TOWARDS NATURAL LANGUAGE DESCRIPTION OF REAL-WORLD IMAGE SEQUENCES

Bernd Neumann

IfI-HH-M-101/82

November 1982

This paper has been presented at the 12. GI-Jahrestagung, October 5-7, 1982, Kaiserslautern, W. Germany. Proceedings as GI-12.Jahrestagung (J. Nehmer, ed.), Informatik-Fachberichte 57, Springer-Verlag Berlin/Heidelberg/New York, 1982, 349-358

Entwicklung eines Systems zur natürlichsprachlichen Beschreibung von Realwelt-Bildfolgen

Zusammenfassung

Eine Computer-Analyse von Bildfolgen mit Szenen der realen Welt kann Bilddeutungen auf einem genügend hohen Abstraktionsniveau erzeugen, um eine Kopplung mit einem natürlichsprachlichen System zu erlauben. Allerdings können viele natürlichsprachlichen Bewegungskonzepte nicht ausschließlich aufgrund von Ergebnissen der Szenenanalyse berechnet werden. Sie erfordern häufig zusätzliches Wissen und Beschreibungsebenen, die zwischen sprachlichem Konzept und Szenenrepräsentation vermitteln. Diese Mitteilung berichtet über Fortschritte von Projekt NAOS), das sich mit der automatischen natürlichsprachlichen Beschreibung von Verkehrsszenen befaßt. Die Szenenanalyse erzeugt mit Hilfe von mehreren bereits vorhandenen Prozessen zunächst eine geometrische Szenenbeschreibung. Diese stellt die Eingabe für höhere Bilddeutungsprozesse dar, Bewegungsinterpretationen die ihrerseits an ein vorhandenes natürlichsprachliches Dialogsystem weitergeben. Einfache Bewegungen können durch Prozeduren erkannt werden, die direkt mit einer Kasusrahmen-Repräsentation der entsprechenden Verben verknüpft ist. Komplexere Bewegungskonzepte benötigen dagegen zusätzliche Strukturen, insbesondere eine explizite Repräsentation von Erfahrung.*

*) *Dieses Projekt wird teilweise von der Deutschen Forschungsgemeinschaft gefördert.*

TOWARDS NATURAL LANGUAGE DESCRIPTION OF REAL-WORLD IMAGE SEQUENCES

Bernd Neumann

Fachbereich Informatik
Universität Hamburg
Schlüterstrasse 70
D-2000 Hamburg 13

Abstract

Computer analysis of image sequences showing real-world scenes with motion may generate interpretations at a sufficiently high level of abstraction to permit interfacing with a natural language system. Yet many natural language motion concepts cannot be evaluated solely from scene analysis results but require additional knowledge and representations at intermediate levels of abstraction. This paper reports about progress of project NAOS (*) towards natural language description of traffic scenes. In the scene analysis stage several existing processes have been combined to produce a geometrical scene description. This is the input for high-level motion interpretation processes which are linked to an existing natural language dialogue system. While simple motion processes could be implemented using procedures attached to case frame representations of verbs, more complex concepts call for an explicit encoding of experience.

1. Introduction

The task of analyzing an image sequence to derive a natural language description involves two major areas of Artificial Intelligence - image understanding and natural language understanding - which have been studied rather independently from each other. Image understanding aims at computing meaningful interpretations from images. Difficult problems arise already in very early processing stages, e.g. when extracting object contours, and hence much work is centered around raw images and

---

(*) This project is partially supported by the Deutsche Forschungsgemeinschaft.

their properties. If interpretations are attempted, they are usually given in terms of identified objects or object configurations. While this is a level of abstraction where object names can be given and hence natural language comes into play, none of the major experimental image understanding systems (as covered e.g. in HANSON and RISEMAN 78a) renders scene descriptions in natural language. Typically, lists or graphics are provided from which the system performance can be inferred.

Similarly, in natural language research work centered mainly around language phenomena and, if concerned with images at all, pictorial input was simulated at the level of abstraction required by the particular world of discourse. When talking about a scene with the dialogue system HAM-RPM, for example, a symbolic data base is addressed where objects are represented as points in bird's-eye view augmented by a rich semantic description in terms of natural language concepts (v.HAHN et al. 80). Another well-known example is Winograd's SHRDLU (WINOGRAD 72) where one can talk about a simulated toy scene containing blocks, pyramids and cylinders. Here, spatial relations like "left of" are evaluated making direct use of a geometrical scene description. Similar problems have also been investigated in SWYSS (SCHEFE 81, HANSSMANN 80) for an interieur scene with various objects represented by polygonal contours.

Some new motivation for studying the border area between natural language and image understanding stems from the increased interest in time-varying scenes, which are also the subject of this report. The analysis of image sequences showing the temporal development of a scene raises several issues which could be ignored or given minor attention in single image analysis. First, there is the very practical problem of how to represent the interpretation of a sequence of - say - 250 images covering a time span of 10 sec. Lists or graphics showing, for example, object locations for each individual image, are clearly no solution. NAGEL 77 foresaw this difficulty and proposed natural language descriptions as a means to interface with the human

experimentator.

Underneath this problem, however, lies a more fundamental question which is discussed in WALTZ 79: What is the output of a (complete) vision system? More specifically, what are the levels of abstraction and which concepts should be used for representing the output of a vision system? Waltz suggests to study vision and natural language in common since both may share common concepts. In WALTZ 81 several levels of abstraction are proposed, ranging from geometrical 3D-models at the vision end to case frames at the language end. Waltz also points out kinds of knowledge which are required to judge the plausibility of a natural language description for an imagined scene ('My dachshund bit our mailman on the ear'). He proposes scripts and 'subscripts' to represent events and verb meanings.

One of the primary goals of image sequence analysis is the interpretation of image variations in terms of moving objects. This can be done at a rather low level of abstraction without using special domain knowledge by tracking nonstationary image regions and interpreting these in terms of rigid bodies. An approach like this has been persued by Nagel and collaborators (DRESCHLER and NAGEL 81) and will be described in the next section with more details since it provides part of the input for the scene description project NAOS. Given a particular domain, e.g. street scenes with children playing ball, image sequence analysis may provide motion interpretations in terms of higher level concepts like 'bounce', 'roll', 'drop', etc. which are much closer to but not necessarily identical with natural language concepts. BADLER 75 showed how these kinds of motion concepts can be derived from a sequence of computer generated line drawings. His work focusses on the process of decomposing motion into a sequence of primitives which in turn can be combined into complex motion patterns corresponding to some natural language concept.

The work of Tsotsos (TSOTSOS 80) builds upon some of the motion

concepts developed by Badler but also improves on Badler's framework in several respects. Most importantly, a hierarchy of conceptual motion frames is defined without presupposing domain specific knowledge. Thus certain concepts like 'area-change' may be applied to left ventricular wall motion (Tsotsos' problem domain) as well as street scenes. Problem specific motion frames, e.g. 'beat' (of a heart), may be defined in terms of lower-level concepts. Tsotsos also offers a control structure based on the 'hypothesize-and-test' paradigm and data-directed hypothesis selection.

It is interesting to note that Tsotsos, who did not attempt to provide natural language output, based his motion concepts on categories developed by MILLER 72 for motion verbs of the English language. Tsotsos realized, however, that some of Miller's categories could not easily be incorporated into a vision system, e.g. 'causative' or 'permissive' motion, and chose to exclude them, while other concepts, e.g. 'inchoative' motion, do not pose severe problems at all. This has also been observed in MARBURGER et al. 81 which is the first report on the NAOS scene description project.

A different set of motion concepts underlies the system SUPP (OKADA 80) which produces sentences from a short sequence of stylized 2D line drawings showing, for example, a bird landing on a tree, a man entering a car, a dog sliding down a ramp. Okada uses 20 semantic features, e.g. 'displacement', 'deformation', 'change in quality', 'start and stop', to decide which of a set of about 1200 primitive verb concepts applies to a given scene. Usually, many concepts qualify giving rise to as many simple sentences describing the same event.

This concludes the presentation of previous work related to natural language description of image sequences. Several worthy contributions have not been mentioned, e.g. the work of Tsuji and collaborators (ABE et al. 81). It is felt, however, that some of the interesting problems specific to this area have been

exposed. They can be summarized by the following questions:

1. Which concepts underly natural language descriptions of motion?

2. What are the motion concepts computable from image sequences?

3. What knowledge is required in addition to the image data?

The following sections refer to project NAOS which - in its current phase - tries to answer these questions for a special domain: traffic scenes as viewed from a stationary observer. Some of the typical motions can be seen in Fig. 1 showing 4 out of a sequence of 64 images with an overall duration of approximately 13 seconds. Pedestrians walk, cross the street, cars start and stop, turn off right, etc.

Figure 1: Images of a traffic scene to be described by NAOS

Scenes like this have been used for many years as experimental data for image understanding research at Hamburg University. For this project, the scene analysis processes developed so far are combined to produce a low-level scene interpretation in terms of object names and trajectories called 'geometrical scene description'. This is the first stage of processing towards obtaining a natural language description. Many of the processes of this stage are subject to ongoing research, but nevertheless a certain architecture and certain representational forms are emerging. They are described in section 2.

In the second stage of processing the geometrical scene description is taken as input for processes which connect to the natural-language dialogue system HAM-ANS (the successor of HAM-RPM, v.HAHN et al. 79). These processes will be the core of NAOS. So far, only yes/no questions pertaining to simple motion concepts are implemented (MARBURGER et al. 81). More complex motion verbs have been theoretically analyzed, however, and in section 3 some results concerning the problems raised earlier in this paper are reported. It is shown that many motion verbs of the traffic domain are not 'observable', i.e. their application requires more than a geometrical scene description. One of the additional knowledge sources which must be provided is experience, i.e. some encoding of (abstractions of) past observations.

2. Obtaining a geometrical scene description

This section gives an overview of the first stage of NAOS. Its input is a TV-image sequence, and its output is the geometrical scene description, where objects are identified and 3D object positions are given at each instance of time. The main processes employed for this task can be conveniently visualized within the framework of a general vision system as depicted in Fig. 2. Several boxes representing data structures are shown in perspective. They are arranged vertically according to their

level of abstraction and in depth corresponding to the time slices which they describe. Some boxes extend along the time axis to indicate time-specific descriptions, others correspond to a single time slice only. In fact, the traditional framework for single-image interpretation (KANADE 80, NAGEL 79, HANSON and RISEMAN 78b) can be obtained from this diagram by separating out all boxes associated with a single time slice.

Boxes are connected by directed arrows corresponding to processes. Double arrows indicate that processing may take place in both directions. The order in which processes are carried out is not specified. As a matter of fact, the diagram lacks any representation of data structures or processes related to control strategy. It should be clear, however, that in general the bottom-up direction corresponds to hypothesis generation and the top-down direction to hypothesis verification.

The following is a description of processes currently used in NAOS. The numbers refer to the corresponding arrows in the diagram.

1: Change detection. The greyvalue characteristics of consecutive images are compared. Differences give rise to nonstationary image regions which may correspond to moving objects (NAGEL and REKERS 82).

2: Extraction of straight lines and corner points. Straight line segments are extracted using an algorithm described in NEUMANN 78. They are good descriptors for many stationary objects of the scene, e.g. houses, streets, light poles. Corner points are extracted to measure the exact displacement of nonstationary image components.

3: Interimage matching of corner points. Corner points of consecutive images are chained together if they depict the same physical surface point (DRESCHLER 81).

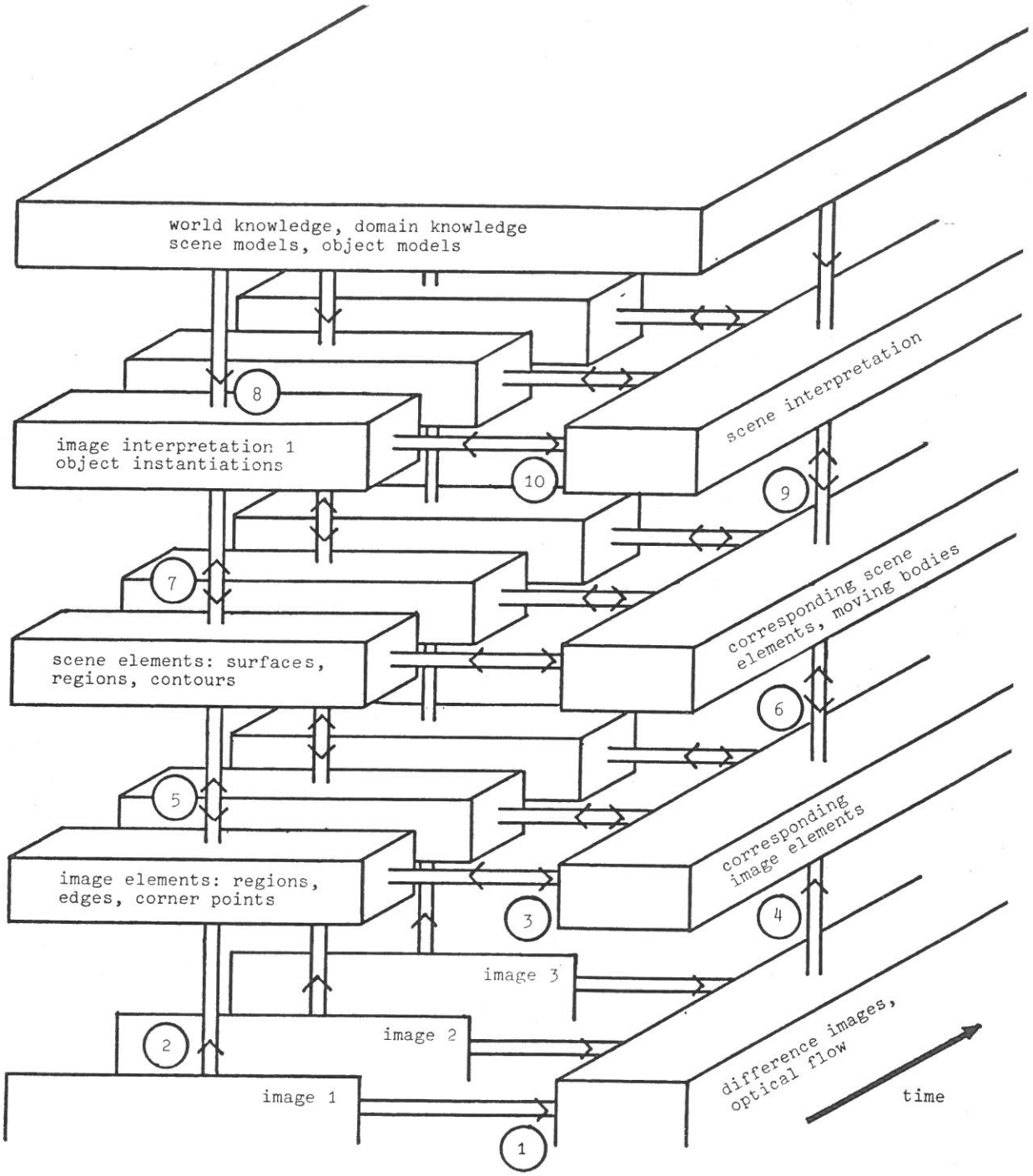

Figure 2: Basic components of a system for image sequence analysis

4: Chains are grouped according to connected nonstationary image regions.

5: Interpreting 2D lines as 3D lines. A 2D image line may be the projection of any 3D scene line assuming certain camera

parameters. Collinear 2D line segments may correspond to a single 3D line. More than two 2D lines intersecting in one point are assumed to be parallel in 3D.

6: Motion stereo. From the 2D displacements of corner points belonging to a nonstationary component one can compute their 3D structure and trajectory assuming a rigid configuration (BONDE and NAGEL 79).

7,8: Object identification. 3D lines are matched to corresponding descriptors of an extensive street model and the camera parameters, i.e. view point and direction from which the scene is observed, are determined. Major stationary image components are identified as named objects, e.g. 'Post',' Schlueterstrasse', 'Gehweg1'.

9: Moving objects like cars and pedestrians are identified and named by human interaction.

10: Scene interpretation. While in general this process may extract higher level motion concepts, it is a simple step in NAOS. A symbolic database is created partitioned along the time axis. The first time slice contains the instantiated models of all stationary objects, i.e. 3D-surface descriptions along with viewpoint information, as well as polyhedral representations of the moving objects visible at this moment. Further time slices describe changes with respect to the initial description, i.e. objects are only entered if they have changed location, appeared or disappeared. This database is called geometrical scene description and represents the output of the first stage of NAOS.

This concludes the overview of the scene analysis subsystem of NAOS. Most of the processes are the result of independent research in image sequence analysis at Hamburg University and I

gratefully acknowledge the friendly support of all people in the Cognitive Systems group.

3. Higher level motion concepts

The geometrical scene description represents a scene interpretation at a rather low level of abstraction in comparison with natural language concepts like 'ueberholen' ('overtake') or ausweichen ('give way to'), and it is not immediately clear how to compute the latter from the former. In fact, the verb 'ausweichen' seems to imply a certain intensionality which may not be computable from the geometrical scene description at all. On the other hand, there are examples like 'anhalten' ('stop') or beschleunigen ('accelerate') which may be readily evaluated from the data captured in the geometrical scene description.

As a first step towards gaining some understanding of these problems a simplified paradigm was adopted. The system simulates a human observer who may be asked decision questions about a scene, e.g.

'Ist ein Lastwagen von der Schlueterstrasse abgebogen?'
('Has a truck turned off Schlueterstrasse?')

Thus only top-down processes have to be devised which determine the applicability of a given natural language expression. On the other hand, additional complexities arise from the free use of natural language, e.g. time adverbials. They define an 'interval of consideration', i.e. a subsequence of images to which the natural language expression may apply. In bottom-up processing intervals of consideration may be selected according to other, possibly simpler criteria.

A system for answering decision questions was implemented using components of HAM-RPM (now HAM-ANS) and several newly devised modules which are described in MARBURGER et al. 81 in some

detail. Several simple motion concepts like 'anfahren' ('start'), 'anhalten' ('stop'), 'fahren' ('drive'), 'gehen' ('go'), 'abbiegen' ('turn off') are implemented using a case frame representation associated with a predicate to determine the applicability of the motion concept. The predicate is implemented as a procedure which can be directly applied to the geometrical scene description, restricted to the interval of consideration.

This system demonstrates the possibilities but also the limitations of obtaining natural language descriptions by linking a natural language system to a scene analysis system via verb expert procedures. On the positive side, it has been shown that scene analysis and natural language processing have progressed far enough to make natural language description of image sequences feasable. Furthermore, the data supplied by the geometrical scene description suffice to answer decision questions about certain motion concepts, and solutions are offered to achieve this. On the other hand, not all verbs can be processed in this framework, also the procedural representation of motion concepts may become unwieldy in complex cases.

Before extending a vocabulary it is useful to determine criteria for its completeness. This has been carried out recently for the domain of traffic scenes (NOVAK 82). Novak shows that natural language descriptions can be produced with an extremely limited vocabulary (such descriptions are essentially a natural language rendering of the geometrical scene description). On the other hand, there is an undeterminable number of verbs which may potentially be applied to a scene, particularly in metaphorical use. Many of these verbs express more than a description of trajectories, i.e. more than can be gathered from the geometrical scene description alone. This will be elaborated in the remainder of this section.

Let us consider the verb 'rasen' ('speed'). As a first try one might define 'rasen' as the motion of vehicles at a velocity

exceeding the local speed limit. Thus a representation could be chosen comparable to a motion frame in TSOTSOS 80 where domain knowledge (here: the local speed limit) is encoded via constraints on certain motion properties (here: velocity). But how about wheel chairs? Clearly, some other constraints would be in order, possibly derived from a 'typical-speed' slot of a wheel-chair frame. Consider now a parking area or a road with children playing ball. Regardless of speed limit and typical vehicle speed 'rasen' would apply if the speed exceeded constraints imposed by the locality. To do justice to all of these uses of 'rasen' some representation of typical scenes is required which does not only associate typical velocities with the respective agents but also captures other scene characteristics on which the typical velocity might depend.

As another example consider 'abbiegen' ('turn off'). In a simple definition 'abbiegen' would apply whenever the trajectory of a vehicle suddenly changed direction by a certain amount. This would also apply, however, to cars following a sharp turn of a road, hence the existence of a road continuing in the old direction should be postulated. Again this does not suffice since one should not use 'abbiegen' if the continuation is a side road and the through traffic turns. To remedy this, one is led to define 'abbiegen' with respect to the typical traffic flow pattern at this locality.

Several other verbs, e.g. 'begegnen' ('meet') or 'ankommen' ('arrive'), can also be shown to refer to typical aspects of a scene. Hence a representational scheme is emerging which allows to determine verb applicability by comparing the actual scene description with certain abstractions of experienced scenes. This is in perspicuous analogy to suggestions of Waltz regarding shape representation (WALTZ 79). He proposes to describe the shape of an unfamiliar object as a mapping with respect to the shape of well-known objects.

4. Conclusions

The task of computing natural language descriptions for real-world scenes has been addressed by characterizing the level of abstraction achieved by existing scene analysis processes as opposed to motion concepts used in natural language. To bridge this gap additional knowledge sources are required. It has been shown that an explicite encoding of experience may be the key to verb sense representation. Other kinds of knowledge which have not been addressed in this paper are also required, e.g. a communication model or hierarchical relations between concepts. Further work in NAOS is intended to lead to concrete representational structures for experience.

References

Abe et al. 81
A Plot Understanding System on Reference to Both Image and Language
N. Abe, I. Soga, and S. Tsuji
IJCAI-81, pp. 77-84

Badler 75
Temporal Scene Analysis: Conceptual Descriptions of Object Movements
N.I. Badler
Technical Report No. 80, Dept. Computer Science, University of Toronto
February 1975

Bonde and Nagel 79
Deriving a 3-D Description of a Moving Rigid Object from Monocular TV-Frame Sequences
T. Bonde and H.-H. Nagel
WCATVI-79, pp. 44-45

Dreschler 81
Ermittlung markanter Punkte auf den Bildern bewegter Objekte und Berechnung einer 3D-Beschreibung auf dieser Grundlage
L. Dreschler
Dissertation (Juni 1981), Fachbereich Informatik der Universitaet Hamburg

Dreschler and Nagel 81
Volumetric Model and 3D-Trajectory of a Moving Car Derived from Monocular TV-Frame Sequences of a Street Scene
L. Dreschler and H.-H. Nagel
IJCAI-81, pp. 692-697

Hanson and Riseman 78a
Computer Vision Systems

A.R. Hanson and E.M. Riseman (eds.)
Academic Press New York 1978

Hanson and Riseman 78b
VISIONS: a Computer System for Interpreting Scenes
A.R. Hanson and E.M. Riseman
in A.R. Hanson and E.M. Riseman (eds.): Computer Vision Systems, Academic Press New York 1978, pp. 303-333

Hanssmann 80
Sprachliche Bildinterpretation fuer ein Frage-Antwort-System
H.-J .Hanssmann
IfI-HH-M-74/80, Fachbereich Informatik, Universitaet Hamburg, 1980

Kanade 80b
Region Segmentation: Signal vs Semantics
T. Kanade
Computer Graphics and Image Processing 13 (1980) 279-297

Marburger et al. 81
Natural Language Inquiries about Motion in an Automatically Analyzed Traffic Scene
H. Marburger, B. Neumann, and H.J. Novak
in J. Siekmann (ed.): GWAI-81, Informatik-Fachberichte 47, Springer Berlin 1981, pp. 79-87

Miller 72
English Verbs of Motion: A Case Study in Semantics and Lexical Memory
G. Miller
in A.W. Melton and E. Martin (eds.): Coding Prowesses in Human Memory, Winston and Sons, Washington/DC 1972, 335-372

Nagel 77
Analysing Sequences of TV-Frames: System Design Considerations
H.-H. Nagel
IJCAI-77, p. 626, IfI-HH-B-33/77 (March 1977) Fachbereich Informatik, Universitaet Hamburg

Nagel 79
Ueber die Repraesentation von Wissen zur Auswertung von Bildern
H.-H. Nagel
in J.P. Foith (ed.): Angewandte Szenenanalyse, Informatik Fachberichte 20, Springer Verlag, Berlin-Heidelberg-New York 1979, 3-21

Nagel and Rekers 82
Moving Object Masks Based on an Improved Likelihood Test
H.-H. Nagel and G. Rekers
ICPR-82, to be published

Neumann 78
Interpretation of Imperfect Object Contours for Identification and Tracking
B. Neumann
IJCPR-78 Nov. 7-10, 1978 Kyoto/Japan, pp. 691-693

Novak 82
On the Selection of Verbs for Natural Language Description of Traffic Scenes
H.-J. Novak
GWAI-82, to be published

Okada 80a
Conceptual Taxonomy of Japanese Verbs for Understanding Natural Language and Picture Patterns
N. Okada
Proc. COLING-80, 127-135

Okada 80b
Conceptual Taxonomy of Japanese Verbs and Sentence Production from Picture Pattern Sequences
N. Okada
Information Science and Systems Engineering, Oita University, Japan (December 1980)

Schefe and Pretschner 81
SWYSS - A Natural Language Question-Answering System for Scene Analysis
P. Schefe, B. Pretschner
GWAI-81, Informatik Fachberichte 47, Springer 81, 69-78

Tsotsos 80
A Framework for Visual Motion Understanding
J.K. Tsotsos
TR CSRG-114, University of Toronto, 1980

Tsotsos et al. 80
A Framework for Visual Motion Understanding
J.K. Tsotsos, J. Mylopoulos, H.D. Covvey, S.W. Zucker
IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980) 563-573

Waltz 79
Relating Images, Concepts, and Words
D.L. Waltz
Proc. NSF Workshop on the Representation of Three-Dimensional Objects, R. Bajcsy (ed.), Philadelphia/PA, May 1-2, 1979

Waltz 81
Toward a Detailed Model of Processing for Language Describing the Physical World
D.L. Waltz
IJCAI-81, (1981) 1-6

Winograd 72
Understanding Natural Language
T. Winograd
Academic Press, New York 1972

v.Hahn et al. 80
The Anatomy of the Natural Language Dialogue System HAM-RPM
W. v.Hahn, W. Hoeppner, A. Jameson, W. Wahlster
in L.Bolc (ed.): Natural Language Based Computer Systems
Muenchen, Hanser/McMillan 1980, 119-253