In: Computers and Artificial Intelligence, 8:479-492.

One Word Says More Than a Thousand Pictures

On the Automatic Verbalization of the Results of Image Sequence Analysis Systems

Wolfgang Wahlster
Computer Science Department
University of Saarbrücken Federal Republic of Germany

Abstract

This paper provides a compact introduction into the goals of the VITRA research project examining the possibilities of natural language description of images and presenting some of the project's results. In the combination of image understanding and natural language understanding systems, two important areas of research within Artificial Intelligence are brought together. After a description of the different domains of discourse of VITRA, the internal representation of spatial relations in the system is explained. This is followed by a few examples of how the knowledge of the meaning of natural language expressions pertaining to spatial concepts is used by the system in generating image descriptions in German. Since this brief presentation can only provide a first orientation, the paper contains numerous references to more detailed research reports.

1. Introduction

When seeing a series of TV pictures showing a part of a freeway where several hundred vehicles are lined up one behind the other, each one moving forward only at a snail's pace, we can sum up the scene with the expression 'traffic-jam'. This is a typical example of a large class of situations that can be described with the statement "one word says more than a thousand pictures" - a reversal of the classical saying.
One of the goals within Artificial Intelligence is to develop systems capable of translating visual information into natural language. This paper reports on the current state of development of the AI system VITRA (VIsual TRAnslator). The project, which has been funded by the German Science Foundation (DFG) since 1985, aims to make a contribution to basic research concerning the combination of image understanding and natural language understanding systems.

In the long run, two main goals are pursued in this research field:

(al) The complex information processing of humans underlying the interaction of natural language production and visual perception is to be described and explained exactly by means of the tools of computer science.

(a2) The natural language description of images is to provide the user with an easier access to, and a better understanding of, the results of an image understanding system.

It is characteristic of AI research that, apart from the cognitive science perspective (al), an application-oriented objective is also pursued (a2).
A great practical advantage of natural language image description is the possibility of the application-specific selection of varying degrees of condensation of visual information. The vast amount of visual data accumulating in medical technology, remote sensing and traffic control, for example, can only be handled by machine. As opposed to a representation of the results of processing digitized image sequences in the form of graphical output, a natural language description o I images can provide the user with more information in less time. If an AI system is capable of describing the results of interpreting an image sequence in a medical context as 'a stricture of the left kidney artery', the doctor can immediately classify this description according to the diagnostic context and later go back to specific segments of single relevant images, if necessary.

2. The Domains of Discourse in VITRA

At present, we are still far from a universally applicable AI system capable of describing an arbitrary sequence of images. We must concentrate on restricted domains of discourse (the possibilities and limits of knowledge-based systems in image and natural language understanding are presented in [Nagel 1985] and [Wahlster 1982]) in the system development process. In the VITRA project, four different domains of discourse and two communicative situations are examined in order to be able to verify the domain independence of the developed concepts and methods as early as possible:

Communicative Situation Cl:	answering of natural language queries about spatial relations and trajectories after processing a sequence of images
Domain of Discourse Dl:	CITYTOUR - segments of the city map of Saarbrücken with trajectories of moving objects
Domain of Discourse D2:	UNITOUR - map of the University of Saarbrücken campus
Domain of Discourse D3:	DURLACHER TOR - street scene in Karlsruhe with trajectories of moving objects
Communicative Situation C2:	simultaneous report about events observed while processing an image sequence
Domain of Discourse D4:	SOCCER - clips from television broadcasts of soccer matches

While the role of the AI system in (Cl) is similar to that of a local person giving directions, the system's role in (C2) is comparable to that of a radio commentator. The description is a posteriori in (Cl) as in the earlier systems HAM-ANS (cf. [Wahlster et al. 1983]) and NAOS (cf. [Neumann/Novak 1986]), while in (C2) the simultaneity of description poses entirely new problems (cf. [André et al. 1987] and [Rist et al. 1987]). For both types of situations, there are numerous scenarios for applications in real life. A biologist, as an example of the first type of situation, might ask "Where has there been damage to birches?" after the evaluation of a series of aerial photographs. In the case of the second type of situation, the operator in the control room of a complex technical system might require the description of an imminent malfunction or warning of a potential system breakdown.

While synthetic images are used in (Dl) and (D2) for which a geometrical description of scenes is given, the image sequences used in (D3) and (D4) are taken from natural surroundings. For (D3), the trajectories of moving objects are extracted by means of an image sequence analysis system developed by a research group at the ÜTB Karlsruhe (cf. [Schirra et al. 1987]), so that part of the geometrical description of scenes can be generated algorithmically (the geometrical description of the static background is still hand-coded). In the near future, we hope to be able to use the same procedure for (D4) as well, but the realization of this objective is made considerably more complicated by the fact that here, unlike in the case of (D3), no fixed camera position may be assumed and the moving objects are generally non-rigid. At present, examples of the soccer domain are still built up by means of a menu-driven graphic trajectory editor (cf. [Herzog 1986]).

The successful combination of the UTB image understanding system and VITRA has led, for the first time world-wide, to the development of an AI system capable of producing a natural language description of the motion of objects in a sequence of TV frames without any human intervention. In the course of this process, the visual data (between 250 K and 1 megabyte per image) and the resulting symbolic output (7 files with 12 K each) are transferred from a VAX computer at the ÜTB Karlsruhe to the Symbolics computer at the AI-Lab in Saarbrücken via three interconnected networks (DFN, CANTUS and Ethernet).

For the domain of discourse (D3), the traffic at the intersection in front of the Durlacher Tor in Karlsruhe was recorded using a stationary video camera from the roof of a 35-meter-high building (cf. [Schirra et al. 1987]). A sequence of 130 images (5.2 seconds total) were digitized

(512 by 512 pixels with an 8-bit greyscale) and processed by the ÜTB image analysis system. The system recognized ten candidates for moving objects and their trajectories, although they were partially occluded by trees and street-lamps. In the central lanes, a streetcar was moving from left to right and two small trucks, three cars and a bicycle were moving from right to left. A fourth car had appeared at the beginning of the processed image sequence and was therefore not considered by the motion detection system.

VITRA recognizes, based on the trajectories extracted, such higher-level motion concepts as three cars stopping at a traffic light or the streetcar going up Kaiserstrasse. Fig. 1 shows a hardcopy of the high-resolution graphics monitor of the LISP machine on which VITRA is implemented. The large graphics window on the right is displaying a digitized image from the processed sequence. The trajectories of the vehicles recognized are graphically superimposed over the original image. Each object trajectory, represented internally as a list of position-time pairs, is displayed graphically as a labeled edge containing numbered time markers, the distance and sequencing of which code the speed and direction of the observed motion, respectively. The small menu superimposed onto the graphics window allows the selection of a different domain of discourse via the mouse.

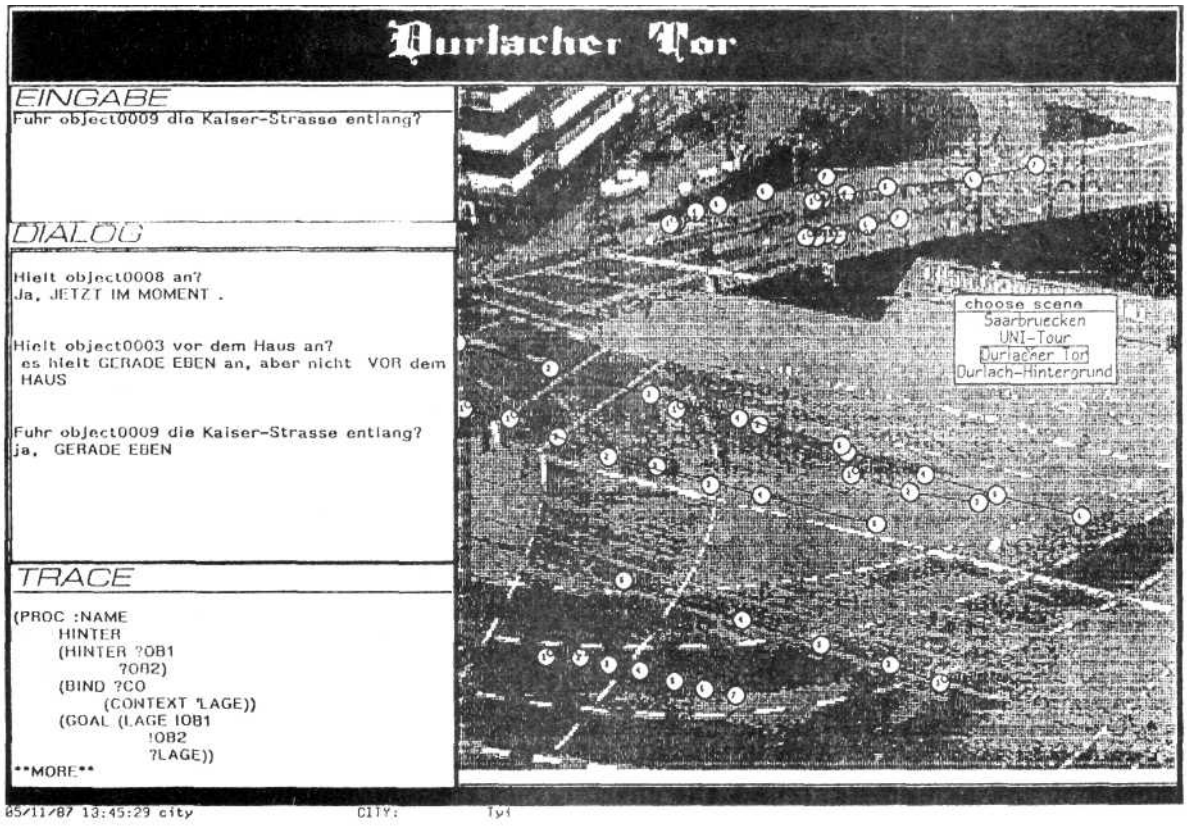
The left third of the screen contains three text windows: input, dialog and trace. German-language user queries are typed in the input window. The answer from VITRA appears in the dialog window along with a few preceding question-answer pairs forming part of the dialog context. In the trace window, it is possible for the user to follow the internal processes or to have the knowledge sources used by VITRA displayed.

The trace window and the graphics window are of particular importance for the further development of VITRA because it is with these aids that knowledge gaps of the system can be detected and the reasons for incorrect or inappropriate responses can be analyzed. In real life application environments for communicative situation (Cl), only the input and dialog windows are of interest, of course.

In Fig. 1, system-internal object labels are used (e.g. object 0009 instead of 'streetcar'). As the dialog examples in Figs. 2 and 3 show, this is not a restriction of VITRA but is merely due to the fact that the cooperating research group at ÜTB has not yet completed the object identification for this scene. In principle, there are several model-driven AI procedures potentially capable of assigning objects to the classes 'car', 'truck' and 'streetcar'. In the course of our project, however, we have so far concentrated more on the analysis of motion due to the methodological deficit in basic research in this area. The system's answers as displayed in the dialog window in Figs. 1 to 3 show that, apart from direct answers to user queries, some additional information is also generated (e.g. time-frame information, presupposition failures recognized, starting-point and end-point of a movement, modification of the predication queried). The additional informativeness of such answers to yes-no questions (cf. [Wahlster et al. 1983]) is an important prerequisite for the development of cooperative access systems capable of dynamically adapting their responses to the goals and knowledge of the user (cf. [Wahlster 1984]).

The trace window in Fig. 1 shows a segment of a knowledge source of VITRA in which the formal semantics of those spatial prepositions is defined which VITRA can both understand and generate. It is part of the definition of the relation 'behind', which has been coded in the AI programming language FUZZY. FUZZY is embedded in LISP and contains additional mechanisms for automatic inferencing, access to associative databases and pattern-directed procedure invocation, among others.

Before the representation of the semantics of such prepositions is illustrated with examples in section 4, the geometrical description of scenes used in VITRA as the basis for processing spatial relations will first be briefly outlined in the following section using the domains UNITOUR and CITYTOUR.



Translation of the German examples in the dialog window

Did object0008 stop?
Yes, JUST NOW.

Did object0003 stop in front of the house?
It stopped A MOMEMT AGO, but not IN FRONT OF the HOUSE.

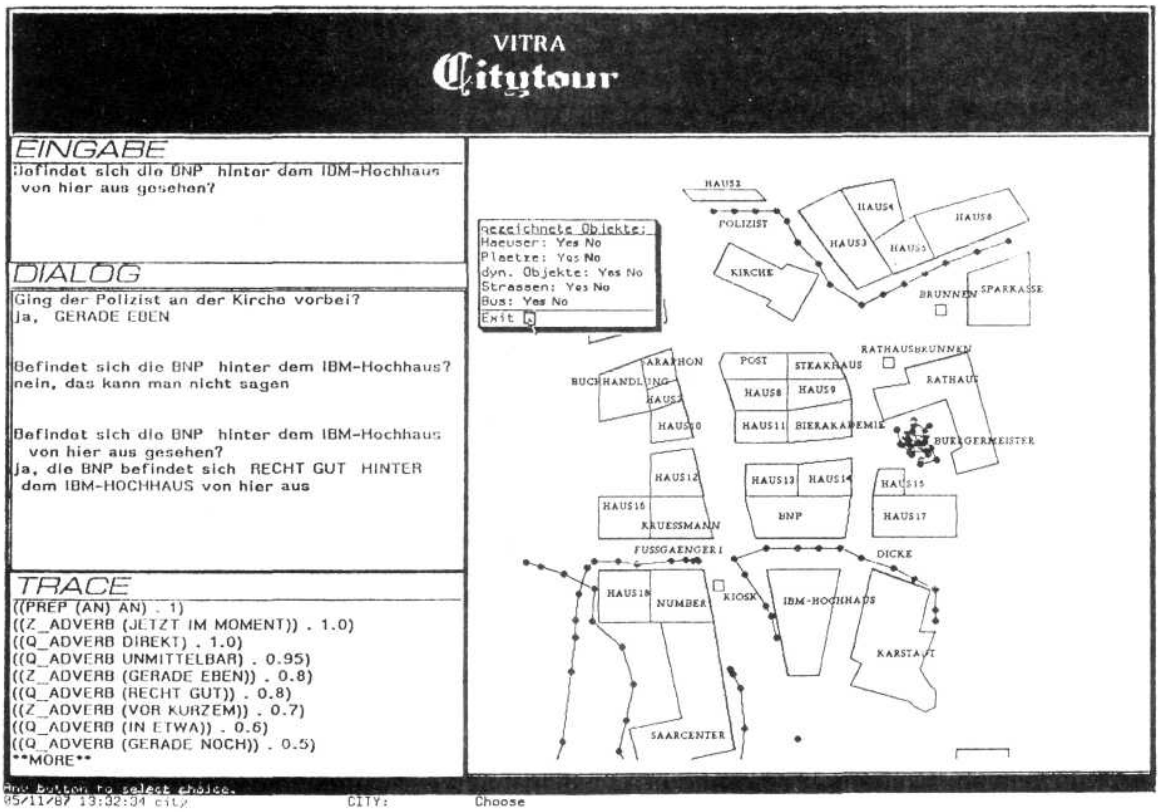
Did object0009 go along Kaiserstreet?
Yes, A MOMENT AGO.

Fig. 1: Example dialog with VITRA about motions at the Durlacher Tor

3. The Representation of Spatial Relations

In CITYTOUR, a distinction is made between static (i.e. immobile) and dynamic (i.e. mobile) objects. Examples of static objects are, among others, streets, houses and squares (cf. Fig. 2). Streets are represented by their left and right curbs as well as by the centerline. All other static objects are represented as polygons in the geometrical scene description. From the polygons, delineative rectangles and their centroids can be calculated, which provide coarser-grained forms of representation for the fast approximate computation of spatial relations. A further special feature defined for certain types of static objects is their prominent front (such as the main entrance of the post office in Fig. 2), displayed in the graphics window as a bold edge (cf. Fig. 2). By means of a mouse-controlled menu which can be called up in the graphics window, the set of object classes to be displayed can be defined. In Fig. 2 all internally represented objects except the streets are visible.
The dynamic objects in CITYTOUR are pedestrians, bicyclists and motor vehicles (cars, buses and streetcars). They are not differentiated in the representation; they are all represented as centroids. Their trajectories are represented as lists of position-time pairs, with the positions themselves being represented as x-y coordinate pairs. On the screen, the trajectories are projected onto the static background for the entire duration of a scene (cf. Fig. 2). A spatial relation is represented as an atomic formula, in the sense of predicate logic, consisting of one

predicate and several terms. The predicate corresponds to a preposition in natural language (e.g. 'in front of, 'at' and 'to the left of). The first argument is referred to as the 'subject'. It is the object which is to be localized via its relation to one (as in most cases) or more (e.g. in the case of' between', cf. Fig. 4) reference objects.
A distinction is made between static relations, in which the subject as well as the reference object are immobile, and dynamic relations, in which the subject is mobile. So far, only immobile reference objects can be handled.
While static relations indicate the position of immobile objects, both the direction (e.g. 'to go behind the city hall') and course of the trajectories (e.g. 'to go past the Saar Center' or 'to turn off) of mobile objects can be described by means of dynamic relations. Apart from the dynamic relations, the semantics of some verbs of motion such as 'to stop' and 'to start' is implemented. Events characterized by verbs of motion as well as immobile objects can be localized via static relations (e.g. 'to stop in front of the post office').
Some prepositions permit a localization of the subject either with respect to intrinsic qualities of the reference object or in relation to another observer viewpoint. In the first case, we speak of an intrinsic reading of the preposition and of an extrinsic reading of the preposition in the second case. If the viewpoint of the observer is the same as that of the speaker or hearer, we speak of a deictic reading (cf. (Retz-Schmidt 1986b] for an extensive discussion of this distinction).



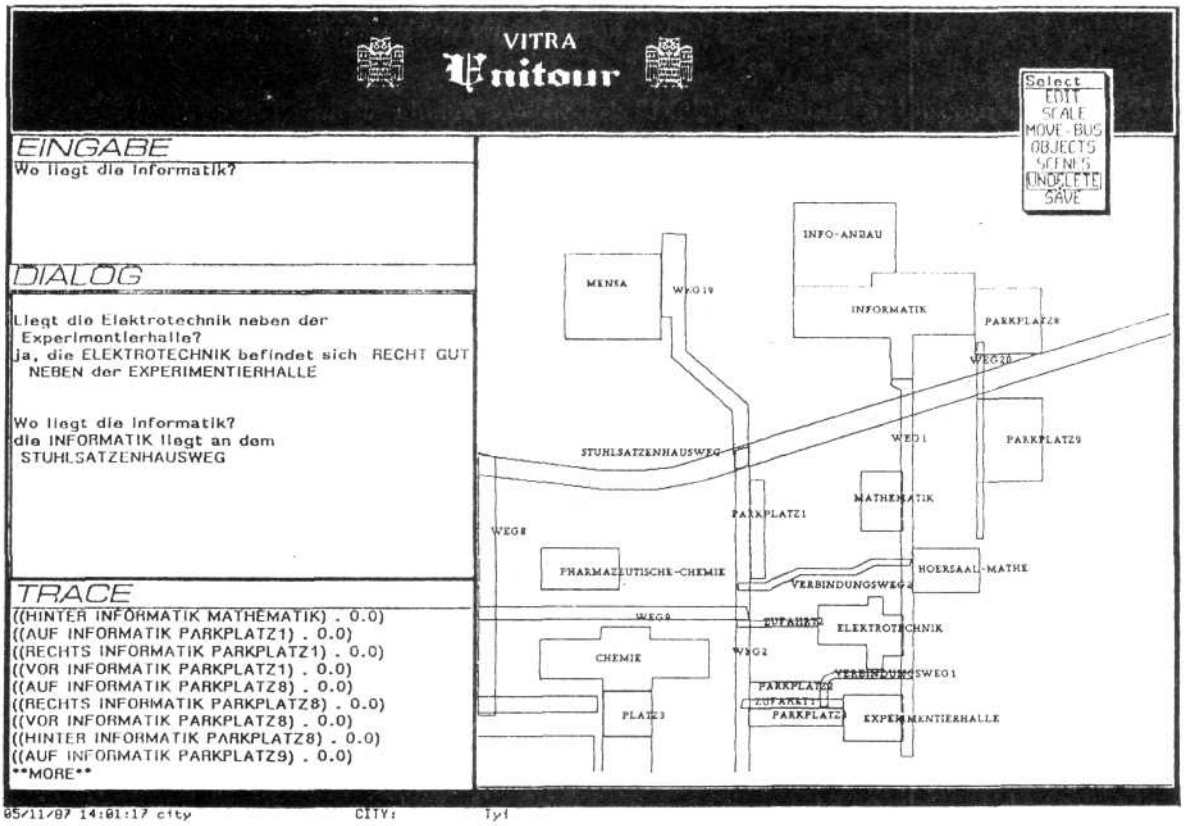
Translation of the German examples in the dialog window
Did the policeman go past the church?
Yes, JUST NOW.

Is the BNP behind the IBM-Tower?
No, I would not say so.

Is the BNP behind the IBM-Tower?
Yea, the BNP is WELL BEHIND the IBM-Tower from here.

Fig. 2: Example Dialog for the CITYTOUR Domain of Discourse

An example illustrating this difference can be seen in Fig. 2. The speaker's as well as the hearer's viewpoint is assumed to be the single thick black dot (cf. the dot below the polygon marked 'IBM Hochhaus' in the lower center of the graphics window in Fig. 2). They can be imagined as being tourist and guide on a city tour, for example, with the tourist asking questions about the location of various buildings in the city. In communicative situation (Cl), VITRA permits the free positioning of the fictitious speaker or hearer viewpoint by means of graphic operations (moving the corresponding dot) in order to test and demonstrate the situational dependency of spatial descriptions.
VITRA permits the intrinsic and deictic reading of the following prepositions: 'in front of, 'behind', 'to (he right of, 'to the left of and 'next to' in their locative and directional readings and also in combination with 'past'.
In the case of locative reading, the system analyzes whether the respective relation holds as follows: The area surrounding the reference object is divided into four half planes - front, rear, left and right. These four half planes are defined by the four sides of a rectangle delineating the reference object. In the case of the intrinsic reading, the rectangle is oriented with respect to the front of the reference object. In the case of the deictic reading, the delineative rectangle is oriented with respect to the line of sight from the observer's viewpoint to the reference object (for further details, see [André et al. 1986b]). In both cases it is then possible to determine which of the four half planes the subject is located in.



Translation of the German examples in the dialog window

Is Electrical Engineering next to the experiment building?
Yes, ELECTRICAL ENGINEERING is WELL NEXT TO the EXPERIMENT BUILDING.

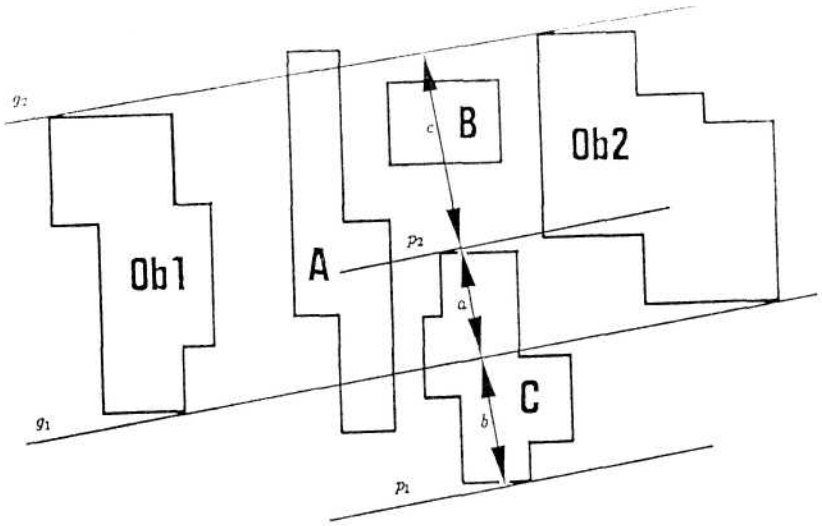
Where is Computer Science?
COMPUTER SCIENCE is on STUHLSATZENHAUSWEG.

Fig. 3: Example Dialog for the UNITOUR Domain of Discourse

When deciding as to the intrinsic vs. the deictic reading, VITRA applies the following strategy - after [Miller/Johnson-Laird 1976]: The intrinsic reading is assumed as long as the deictic reading is not explicitly indicated (e.g. through an expression such as 'from here') and the reference object has a prominent front. Otherwise, the deictic reading is assumed.
As the examples in Fig. 2 show, the system states that, in an intrinsic interpretation, the BNP does not lie behind the IBM-Tower whereas the deictic reading leads to an affirmative answer to the input question.

4. The Semantics of Spatial Prepositions

Although the four half planes surrounding the reference objects are infinite, it does not seem plausible that the relation 'in front of', for example, would still be true even if the subject is very far away from the reference object. On the other hand, it appears unreasonable to draw a sharp line between an area for which a spatial relation holds and another area for which this relation no longer holds.
Therefore, so called degrees of applicability are calculated for the relations depending upon the distance between subject and reference object as well as upon the size of the reference object. These degrees of applicability are represented internally as values from the real interval [0,1].



Step 1:

Calculate the two tangents g1and g2 between the reference objects using their closed-polygon representation;

Step 2: If:

A: both tangents cross the subject (also in its polygon representation), the relationship *between* holds with degree 1;

B: the subject is totally enclosed by the tangents and the reference objects, the relationship is also applicable with degree 1;

C: only one of the tangents intersects the subject, the degree of applicability is calculated, depending on its penetration depth in the area between the tangents:

$$\text{applicability degree} = max\left[\frac{a}{a+b}, \frac{a}{a+c}\right]$$

Otherwise:

D: the relationship is not applicable: degree = 0;

Fig. 4: Degrees of Applicability of the Spatial Relation 'between'

When answering yes-no questions, these degrees of applicability are verbalized as linguistic

hedges, as in the example:

'Does the City Hall fountain lie in front of the Beer-Academy?'
'Yes, the City Hall fountain lies approximately in front of the Beer-Academy.'

When answering where-questions (cf. Fig. 3), on the other hand, these degrees of applicability can serve to select the most suitable reference object. In such cases, the applicability of the four basic relations for different nearby reference objects is examined. A measure of saliency is included in the calculation before choosing the reference object with the highest degree of applicability for describing the location.

'Where is the Computer Science building?'
'The Computer Science building is on Stuhlsatzenhausweg.'

As an example for the calculation of degrees of applicability, Fig. 4 illustrates the procedure for llie static relation 'between'. For the three possible subjects A, B and G, the extent to which the relation 'between' holds with respect to reference objects OBI and OB2 is examined. If none of the cases A, B or C is given, the degree of applicability is set to 0 which corresponds to the negation of the atomic formula in question.
For the determination of dynamic relations, it is assumed in communicative situation Gl that the trajectories are known in their entirety. The trajectories are projected into the static scene (cf. Fig. 2) and the last dot in a trajectory line marks the position of the moving object at the tense-logical present.
Earlier work on natural language image description was based on a far simpler - compared with VITRA - geometric scene description in which the static objects were also represented by centroid coordinates (cf. [Wahlster et al. 1978]). A result of this method was that the semantics of path prepositions such as 'past' and 'along', which relate the trajectories of moving subjects to the delimiting line or plane of another object, could not be described adequately. In VITRA, however, the computational semantics of the path prepositions 'along' and 'past' could be examined more closely for the first time, uncovering the following differences between them:
In both cases, the distance between the trajectory of the moving subject and the reference object may not exceed a certain threshold depending on the size of the reference object. In the case of 'along', this threshold is lower than in the case of 'past'. Furthermore, in the case of 'along', the trajectory must follow the contours of the reference object more closely than in the case of 'past'. While a delineative rectangle (the less detailed form of representation) is sufficient for the preposition 'past', the polygonal representation of the reference object is necessary for the calculation in the case of 'along'. Another difference is that, in the case of 'along', the trajectory may not reverse direction in the course of the motion being described. Whereas in the case of 'past the front/the back/the left side/the right side of the entire area between opposite half planes must be crossed; in the case of 'along', only a sufficiently long path along the reference object has to be covered. Fig. 4 shows a few examples of trajectories illustrating the difference between 'along' and 'past'. The differences sketched here were integrated into the semantics of'past' and 'along' as implemented in VITRA (cf. [André et al. 1986a]).

5. Current Research Tasks

The preceding sections have shown that the Al systems developed in VITRA are already capable of generating simple natural language descriptions of image sequences. But the performance of system components developed to date is still far from the verbal, communicative and cognitive performance of a human observer. Therefore, in the second project phase from 1988 to 1990, numerous open questions concerning the formal reconstruction of the interplay between 'seeing' and 'speaking' must be further explored and resolved. In this section, three of these current questions will be discussed in the context of the SOCCER domain:
(a) the recognition and description of groups of moving objects;

(b) the influence of the system's assumptions about the intentions behind the observed actions upon the description of movements; and
(c) the development of an imagination component facilitating a feedback loop.

On a soccer field, twenty-six objects can move simultaneously: twenty-two players, one referee, two linesmen and the ball. A statement made during a TV or radio broadcast can sum up the movement of a single object (1) or of all twenty-six objects together (2) as the extreme cases.

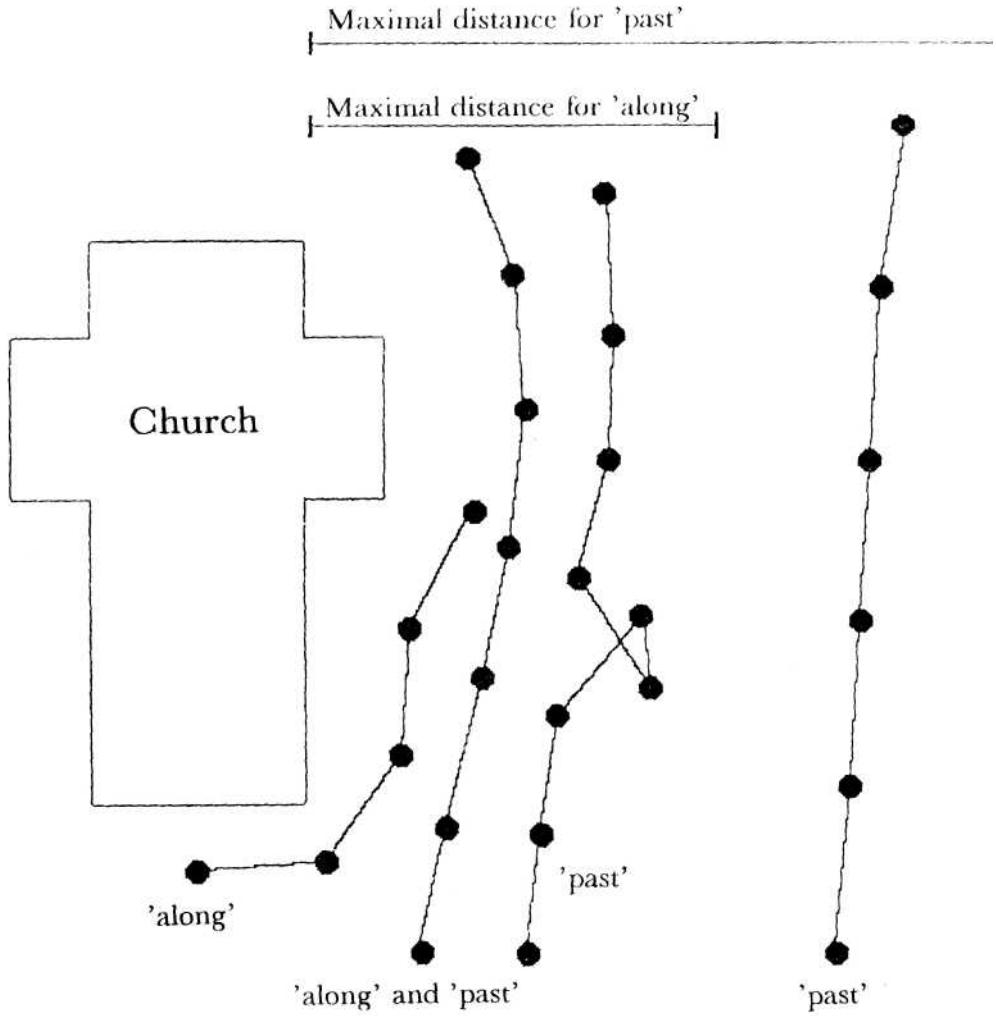


Fig. 5: Example Trajectories for Two Path Prepositions

(1) The ball is rolling out of play.
(2) The game is slowing down.

Note that the descriptions can refer to arbitrary elements of the power set of the moving objects

(3).

(3) The defense of the Birmingham Bombers is forming a wall.

This illustrates the enormity of the task of selection which a human observer must cope with in an extremely short time since, arithmetically, s/he must choose from 2^{26} - 1 (i.e. 67,108,863) possible combinations.

Another problem which already appeared in the case of street scenes and which is generally left unsolved is that not only the course of a trajectory in time and space is of decisive importance for the selection of

an adequate description. For example, even if all temporal and spatial requirements for the description of the observed trajectory of a moving vehicle are met, a description such as (4) might still be felt to be inadequate.

(4) The car is parking in front of the traffic lights.

Only by taking the intention behind an action into consideration (cf. [Retz-Schmidt 1986a]) can an adequate description of the same trajectory be given as in (5).

(5) The car is wailing in front of the traffic lights.

One criterion for the choice of soccer as a domain of discourse was the fact that the influence of the agents' assumed intentions on the description is particularly obvious here. Thus, (7) and (8) describe the same process in time-space but imply different team membership for player Meyer.

(7) Meyer kicked the ball out of play next to the goal.

(8) Meyer barely missed the goal.

In (7), the player has no intention of getting the ball into the goal, but deliberately kicks it out of play. In (8), by comparison, the player's kick was clearly aimed at the goal as expressed in the verb 'miss'.

One advantage of this domain of discourse is that, given the position of players, their team membership and the distribution of roles in standard situations (e.g. penalties and corners), stereotypical intentions can be assumed for each situation. Given the current state of plan recognition research, then, the chances of successfully reaching our working goal (b) are better than in other less schematized situations.

A third problem which we are currently analyzing arises from the fact that the system, in order to generate communicatively adequate descriptions, must construct a model of the visual conceptualizations which the system's utterance elicits in the hearer's mind. Such a user model (cf. [Wahlster/Kobsa 1986]) can become relevant for the decision during sentence generation as to whether, instead of a definite description, a pronoun might also be understandable for the hearer.

Let us suppose that the system has just generated the following text as a description of an observed situation:

(9) In the left half, Jones is running toward the goal with the ball. Meyer is chasing him and trying to attack him. But Meyer is too slow.

Since it is not possible for the hearer to visually follow the action on the field, s/he can only form a rough idea of the spatial setting. It is imperative that the system be able to put itself into the hearer's place and take the hearer's possibly imagined conceptualization into account before continuing to generate sentences.

Fig. 6 shows a possible graphic representation of the conceptualization in the hearer's mind elicited by (9). The courses of the trajectories shown must be regarded as prototypical and the surrounding ellipses indicate the degree of leeway in spatial interpretation. If the system is planning to generate sentence (10), it must decide, in order to conform with the conversational maxim of cooperativity, whether the referent of the pronoun 'him' can be unambigously determined by the hearer.

(10) Now only the goal keeper is between him and the goal.

Only 'Jones' and 'Meyer' in the preceding text are possible referents for the pronoun. Since Meyer was mentioned last, this referent is the first to suggest itself to the reader in purely

textual terms. In the sense of an anticipation feedback loop (cf. [Jameson/Wahlster 1982]), however, the system could recognize that this resolution of the anaphora is inconsistent with the assumed spatial conceptualization in the hearer's mind by accessing the imagination component of the user model. Therefore, 'Jones' is the only unambiguous referent for the pronoun that is compatible with the user model. Only after such a successful understanding process has been anticipated should the planned sentence be generated. Otherwise, the system would not be able to employ pronominalization to shorten its sentences but would have to resort to the use of proper nouns, for example.

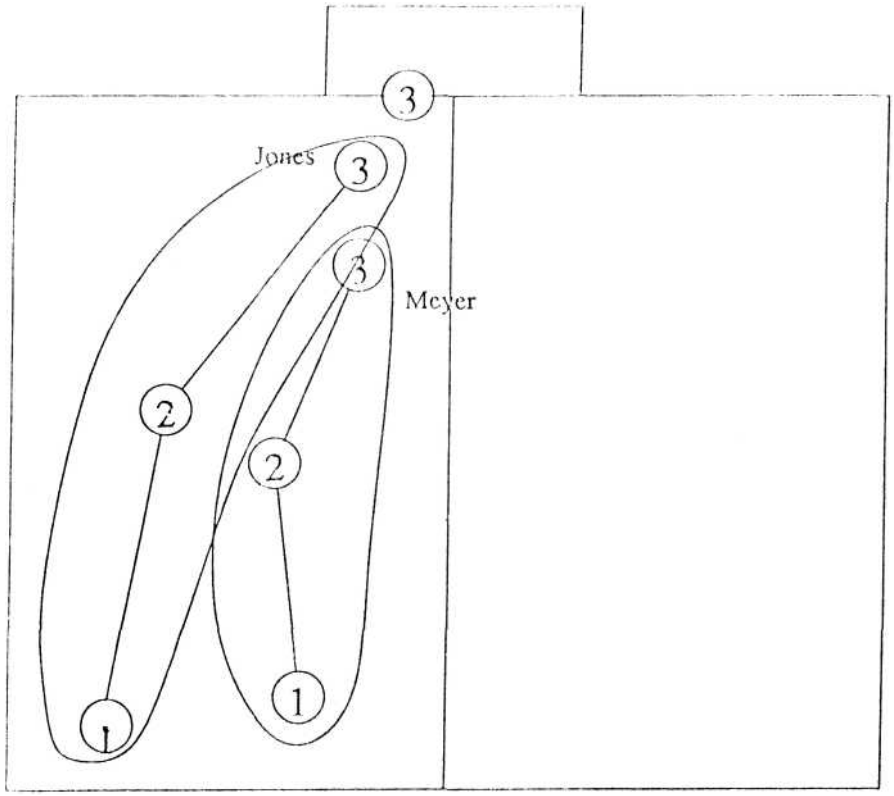


Fig. 6: Imagination Derived from Anticipated Understanding

Acknowledgements

Without the successful cooperation with my colleague H.-H. Nagel at the ÜTB Karlsruhe that began as early as ten years ago at the University of Hamburg, the combination of two AI systems as described here would certainly not have been possible. Our thanks go to his colleagues G. Zimmermann and C. K. Sung for the fruitful cooperation over the last three years. As members of the VITRA project, G. Retz-Schmidt and J. Schirra made fundamental contributions to the design and realization of the project. Important ideas as well as the entire implementation were provided by the research assistants and diploma candidates E. André, G. Bosch, G. Herzog, T. Rist and I. Wellner who contributed decisively to the project's progress. Thanks go also to Mark Line for translating the German version of this paper.

André, E.; Bosch, G.; Herzog, G.; Rist, T. (1986a): Characterizing Trajectories of Moving Objects

Using Natural Language Path Descriptions. Memo No. 5, SFB 314, Computer Science Dept., University of Saarbrücken. Also in: Proc. of the 7th ECAI, July 1986, Brighton, England, Vol. 2, 1-8.

André, E.; Bosch, G.; Herzog, G.; Rist, T. (1986b): Coping with the Intrinsic and Deictic Uses of Spatial Prepositions. Report No. 9, SFB 314, Computer Science Dept., University of Saarbrücken. Also in: Proc. of AIMSA 1986, Varna, Bulgaria.

André, E.; Rist, T.; Herzog, G. (1987): Generierung natürlichsprachlicher Äußerungen zur simultanen Beschreibung von zeitveränderlichen Systemen. Report No. 18, SFB 314, also in: Morik, K. (ed.): GWAI-87, 11th German Workshop on Artificial Intelligence. Berlin/Heidelberg/New York/Tokyo: Springer.

Herzog, G. (1986): Ein Werkzeug zur Visualisierung und Generierung von geometrischen Bildfolgenbeschreibungen. Memo No. 12, SFB 314, Computer Science Dept., University of Saarbrücken.

Jameson, A.; Wahlster, W. (1982): User Modelling in Anaphora Generation: Ellipsis and Definite Descriptions. In: Proc. of 1st ECAI, Orsay 1982, 222-227.

Miller. G.A.; Johnson-Laird, P.N. (1976): Language and Perception. Cambridge: Cambridge University Press.

Nagel, FI.-H. (1985): Wissensgestützte Ansatze beim maschinellen Sehen: Helfen Sie in der Praxis? in: Brauer, W., Radig, B. (eds.): Wissensbasierte Systeme. GI-Kongrefi 1985. Informatik-Fachberichte, Berlin/Heidelberg/New York/Tokyo: Springer.

Neumann, B.; Novak, H.-J. (1986): NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen. In: Informatik - Forschung und Entwicklung (1986) 1, 83-92.

Retz-Schmidt, G. (1986a): Script-Based Generation and Evaluation of Expectations in Traffic Scenes. In: H. Stoyan (ed.) (1986): GWAI-85, 9. Fachtagung über Künstliche Intelligenz, Informatik-Fachberichte, Berlin/Heidelberg/New York/Tokyo: Springer.

Retz Schmidt, G. (1986b): Deictic and Intrinsic Uses of Spatial Prepositions: A Multidiscipli-nary Comparison. Memo No. 13, SFB 314, Computer Science Dept., University of Saarbrücken. Also in: Proc. of the Workshop on Spatial Reasoning and Multi-Sensor Fusion, St. Charles, Illinois, Oct. 1987, Morgan Kaufmann.

Rist, T.; Herzog, G.; André, E. (1987): Ereignismodellierung zur inkrementellen High-level Bildfolgenanalyse. Report No. 19, SFB 314, Computer Science Dept., University of Saarbrücken, also in: Buchberger, E.; Retti, J. (eds.): OGAI-87. 3.Österreichische AI-Tagung. Informatik-Fachberichte, Berlin/Heidelberg/New York/Tokyo: Springer.

Schirra, J.R.J., Bosch, G., Sung, C.K., Zimmermann, G. (1987): From Image Sequences to Natural Language: A First Step towards Automatic Description of Motions. University of Saarbrücken, SFB 314 (VITRA), Report No. 26. To appear in: Applied Artificial Intelligence, 1988.

Wahlster, W (1982): Natürlichsprachliche Systeme. Eine Einführung in die sprach-orientierte KI-Forschung. In: Bibel, W.; Siekmann, J.I1. (eds.): Künstliche Intelligenz. Frühjahrsschule der Gl. Informatik-Fachberichte. Berlin/Heidel-berg/New York/Tokyo: Springer.

Wahlster, W. (1984): Cooperative Access Systems. In: Future Generations Computer Systems, Vol. 1, No. 2, 103-111.

Wahlster, W.; Jameson, A.; Hoeppner, W. (1978): Glancing, Referring and Explaining in the Dialogue System HAM-RPM. In: American Journal of Computational Linguistics, 53-67.

Wahlster, W.; Marburger, H.; Jameson, A.; Busemann, S. (1983): Over-answering Yes-No Questions: Extended Responses in a NL Interface to a Vision System. In: Proc. of the 8th IJCAI, Karlsruhe.

Wahlster, W.; Kobsa, A. (1986): Dialog-based User Models. In: Proceedings of the IEEE 74(7), 948-960 (Special Issue on Natural Language Processing).