In: Ch. Freksa und W. Brauer (Hrsg.), Wissensbasierte Systeme , pp. 153-162, Berlin, Heidelberg: Springer.

Incremental Natural Language Description of Dynamic Imagery

Gerd Herzog,* Chen-Ko Sung,+ Elisabeth André, **
Wilfried Enkelmann,+ Hans-Hellmut Nagel,+ ++ Thomas Rist, **
Wolfgang Wahlster,* ** Georg Zimmermann+

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

* German Research Center for Artificial Intelligence (DFKI)
D-66123 Saarbrücken, Germany

+ Fraunhofer–Institut für Informations- und Datenverarbeitung (IITB)
D-76131 Karlsruhe, Germany

++ Fakultät für Informatik, Universität Karlsruhe (TH)
D-76128 Karlsruhe, Germany

Abstract

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

The aim of our joint efforts at combining a vision system and a natural language access system is the automatic *simultaneous* description of dynamic imagery, i.e., we are interested in image interpretation and language processing on an *incremental* basis. In this contribution[1] we sketch an approach towards the integration of the Karlsruhe vision system called *Actions* and the natural language component *Vitra* developed in Saarbrücken. The steps toward realization, based

[1] The work described here was partly supported by the Sonderforschungsbereich 314 der Deutschen Forschungsgemeinschaft, “Künstliche Intelligenz und wissensbasierte Systeme”, projects V1 (IITB, Karlsruhe) and N2: VITRA (Universität des Saarlandes).

on available components, are outlined and the capabilities of the current system are demonstrated.

Zusammenfassung

Obwohl das Bildverstehen und die Verarbeitung natürlicher Sprache zwei der Kerngebiete im Bereich der KI darstellen, wurden sie bisher nahezu unabhängig voneinander untersucht. Nur sehr wenige Ansätze haben sich mit der Intergration von maschinellem Sehen und der Generierung natürlichsprachlicher Äußerungen zur Beschreibung von Bildfolgen beschäftigt.

Das Ziel unserer Zusammenarbeit bei der Kopplung eines bildverstehenden Systems und eines natürlichsprachlichen Zugangssystems ist die automatische *simultane* Beschreibung zeitveränderlicher Szenen, d.h. wir sind interessiert an Bildfolgeninterpretation und Sprachverarbeitung auf inkrementeller Basis. In diesem Beitrag beschreiben wir einen Ansatz zur Integration des Karlsruher Bildfolgenanalysesystems *Actions* und der natürlichsprachlichen Komponente *Vitra*, die in Saarbrücken entwickelt wird. Die Schritte hin zur Realisierung, basierend auf bereits verfügbaren Komponenten, werden dargestellt und die Fähigkeiten des derzeit vorhandenen Systems demonstriert.

This paper appeared in: In: C. Freksa and W. Brauer (eds.), Wissensbasierte Systeme. 3. Int. GI-Kongreß, pp. 153–162. Berlin, Heidelberg: Springer, 1989.

1 Introduction

Image understanding and natural language processing are two major areas of research within AI that have generally been studied independently of one another. Advances in both technical fields during the last 10 years form a promising basis for the design and construction of integrated knowledge-based systems capable of translating visual information into natural language descriptions. From the point of view of cognitive science, anchoring meaning in a referential semantics is of theoretical as well as practical interest. From the engineering perspective, the systems envisaged here could serve such practical purposes as handling the vast amount of visual data accumulating, for example, in medical technology, remote sensing, and traffic control.

The goal of our joint efforts at combining a vision system and a natural language access system is the automatic *simultaneous* description of dynamic imagery, i.e., we are interested in image interpretation and language processing on an *incremental* basis. The conversational setting is this: the system provides a running report of the scene it is watching for a listener who cannot see the scene her/himself, but who is assumed to have prior knowledge about its static properties. In this paper we describe the integration of the Karlsruhe vision system *Actions* and the natural language component *Vitra* developed in Saarbrücken.[2] The steps toward realization, based on available components, are outlined, and results already obtained in the investigation of traffic scenes and short sequences from soccer matches will be discussed.

2 Relations to Previous Research

Following Kanade (see Kanade [1980]), it is advantageous for a discussion of machine vision to distinguish between the 2-D picture domain and the 3-D scene domain. So far, most machine vision approaches have been concerned (i) with the detection and localization of significant grey value variations (corners, edges, regions) in the picture domain, and in the scene domain (ii) with the estimation of 3-D shape descriptions, as well as—more recently—(iii) with the evaluation of image sequences for object tracking and automatic navigation. Among the latter approaches, the estimation of relative motion between camera(s) and scene components as well as the estimation of spatial structures, i.e., surfaces and objects, are focal points of activity (see Ayache and Faugeras [1987], Faugeras [1988], Nagel [1988b]). Few research results have been published about attempts to associate picture domain cues extracted from image sequences with conceptual descriptions that could be linked directly to efforts at algorithmic processing of natural language expressions and sentences. In this context, computer-based generic descriptions for complex movements become important. Those accessible in the image understanding literature have been surveyed in Nagel [1988a]. Two even more recent investigations in this direction have been published

[2] The acronyms stand for `Automatic Cueing and Trajectory estimation in Imagery of Objects in Natural Scenes' and `VIsual TRAnslator'.

in Witkin et al. [1988] (in particular Section D) and Goddard [1988]. A few selected approaches from the literature are outlined in the remainder of this section to provide a background for the ideas presented here.

In Badler [1975], Badler studied the interpretation of simulated image sequences with object motions in terms of natural language oriented concepts. His approach has been improved by Tsotsos, who proposed a largely domain-independent hierarchy of conceptual motion frames which is specialized further within the system *Alven* to analyze X-ray image sequences showing left ventricular wall motion (see Tsotsos [1985]). Later, a similar system for the analysis of scintigraphic image sequences of the human heart was developed by Niemann et al. (see Niemann et al. [1985]). Based on a study of Japanese verbs, Okada developed a set of 20 semantic features to be used within the system *Supp* to match those verb patterns, that are applicable to simple line drawings (see Okada [1979]). Traffic scenes constitute one of the diverse domains of the dialog system *Ham-Ans* (see Wahlster et al. [1983]). Based on a procedural referential semantics for certain verbs of locomotion, the system answers questions concerning the motions of vehicles and pedestrians. The system *Naos* (see Neumann [1984], Novak [1986]) also allows for a retrospective natural language description. 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. The more recent *Epex* system (see Walter et al. [1988]) studies the handling of conceptual units of higher semantic complexity, but still in an *a posteriori* way.

The natural language interfaces mentioned so far have not been connected to real vision components, they use only simulated data. Apart from our previous results (see André et al. [1986], Schirra et al. [1987]) the *LandScan* system (see Bajcsy et al. [1985]) constitutes the only approach in which processing spans the entire distance between raw images and natural language utterances but it deals only with static scenes.

3 Simultaneous Evaluation and Natural Language Description of Image Sequences

The main goal of our cooperation is the design and implementation of an integrated system that performs a kind of simultaneous reporting, that is, evaluating an image sequence and immediately generating a natural language description of the salient activities corresponding to the most recent image subsequence. It is not (yet) real-time evaluation, but our approach emphasizes concurrency of image sequence evaluation and natural language generation.

In order to gain a realistic insight into the problems associated with such an endeavor, we decided to evaluate real-world image sequences with multiple mobile agents or objects, based on system components which are already partially available due to previous research efforts in the laboratories involved. Since the analysis of complex articulated movements still exceeds our capabilities given the computational resources available today, we concentrate initially on the picture domain in order to detect and track projected object candidates, which are considered to be essentially rigid. The crucial links between the picture domain results and the natural language processing steps are provided by *complex events*, i.e., higher conceptual units capturing the spatio-temporal aspects of object motions. A *complex event* should be understood as an `event' in its broadest sense, comprising also notions like `episode' and `history' (see Nagel [1988a]). The recognition of intentions and plans (see Retz-Schmidt [1988]) is, however, outside the scope of this paper. In what follows, the term `event' will be used to refer to *complex events*.

3.1 Overall Structure of the Approach

The task of generating natural language descriptions based on visual data can roughly be subdivided into three parts: (1) constructing an abstract propositional description of the scene, the so-called *Geometrical Scene Description* (GSD, see Neumann [1984]), (2) further interpretation of this intermediate geometrical representation by recognizing *complex events*, and (3) selection and verbalization of appropriate propositions derived in step 2 to describe the scene under discussion. Because of the simultaneity of the description in our case, the three steps have to be carried out incrementally.

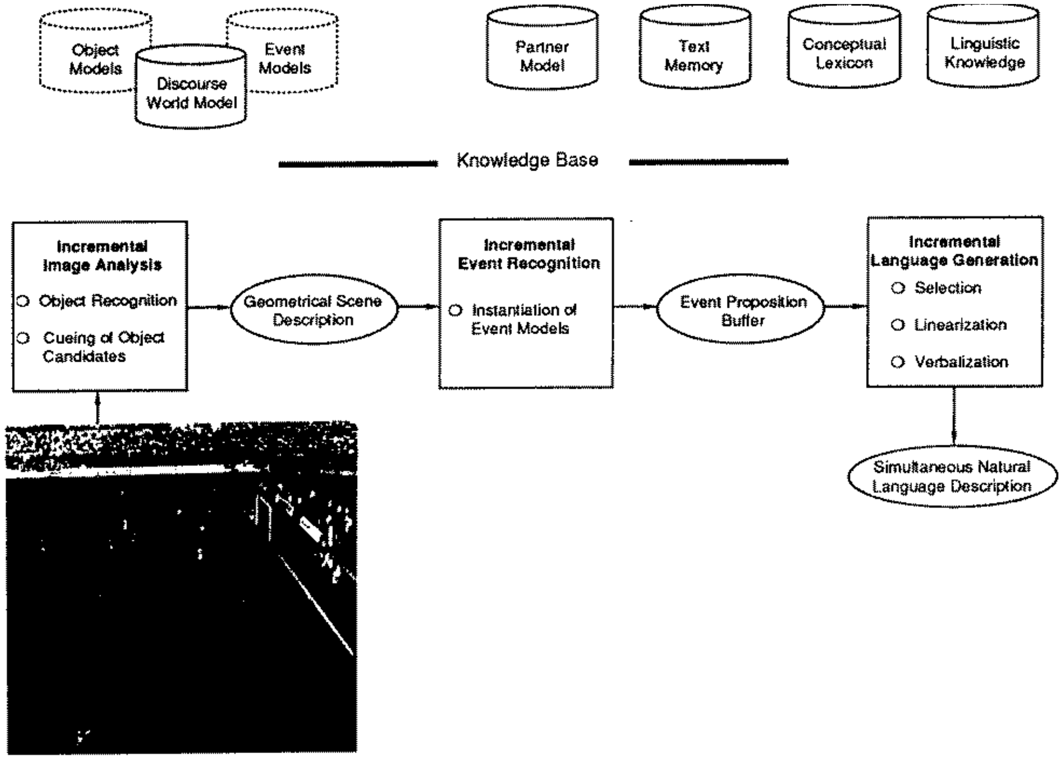

Object Models
Event Models
Discourse World Model
Partner Model
Text Memory
Conceptual Lexicon
Linguistic Knowledge
Knowledge Base
Incremental Image Analysis
○ Object Recognition
○ Cueing of Object Candidates
Geometrical Scene Description
Incremental Event Recognition
○ Instantiation of Event Models
Event Proposition Buffer
Incremental Language Generation
○ Selection
○ Linearization
○ Verbalization
Simultaneous Natural Language Description

Figure 1: The architecture of the integrated system

Fig. 1 gives an overview of the architecture of our integrated system. An image sequence, i.e., a sequence of digitized video frames, forms the input for the system. Based on these incoming visual raw data, the image analysis component constructs a geometrical representation of the scene, stating the locations of the visible objects at consecutive points in time. The contents of the GSD, which is constructed incrementally, as new visual data arrive, are further interpreted by the event recognition component. Information about recognized and partly recognized events is stored in the event proposition buffer and updated continuously as the scene progresses. The language generation component selects relevant propositions from this buffer, orders them and finally transforms the non-verbal information into an ordered sequence of either written or spoken German words.

In order to guarantee that current events can immediately influence language generation, image analysis and event recognition must continue during the course of language generation. Thus, the different processes need to be implemented at least partly in parallel.

3.2 Incremental Picture Domain Analysis of Image Sequences

Image sequences of thousands of frames have to be analysed in order to provide input for the generation of non-trivial natural language descriptions. In order to limit the required computations, a very robust method for the estimation of displacement vector fields (see Kories and Zimmermann [1986]) has been applied in the *Actions* system. The digitized TV-frame is subjected to a bandpass filter. Blobs representing local maxima and minima are then determined as features and tracked through subsequent frames, resulting in displacement vectors. The method has been successfully applied to several ten thousands of images taken from various sources without any change of parameters. Its first steps are now implemented in a VLSI-chip working at video rate.

The displacement vectors are clustered in order to incrementally create `candidate' moving objects in the picture domain. Tracking such object candidates through extended image subsequences allows us to incrementally build up (projected) trajectories which—together with additional attributes like size, speed, orientation, and internal blob structure of object candidates—provide the input data for the natural language generation steps (see Sung and Zimmermann [1986], Schirra et al. [1987]).

3.3 Incremental Event Recognition

In order to be able to talk about incomplete events while they are happening, we have to recognize them `stepwise', as they progress; event instances must be made available for further processing from the moment they are first noticed. The different approaches mentioned in Section 2, as well as formalisms like Allen's temporal logic (see Allen [1984]), only distinguish between events that have occurred and those that have not, thus our specific requirements for the modeling of events are not met. To clarify this, consider the short image sequence shown in Fig. 2.

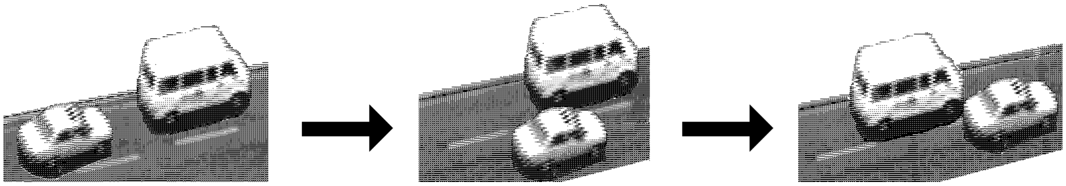


Figure 2: A `passing' event in a traffic scene

It is only after the final image that all necessary subevents have been detected and the `passing' event can be recognized. For an incremental description, however, various phases of an event must be distinguished. After the first image one could already say that a `passing' event seems to be starting, i.e., the `passing' event is *triggered* by the `swing-out' event. In the second image there is a `drive-beside' event and the `passing' event *proceeds*. Finally the `passing' event *stops* because the expected `swing-into-line' event is recognized.

In the *Vitra* system the additional event predicates *trigger*, *proceed*, and *stop* can be used to characterize the progression of such a `passing' event with greater precision. In order to model durative events like `move', a further predicate called *succeed* was introduced to express the continuation of an event. By means of an incremental recognition strategy based on these predicates, events can be recognized simultaneously as they occur in the scene and additional information concerning partly-recognized events can be provided.

3.4 Incremental Natural Language Generation

Automatic generation of simultaneous descriptions for dynamic imagery reveals a problem that has not heretofore been dealt with within generation systems. On the one hand, temporal aspects such as the time required for text generation and decoding time of the listener or reader have to be considered for the coordination of perception and language production. On the other hand, automatic generation of simultaneous descriptions has consequences for the planning and realization of natural language utterances. Since a scene is not described *a posteriori* but rather as it progresses, the entire scene itself would only be known after it is complete. 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 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. Other characteristics of simultaneous reporting besides incremental generation of utterances need to be dealt with. The description often lags behind with respect to the events in the scene and unexpected topic shifts occur very frequently.

Language generation in *Vitra* includes processes that handle the selection, linearization and verbalization of event propositions (see André et al. [1987]). After relevant propositions are selected and ordered, they are passed on to the encoding process. Additional selection processes are used to determine deep cases and to choose descriptions for objects, locations, and time; in these choices the contents of the text memory and the partner model must also be considered. Encoding includes lexicalization, the determination of morphosyntactic information, and surface transformations. Lexicalization is based on the conceptual lexicon, which constitutes the connection between non-linguistic and linguistic concepts.

4 Details of Realization

Using an example from the soccer domain, the present section elaborates in more detail the various steps in transforming the results of the picture domain analysis of an image sequence into a simultaneous natural language description.

4.1 Event Models and Event Recognition

Events are described conceptually by means of event models. In addition to a specification of *roles* denoting participating objects, which must be members of specified object classes, an event model includes a *course diagram*, used to model the prototypical progression of an event. We have defined course diagrams as labeled directed graphs with typed edges (see André et al. [1988]). Fig. 3 shows a simplified course diagram for the concept *`ball-transfer'*. It describes a situation in which a player passes the ball to a teammate. The event is *triggered* if a `have-ball' event is stopped and the ball is free. The event *proceeds* as long as the ball is moving free and *stops* when the recipient has gained possession of the ball. The recognition of an occurrence can be thought of as traversing the course diagram, where the edge types (*:trigger*, *:proceed*, etc.) are used for the definition of our basic event predicates (see Section 3.3). Fig. 4 shows how an interval-based representation of an event can easily be translated into a course diagram.

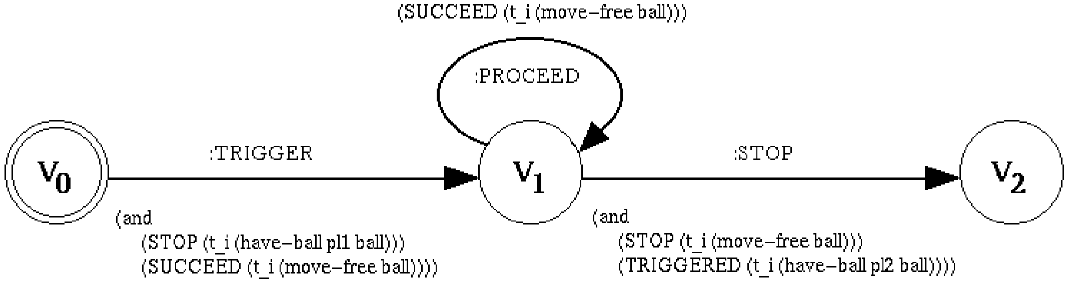

(SUCCEED (t_i (move–free ball)))
:PROCEED
:TRIGGER
V0
V1
:STOP
V2
(and
(STOP (t_i (have–ball pl1 ball)))
(SUCCEED (t_i (move–free ball))))
(and
(STOP (t_i (move–free ball)))
(TRIGGERED (t_i (have–ball pl2 ball))))

Figure 3: Course diagram for `ball-transfer'

Using course diagrams guarantees that primitive motion concepts as well as complex activities can be defined in an uniform and declarative way. Course diagrams

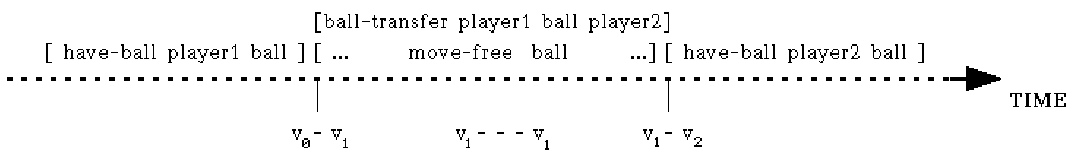

[ball-transfer player1 ball player2]
[have-ball player1 ball] [... move-free ball ...] [have-ball player2 ball]
TIME
$v_0 - v_1$
$v_1 - - - v_1$
$v_1 - v_2$

Figure 4: Corresponding interval-based representation

allow for incremental event recognition, since exactly one edge per unit of time is traversed. As soon as new input data are provided by the vision system, the recognition component continues traversing the course diagrams that are already activated and tries to *trigger* new ones. In order to allow for a uniform data-driven recognition strategy, each class of events has at all times one additional instance, the *demon event*, whose task is to wait for the *trigger*-condition to become true so that a new instance of the event can be created. Each recognition cycle starts at the lowest level of the event hierarchy: first, the traversal of course diagrams corresponding to basic events is attempted; later, more complex event instances can look at those lower levels to verify the existence of their necessary subevents.

4.2 Selection and Linearization of Propositions

Because of the strong temporal restrictions the system cannot talk about all recognized events, thus it 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 (see Grice [1975]), the listener should be informed about all relevant events and redundancy should be avoided. The relevance of an event depends on factors like: (i) salience, which is determined by the frequency of occurrence and the complexity of the generic event model, (ii) topicality, and (iii) current state, i.e., events with state *succeed* or *stop* are preferred. As the scene progresses topicality decreases for stopped events and events enter different states, thus relevance changes continually. To avoid redundancy, an event will not be mentioned if it is implied by some other event already verbalized, e.g., a `have-ball' event following a pass will not be selected for verbalization.

The linearization process determines the order in which the selected propositions should be mentioned in the text. The temporal ordering of the corresponding events is the primary consideration for linearization; secondarily, focusing criteria are used to maintain discourse coherence. The need to change this preliminary text plan arises when an outstanding event (e.g., a goal kick) occurs, or because the topicality of events already selected has fallen below a certain threshold.

4.3 Verbalization of Event Propositions

In the process of transforming symbolic event descriptions into natural language utterances, first a verb is selected by accessing the concept lexicon, and the case-roles associated with the verb are instantiated. Control passes back to the selection component, which decides 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 indicated by the verb tense and by temporal adverbs; spatial prepositions and appropriate objects of reference are selected to refer to spatial relations. Internal object identifiers are transformed into noun phrases by the selection of attributes that enable the listener to uniquely identify the intended referent. If an object cannot be characterized by attributes stored *a priori* in the partner model, it will be described by means of spatial relations, such as *`the left goal'*, or by means of events already mentioned in which it was (is) involved, e.g., *`the player who was attacked'*. Anaphoric expressions are generated if the referent is in focus and no ambiguity is possible.

To meet the requirements of simultaneous scene description, information concerning partly-recognized events is also provided. Consequently, language generation cannot start from completely worked-out conceptual contents; i.e., the need for an incremental generation strategy arises (see, e.g., Kempen and Hoenkamp [1987]). Consider Fig. 5: at the moment t_1 it has been detected that player 5 is transferring the ball, but the target of the pass has not yet been identified. The system starts to verbalize the proposition, but then the encoding process has to wait until the missing case role is filled. At the moment t_5 the event is completely recognized and the utterance can be continued.

5 Capabilities of our Current System

Since the first results described in Schirra et al. [1987], more than 3000 frames (120 seconds) of image sequences recorded from a major traffic intersection in Karlsruhe have been evaluated by the *Actions* system. Sung [1988] demonstrates with several examples that the results obtained from this image sequence make it possible to recognize complex activities such as driving towards and stopping in front of a traffic light until it changes to green, the length of various traffic light periods as well as turning and passing maneuvers. The calibration of the camera allows for the transformation of trajectories from the image plane into, for example, the street plane and thus a direct comparison of trajectories with a high resolution street map.

Since radio reports of soccer games are a good example of simultaneous descriptions, the method just described has been applied, with only minor changes, to more than 1000 frames of an image sequence recorded from a soccer game. Fig. 1 (see Section 3.1) includes a frame from the soccer sequence and Fig. 6 shows the projected trajectories of various players as they were automatically detected. The trajectories—shown from a bird's eye view in front of a goal—are printed over a map of the soccer field. This scene demonstrates the ability of *Actions* to deal with even non-rigid objects in a very different domain with remarkable results.

The as yet partial trajectories delivered by *Actions* are currently used to synthesize interactively a realistic GSD, with object candidates assigned to previously known players and the ball. Together with an instantiated model of the static background, this information forms the input for the *Vitra* system. Event recognition in *Vitra* is based on the approach described in Section 4.1. So far the role-fillers of events are restricted to being single objects; coordinated motions of groups of objects (e.g., an attack by a team) cannot be recognized yet. The language generation component of *Vitra* incorporates all the different processing modules sketched in Sections 4.2 and 4.3, especially an experimental module for the incremental generation of surface structures, which utilizes morphological processes of the system *Sutra* (see Busemann [1984]). Thus, the *Vitra* system can be regarded as a framework that may be used for further investigation of effects occurring in simultaneous reporting. The output window in Fig. 6 shows part of a typical description. The German text might be translated as: *`Munk, the midfieldman, has the ball. He passes the ball to Brandt, the sweeper. The sweeper cross kicks into the penalty area. Now Binkelmann, the goalie, has the ball.'*

6 Conclusion

We have presented an architecture and the currently available components for an integrated knowledge-based system capable of translating dynamic visual information into a *simultaneous* natural language description of the scene. We have shown that the various processing steps from raw images to natural language utterances, i.e., picture domain analysis of the image sequence, event recognition, and natural language generation, must be carried out on an *incremental* basis. Our approach emphasizes concurrent image sequence evaluation and natural language processing, an important prerequisite for real-time performance, which is the long-term goal of this work.

7 Technical Notes

Image processing has been done with a VTE Digital Video Disk and a VAX-11/780, programmed in Pascal. The current version of *Vitra* was implemented in Commonlisp and Flavors on Symbolics 3600 and 3640 Lisp-machines running Release 7.1. TCP/IP is used to connect the Symbolics machines to a Siemens 7.570 mainframe that serves as a gateway to the German Research Net (DFN) and the VAX in Karlsruhe.

Acknowledgements

We gratefully acknowledge the use of calibration programs developed by G. Saur and discussions with R. Kories. K. Rohr and K. Schreiner helped us to establish communication via DFN and we would also like to thank E. Hays and G. Retz-Schmidt for their helpful comments and suggestions throughout the development and the writing of this paper.

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. Herzog, T. Rist. On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER. In: *Proc. of the 8th ECAI*, pp. 449–454, Munich, 1988.

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, 1987.

N. Ayache, O. D. Faugeras. Building, Registrating, and Fusing Noisy Visual Maps. In: *Proc. of the First Int. Conf. on Computer Vision*, pp. 73–82, London, UK, 1987.

N. I. Badler. Temporal Scene Analysis: Conceptual Description of Object Movements. Technical Report 80, Computer Science Department, Univ. of Toronto, 1975.

R. Bajcsy, A. K. Joshi, E. Krotkov, A. Zwarico. LandScan: A Natural Language and Computer Vision System for Analyzing Aerial Images. In: *Proc. of the 9th IJCAI*, pp. 919–921, Los Angeles, CA, 1985.

S. Busemann. Surface Transformations during the Generation of Written German Sentences. In: L. Bolc, ed., *Natural Language Generation Systems*, pp. ??–??, Springer, Berlin, Heidelberg, New York, 1984.

O. D. Faugeras. A Few Steps toward Artificial 3D Vision. Report 790, Institut National de Recherche en Informatique et en Automatique INRIA, Domaine de Voluceau, Rocquencourt, Le Chesnay, France, 1988.

N. H. Goddard. Recognizing Animal Motion. In: *Proc. of Image Understanding Workshop*, pp. 938–944, San Mateo, CA, 1988.

H. P. Grice. Logic and Conversation. In: P. Cole, J. L. Morgan, eds., *Speech Acts*, pp. 41–58, Academic Press, London, 1975.

T. Kanade. Region Segmentation: Signal versus Semantics. *Computer Graphics and Image Processing*, **13**, 279–297, 1980.

G. Kempen, E. Hoenkamp. An Incremental Procedural Grammar for Sentence Formulation. *Cognitive Science*, **11**(2), 201–258, 1987.

R. Kories, G. Zimmermann. A Versatile Method for the Estimation of Displacement Vector Fields from Image Sequences. In: *Proc. of Workshop on Motion: Representation and Analysis*, pp. 101–106, Kiawah Island, Island Resort, Charleston, SC, 1986.

H.-H. Nagel. From Image Sequences Towards Conceptual Descriptions. *Image and Vision Computing*, **6**(2), 59–74, 1988a.

H.-H. Nagel. Image Sequences - Ten (Octal) Years - From Phenomenology Towards a Theoretical Foundation. *Int. Journal of Pattern Recognition and Artificial Intelligence*, **2**, 459–483, 1988b.

B. Neumann. Natural Language Description of Time-Varying Scenes. Report 105, Fachbereich Informatik, Univ. Hamburg, 1984.

H. Niemann, H. Bunke, I. Hofmann, G. Sagerer, F. Wolf, H. Feistel. A Knowledge Based System for Analysis of Gated Blood Pool Studies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **7**, 246–259, 1985.

H.-J. Novak. Generating a Coherent Text Describing a Traffic Scene. In: *Proc. of the 11th COLING*, pp. 570–575, Bonn, FRG, 1986.

N. Okada. SUPP: Understanding Moving Picture Patterns Based on Linguistic Knowledge. In: *Proc. of the 6th IJCAI*, pp. 690–692, Tokio, Japan, 1979.

G. Retz-Schmidt. A REPLAI of SOCCER: Recognizing Intentions in the Domain of Soccer Games. In: *Proc. of the 8th ECAI*, pp. 455–457, Munich, 1988.

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. Extraktion von typischen und komplexen Vorgängen aus einer langen Bildfolge einer Verkehrsszene. In: H. Bunke, O. Kübler, P. Stucki, eds., *Mustererkennung 1988; 10. DAGM Symposium*, pp. 90–96, Springer, Berlin, Heidelberg, 1988.

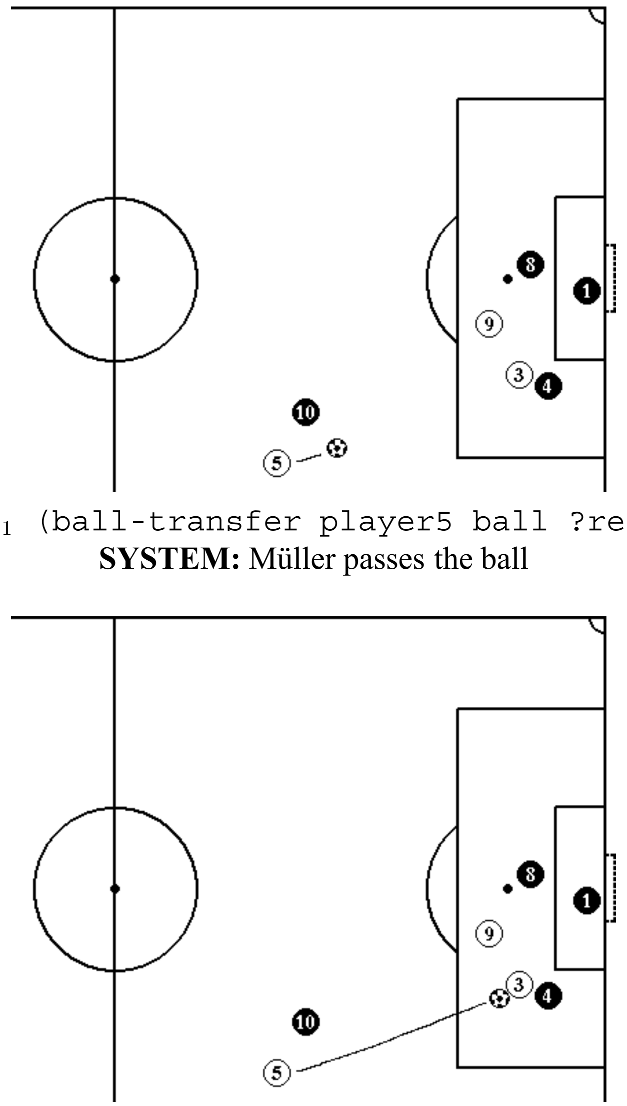
C.-K. Sung, G. Zimmermann. Detektion und Verfolgung mehrerer Objekte in Bildfolgen. In: G. Hartmann, ed., *Mustererkennung 1986; 8. DAGM–Symposium*, pp. 181–184, Springer, Berlin, Heidelberg, 1986.

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.

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, P. C. Lockemann, H.-H. Nagel. Database Support for Knowledge-Based Image Evaluation. In: P. M. Stocker, W. Kent, R. Hammersley, eds., *Proc. of the 13th Conf. on Very Large Databases, Brighton, UK*, pp. 3–11, Morgan Kaufmann, Los Altos, CA, 1988.

A. Witkin, M. Kass, D. Terzopoulos, K. Fleischer. Physically Based Modelling for Vision and Graphics. In: *Proc. of Image Understanding Workshop*, pp. 254–278, San Mateo, CA, 1988.



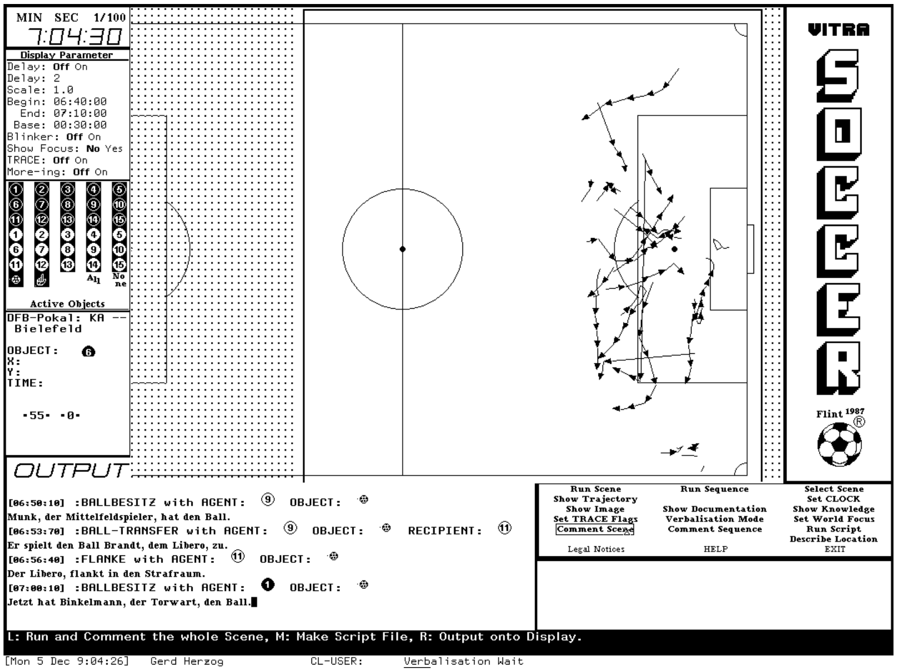
PROCEED(t_1 (ball-transfer player5 ball ?recipient))

SYSTEM: Müller passes the ball

STOP (t_5 (ball-transfer player5 ball player3))

SYSTEM: ... to the striker.

Figure 5: Incremental language generation



```
MIN  SEC  1/100
7:04:30

Display Parameter
Delay: Off On
Delay: 2
Scale: 1.0
Begin: 06:40:00
  End: 07:10:00
 Base: 00:30:00
Blinker: Off On
Show Focus: No Yes
TRACE: Off On
More-ing: Off On

Active Objects

DFB-Pokal: KA --
 Bielefeld

OBJECT:   ❻
X:
Y:
TIME:

  ▪55▪ ▪0▪

OUTPUT

[06:50:10] :BALLBESITZ with AGENT: ⑨ OBJECT: ⚽
Munk, der Mittelfeldspieler, hat den Ball.
[06:53:70] :BALL-TRANSFER with AGENT: ⑨ OBJECT: ⚽ RECIPIENT: ⑪
Er spielt den Ball Brandt, dem Libero, zu.
[06:56:40] :FLANKE with AGENT: ⑪ OBJECT: ⚽
Der Libero, flankt in den Strafraum.
[07:00:10] :BALLBESITZ with AGENT: ❶ OBJECT: ⚽
Jetzt hat Binkelmann, der Torwart, den Ball.█

L: Run and Comment the whole Scene, M: Make Script File, R: Output onto Display.
[Mon 5 Dec 9:04:26]   Gerd Herzog          CL-USER:     Verbalisation Wait
```

Figure 6: The basic windows of *Vitra-Soccer*