currently implemented, these implication relations are stored in the database. It is planned, however, for this information to be computed from the event models.

consequence is that language generation cannot start from completely worked-out conceptual contents; i.e. the need for an incremental generation strategy arises. Currently, an experimental version for the generation of surface structures is used which utilizes morphological processes of the system *Sutra* (Busemann [1983]).

5 The User Interface

Utilization of the window system and menu techniques facilitate system handling. In order to allow for the improvement of the approach realized in *Soccer*, various trace features as well as the possibility of inspection of the different knowledge sources are supplied which increase the perspicuity of the system. By means of the highly interactive trajectory editor, it is possible to synthesize animated displays of various formations and tactical situations. Furthermore, trajectories can be superimposed onto the static background and single preprocessed digitised images can be shown on the screen (cf. Herzog [1986]). Additional graphic features are available in the color version of the *Soccer* system. For instance, parts of trajectories corresponding to certain events can be highlighted using different colors.

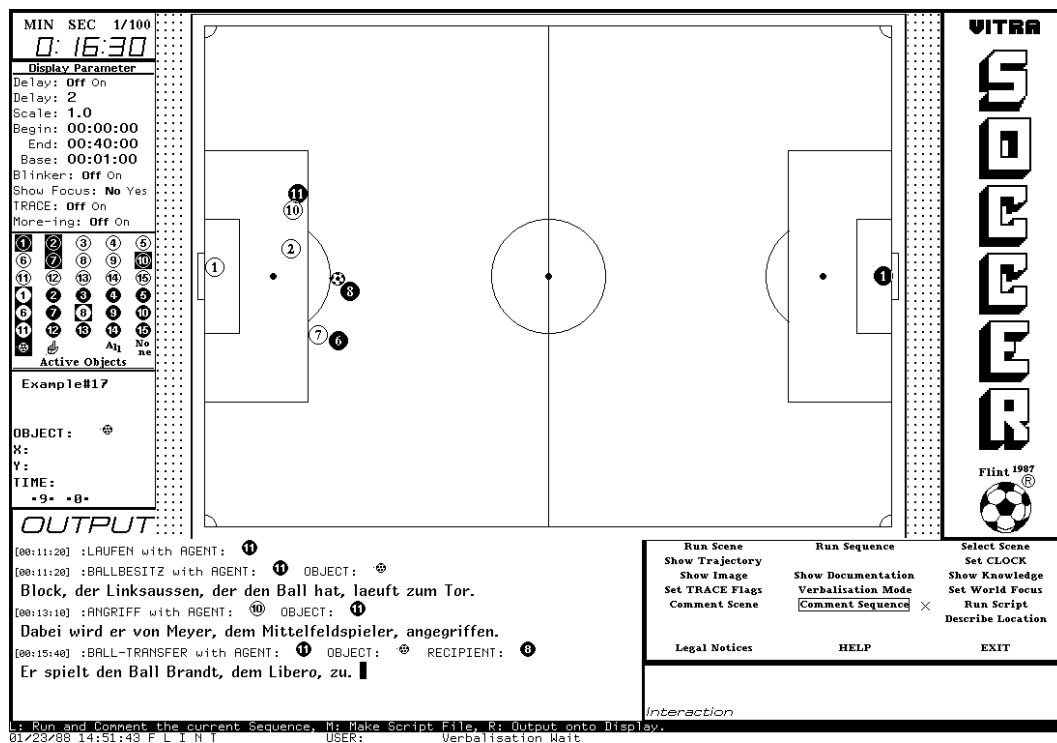


Figure 4: The basic windows of the *Soccer* system

The performance of the *Soccer* system is demonstrated by displaying short scenes within the graphics window and describing them simultaneously in natural language.

The scene is shown from an idealized birds-eye view; the dynamic objects are represented as icons. Fig. 4 shows part of a typical description as an example: At time [00:11:20] the language generation component has selected that *player#11* of the team in the dark shirts has the ball and that he is running towards the goal. While encoding this information, the system recognizes that *player#10* of the opposing team is attacking him. The temporal relation between these two events is expressed by the temporal adverb '*dabei*' (meanwhile). Because of the high degree of focus of *player#11* the passive voice has been chosen. At time [00:15:40] the event has been selected that *player#11* is passing the ball to *player#8* of his own team. The German description in the output window might be translated as follows: '*Block, the outer left, who has the ball, is running towards the goal. Meanwhile he is attacked by Meyer, the midfielder. He passes the ball to Brandt, the sweeper!*'

6 Future Work

Our system *Soccer* should be regarded as a framework which allows the examination of typical effects which occur in producing simultaneous scene descriptions. In order to improve the quality of text produced, our future work will concentrate on the following issues:

- So far, the role-fillers of events are restricted to single objects. The next step will be the recognition of coordinated motions of groups of objects (e.g. an attack of a team). As there are millions of possible sets of objects, heuristics for detecting interesting situations have to be formulated.
- Apart from visual motion concepts, non-visual concepts such as intentions and plans are also to be used for the selection of an adequate natural language description of observed movements; i.e. the same trajectory will be mapped onto completely different motion descriptions, depending on the intention assumed of the actor.
- Another focus of research is the development of a pictorial partner model representing the listener's temporal and spatial conceptualization of the scene. A match between the actual scene observed by the system and the imagined scene allows for checking whether an utterance will have the intended effect on the listener's imagination.

7 Technical Notes

The system *Soccer* has been developed on Symbolics 3600 and 3640 Lisp-machines running Release 6.1. One of the machines is equipped with an OP36-C108 Color Option, which includes an additional high-resolution color screen. A special version of the *Soccer* system which utilizes the color screen is available.

Speech production is done by means of an *AEG-SVS* speech synthesis module connected to the Symbolics via a serial port. The speech synthesis module expects ASCII characters as input and generates spoken German.⁵

The system is implemented in a strictly modular object-oriented style, using the Flavor System embedded in Zetalisp. The program consists of almost 100 flavors and about 1000 Lisp functions, including Flavor methods. The complete source code requires about 750 Kbytes in the file system plus 1 Mbyte online documentation. An incremental world save of the loaded system occupies about 7000 blocks, i.e. 7.5 Mbyte disk space. The program is currently being ported to Commonlisp and the New Flavor System under Release 7.1.

References

- J. F. Allen.** Towards a General Theory of Action and Time. *Artificial Intelligence*, **23**(2), 123–154, 1984.
- E. André, G. Bosch, G. Herzog, T. Rist.** Characterizing Trajectories of Moving Objects Using Natural Language Path Descriptions. In: *Proc. of the 7th ECAI*, vol. 2, pp. 1–8, Brighton, UK, 1986.
- E. André, G. Bosch, G. Herzog, T. Rist.** Coping with the Intrinsic and the Deictic Uses of Spatial Prepositions. In: K. Jorrand, L. Sgurev, eds., *Artificial Intelligence II: Methodology, Systems, Applications*, pp. 375–382, North-Holland, Amsterdam, 1987a.
- E. André, T. Rist, G. Herzog.** Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung zeitveränderlicher Szenen. In: K. Morik, ed., *GWAI-87. 11th German Workshop on Artificial Intelligence*, pp. 330–337, Springer, Berlin, Heidelberg, 1987b.
- N. I. Badler.** Temporal Scene Analysis: Conceptual Description of Object Movements. Technical Report 80, Computer Science Department, Univ. of Toronto, 1975.
- R. J. Brachman, J. G. Schmolze.** An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, **9**(2), 171–216, 1985.
- W. Brandt.** *Zeitstruktur und Tempusgebrauch in Fußballreportagen des Hörfunks*. Elwert, Marburg, 1983.
- S. Busemann.** Oberflächentransformationen bei der Generierung geschriebener Deutscher Sprache. In: B. Neumann, ed., *GWAI-83. 7th German Workshop on Artificial Intelligence*, pp. 90–99, Springer, Berlin, Heidelberg, 1983.

⁵We would like to thank the AEG research center, Ulm, for giving us one gadget free of charge.

- H. P. Grice.** Logic and Conversation. In: P. Cole, J. L. Morgan, eds., *Speech Acts*, pp. 41–58, Academic Press, London, 1975.
- G. Herzog.** Ein Werkzeug zur Visualisierung und Generierung von geometrischen Bildfolgenbeschreibungen. Memo 12, Universität des Saarlandes, SFB 314 (VI-TRA), Saarbrücken, 1986.
- W. Hoepfner, T. Christaller, H. Marburger, K. Morik, M. O'Leary, W. Wahlster.** Beyond Domain-Independence: Experience with the Development of a German Language Access System to Highly Diverse Background Systems. In: *Proc. of the 8th IJCAI*, pp. 588–594, Karlsruhe, FRG, 1983.
- D. McDermott.** A Temporal Logic for Reasoning about Processes and Plans. *Cognitive Science*, **6**, 101–155, 1982.
- G. A. Miller.** English Verbs of Motion: A Case Study in Lexical Memory. In: A. W. Melton, E. Martin, eds., *Coding Processes in Human Memory*, pp. 335–372, Winston, Washington, DC, 1972.
- B. Neumann.** Natural Language Description of Time-Varying Scenes. Report 105, Fachbereich Informatik, Univ. Hamburg, 1984.
- H.-J. Novak.** Strategies of Generating Coherent Descriptions of Object Movements in Street Scenes. In: G. Kempen, ed., *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pp. 117–132, Nijhoff, Dordrecht, Boston, Lancaster, 1987.
- N. Okada.** SUPP: Understanding Moving Picture Patterns Based on Linguistic Knowledge. In: *Proc. of the 6th IJCAI*, pp. 690–692, Tokio, Japan, 1979.
- T. Rist, G. Herzog, E. André.** Ereignismodellierung zur inkrementellen High-level Bildfolgenanalyse. In: E. Buchberger, J. Retti, eds., *3. Österreichische Artificial-Intelligence-Tagung*, pp. 1–11, Springer, Berlin, Heidelberg, 1987.
- D. Rosenbaum.** *Die Sprache der Fußballreportage im Hörfunk*. Ph.D. thesis, Fachbereich Germanistik, Univ. des Saarlandes, 1969.
- J. R. J. Schirra, G. Bosch, C.-K. Sung, G. Zimmermann.** From Image Sequences to Natural Language: A First Step Towards Automatic Perception and Description of Motions. *Applied Artificial Intelligence*, **1**, 287–305, 1987.
- C.-K. Sung, G. Zimmermann.** Detektion und Verfolgung mehrerer Objekte in Bildfolgen. In: G. Hartmann, ed., *Mustererkennung 1986; 8. DAGM-Symposium*, pp. 181–184, Springer, Berlin, Heidelberg, 1986.

- J. K. Tsotsos.** Knowledge Organization and its Role in Representation and Interpretation for Time-Varying Data: the ALVEN System. *Computational Intelligence*, **1**, 16–32, 1985.
- J. K. Tsotsos, J. Mylopoulos, H. D. Covvey, S. W. Zucker.** A Framework for Visual Motion Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 563–573, 1980.
- W. Wahlster, H. Marburger, A. Jameson, S. Busemann.** Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. In: *Proc. of the 8th IJCAI*, pp. 643–646, Karlsruhe, FRG, 1983.
- I. Walter.** EPEX: Bildfolgendeutung auf Episodenebene. In: K. Morik, ed., *GWAI-87. 11th German Workshop on Artificial Intelligence*, pp. 21–30, Springer, Berlin, Heidelberg, 1987.