Natural-Language Oriented Event Models For Image Sequence Interpretation: The Issues

by

Bernd Neumann and
Hans-Joachim Novak

CSRG Technical Note #34
May, 1983

Abstract

Language is a natural medium for communicating the results of high-level image sequence interpretation. This report investigates events as conceptual units mediating between vision and natural language. Events are similar to motion concepts used in earlier approaches to image sequence interpretation, except that they are linked to the deep case structure of corresponding natural language utterances. Several representational and procedural requirements are discussed, which follow from the communication situation. The requirements are compared with the framework for motion understanding developed by Tsotsos.

NATURAL-LANGUAGE ORIENTED EVENT MODELS FOR IMAGE SEQUENCE INTERPRETATION: THE ISSUES

Bernd Neumann* and Hans-Joachim Novak*

Department of Computer Science
University of Toronto, Canada

1. *Introduction*

This report is concerned with high-level interpretation of image sequences containing motion. By this we mean the recognition of concepts which go beyond the traditional paradigm of object recognition (location, shape description, classification and identification of individual objects in a still-frame), in particular we are concerned with the recognition of temporal concepts like object motion. Clearly, an interpretation of a scene with motion would be incomplete in all but exceptional cases, if object descriptions were given only for individual instances of time. We would like a vision system to interpret a time-varying scene in terms of meaningful concepts extending over time.

While the need for higher-level conceptual units appears to be indisputable, it is not at all clear what these should be. Beginning with basic concepts of change (e.g. shrink, translate) one could proceed to recognize composite motion patterns (e.g. swing, roll). In addition to general purpose concepts one might try to recognize domain specific motion, e.g. overtaking cars in a traffic scene, a cell dividing under the microscope, a heart beat anomaly in an X-ray sequence. Yet concepts like these seem to be only at the beginning of still higher levels of interpretation. If a vision system is ever to rival human capabilities of interpreting silent movies (which are no more than image sequences), it must be able to assign significance to motion in terms of causality, intentionality, social acceptability, etc. For example, if in a scene person A inserts a knife into person B, the interpretation should not be homomorphic to person C inserting a spoon into pudding. Similarly, if in a traffic scene a car yields to another car, the danger of collision should somehow be explicit. One might argue that high-level concepts like "yield" do not necessarily constitute a vision concept since much of the significance is not "observable" but derives from general knowledge about people, their typical behavior, etc. On the other hand, a vision system may very well be given the task of finding a "yield" situation much like finding a particular shape. Hence, when thinking about high-level vision it does not seem wise to restrict oneself to the more readily observable concepts like "overtake".

There have been only few approaches to high-level interpretation of image sequences so far. Badler [BADLER 75] investigated the recognition of motion concepts like "swing" and "bounce" as well as associated adverbials. As far as the introduction of conceptual units is concerned, his approach is noteworthy for the taxonomical representation of motion concepts (using a specialization hierarchy) as well as their orientation towards natural-language (using verbs of motion and adverbials). The work of Okada [OKADA 80] is similar in the respect that he also considers a conceptual de-

*visiting from the University of Hamburg, W. Germany

composition of motion verbs. His primitives, however, are not the basis of a taxonomy but constitute features for pattern recognition. Tsotsos refines and extends the work of Badler in several ways [TSOTSOS 81]. He creates a framework of general motion concepts (including e.g. lengthen, area-expand, rotate) and defines particular motions of his domain of interest (left ventricular heart motion) by using composition and constraints on more general motion types. Tsotsos is not concerned with verbalization. Most of his concepts, however, correspond to certain natural-language notions of change, some of which being less natural than others (e.g. "posterior-rapid-fill", which is a special leftventricular motion).

Finally, the work of Marburger and Novak [MARBURGER and NOVAK 81] gives an example of strictly natural-language oriented motion recognition. They developed a system which answers decision questions on simple motions in a traffic scene, e.g. "Did the yellow VW turn off Schlueterstreet into Hartungstreet?". Motion concepts are represented in terms of top-down recognition procedures attached to a verb's case frame. Furthermore, the deep cases extracted from the question (e.g. source and goal) can be used as procedure parameters.

This report describes work on project NAOS which has grown out of the decision-question scheme of Marburger and Novak, and aims at bottom-up verbalization as well as top-down question answering for the domain of traffic scenes. In spite of the commitment to natural-language communication, a major concern of NAOS is high-level vision. Scene data are provided in terms of an image sequence interpreted up to the level of object recognition. Hence for each instance of a traffic scene we assume complete knowledge about class, identity, 3D-shape and 3D-location of every object which might be talked about. This representation is termed "geometrical scene description" (GSD). The first step, then, in obtaining a natural-language description is the recognition of higher-level concepts from the GSD, which is a vision task. From this we instantiate case frames for simple natural-language utterances. Selecting from instantiated case frames we finally obtain a natural-language description of the scene. Question answering is viewed as a top-down process constraining the steps outlined above.

From this introductory discussion one may rightly conclude that the goals of NAOS are similar to previous work, in particular to the work of Tsotsos, while simultaneously introducing new elements in terms of a different domain (traffic scenes), a natural-language interface, and the necessity of both bottom-up and top-down processing. Opportunity and wisdom led us to study these similarities and differences in some depth while still in the design phase of NAOS. The results of this study are reported in the following sections.

First, we shall investigate the consequences of our natural-language commitment. "Event models" are introduced as conceptual units tailored around verbs for traffic scene description. Many event models are conceptually not much different from the motion concepts considered elsewhere, but some verbs necessitate new methods of representation. Event models must also identify components corresponding to the verb's deep cases. Questions to the system may provide constraints on the event type and on these components. Thus dynamic (i.e. context dependent) constraints are introduced into event models which, in turn, affect the control structure. Finally, several more subtle requirements of natural-language communication are shown to be also connected to control structure issues.

In section 3 we discuss possibilities of realizing NAOS using the framework of the ALVEN system developed by Tsotsos. In particular, we compare representational and control flow requirements. While ALVEN's representational formalism is a demonstrably elegant tool for encoding event models, major difficulties arise when applying ALVEN's interpretation strategy to the NAOS problem domain. Both, the differences in the respective domains and the requirement for a question-dependent control flow, let

ALVEN's interpretation strategy appear unsuited for NAOS.

2. *Talking About Events*

In the previous section we tried to provide some motivation for wondering about high-level conceptual units in a vision system. Several approaches have been sketched, ranging from taxonomically oriented motion concepts to more pragmatically oriented conceptual units. We now investigate the approach followed in NAOS, where the conceptual units - called event models in NAOS - are required to link up with a natural-language system. The idea is not new as has been shown above. The consequences of following this idea up, however, have not yet been analyzed in prior work.

Verb-centered event models

Given a natural-language utterance about an event, many characteristic features of the event are usually captured by the verb. In fact, most of the motion concepts mentioned so far have been referred to using verbs. Other components of the utterance, e.g. adverbials, modifiers, tense, agent and other case specifications, seem to play only subordinate roles, reflecting certain aspects of the event. Hence it seems useful to organize event models around verbs. Whenever a particular event model is instantiated, a natural-language utterance involving the corresponding verb describes this event. One may further differentiate event models according to combinations with adverbials, e.g. "backward swing", "forward swing", as in [BADLER 75]. In fact, many verb-adverb combinations in English translate into composite verbs in German. Once event models are required to represent verb meaning, their structure and properties are no longer free to choose. Given a certain verb vocabulary, the design of event models essentially follows from studying the semantics of these verbs. This has been done for the vocabulary used in NAOS, which includes some 70 verbs for describing locomotion in street traffic [NOVAK 82]. Several observations can be made.

(i) Verbs describe ***interval events*** and ***point events.***

Interval events extend over a definite period of time. They can be thought of as beginning at a certain instance and ending at a later instance. Examples are "walk", "overtake", "cross". It is useful to distinguish the subclass of ***durative*** interval events (similar to the linguistic notion in [MILLER/JOHNSON-LAIRD 75]). An event (over a certain interval) is durative if any subinterval is also such an event. "Walk", "accelerate", "stand" are all durative. Durativity can be exploited during the recognition process, where alternate instantiations of interval boundaries may be required.

Point events (elsewhere called "instantaneous" events) usually mark the transition between two interval events. For example, "reach x" happens when "approach x" ends and "be at x" begins. Although it has been suggested to model point events as extending over some small time interval, we feel that conceptually only a time point is involved and event models should be structured accordingly.

(ii) There may not be verbs for conceptually interesting events.

Natural language has not developed to be easily embedded in a conceptual taxonomy as used, for example, by Tsotsos. When trying to relate verb oriented event models to each other using the is-a and part-of relationship, one frequently stumbles upon concepts which are taxonomically useful but do not correspond to any particular verb.

Consider the is-a parent of "walk" and "drive", which might be called "locomove". "Move" would be too general. Many other verbs involve "locomove" as either part-of or is-a parent, e.g. "overtake", "stop", "speed", etc., hence "locomove" is a useful conceptual unit. Similarly, many geometrical concepts like "rotate" or "translate" do not belong to the natural vocabulary for scene description. In NAOS we take a hybrid approach by having both, verb-oriented and "conceptual" event models.

(iii) Some verbs depend not only on scene data.

There are several verbs used in every-day language for traffic scene description, which express more than can actually be observed in the scene. We shall first give some examples and then suggest an approach to deal with verbs of this kind. Consider the verb "speed". It is natural to model speeding events as locomotions with a velocity exceeding a certain threshold. Unfortunately there does not seem to be an easy way of specifying this threshold. It depends on the agent (pedestrian, car, wheelchair) as well as many environmental factors (local speed limits, type of road, daytime, traffic density, etc.). Clearly, in order to recognize a "speed" event one has to judge "the appropriateness" of the agent's velocity Another example is "continue walking" (which translates into a single verb in German). One of its meanings denotes an uninterrupted walk where stopping had been expected, e.g. "he continued walking in spite of the red light". To recognize such an event, one obviously has to generate expectations about the development of the scene. Finally, consider "avoid" or "yield". These verbs involve judgement of intentions based on necessarily indirect evidence. For example, if car 1 stops at an intersection, then car 2 passes in front, and then car 1 continues, it is likely to be a "yield". If after stopping the driver gets out to mail a letter, it probably is not.

The examples demonstrate that the event models of some verbs require interaction with general knowledge which can not be justly considered part of the verb semantics. Such event models can only be instantiated after certain common-sense reasoning on the scene data has been performed. Thus a departure from the traditional paradigm of pattern matching and constraint satisfaction seems to be indicated. In NAOS we plan to use an approximative construction where ***expectations*** about the development of the scene are generated as a substitute for reasoning. Expectations can be provided without resorting to a large body of general knowledge.

Verb case frame instantiation

The deep case frame of a verb is a useful interface to a natural-language system. Given an instantiated case frame, simple utterances can be produced. Conversely, given an utterance, a case frame can be derived and used to constrain the search for events in a scene [MARBURGER et al. 81]. How can event models be linked to the corresponding case frames? As it turns out, verb-oriented event models can be formulated in such a way that certain token variables correspond to the deep cases of the respective verb. For example, the object of a scene for which a "walk" event can be instantiated will obviously also fill the agent slot of the "walk" case frame. Similarly, the object's location at the beginning of the event is the source case. (A verbalization of the - abstract - source case will usually require a nearby object which specifies this location. This is not part of the event model). Also, start and end time of an interval event or the instance of time of a point event play a role quite analogous to the other deep cases. The corresponding token variables of the event model can be directly used for the selection of tense and time adverbials. In the following example we present a simplified version of the event model for "walk" together with an associated case frame to illustrate these points. All numbered identifiers denote token variables which must be instantiated. The @-operator (introduced by Tsotsos) selects a time

point or time interval for evaluation.

```
(EVENT-MODEL E-WALK
      (PARAMETERS OBJ1 TIME1 TIME2)

      (KERNEL
        (INSTANCE OBJ1 PEDESTRIAN)
        (LOCOMOVE OBJ1)@(TIME1 TIME2))

      (C-FRAME C-WALK))


(CASE-FRAME C-WALK

      (CASES
        (VERB WALK)
        (AGENT OBJ1)
        (START TIME1)
        (END TIME2)
        (SOURCE OBJ1@TIME1)
        (GOAL OBJ1@TIME2)
        (PATH OBJ1@TIME3
          (WITHIN TIME3 (TIME1 TIME2)))
        (LOCATIVE OBJ1@(TIME1 TIME2)))

      (E-MODEL E-WALK))
```

Question answering

While in the case of bottom-up scene description the process of verbalization can be separated from event recognition, this is not desirable for question-answering. Questions may provide strong additional constraints as to which events should be recognized. Consider

"Did the yellow VW stop in Schlueterstreet?" (1)

Clearly, we are interested only in (a) stopping-events of (b) certain agents and (c) in a certain location. The constraints (b) and (c) issued by the parser are of the same nature as those in the event models, except that they are dynamic and combine with an event model to define a ***dynamic event model.*** The dynamic quality of event models and as such of high-level vision models has not become apparent in prior work, which emphasized bottom-up recognition of certain fixed motion concepts.

Apart of representational requirements dynamic event models call for a flexible control structure of the recognition process. Consider the following variation of the above example:

"Did a car stop in front of the post office?" (2)

The order of evaluating the constraints arising from this question should be quite different from the former example, since possible stopping events are narrowed down most effectively by the agent specification "the yellow VW" in question (1) and by the locative "the post office" in question (2). Preliminary investigations indicate that a promising (possibly suboptimal) order of evaluation can indeed be determined by a sufficiently simple algorithm.

The communication task

Verbalization and question answering are usually part of a communication situation and as such governed by certain conventions, some of which also affect the underlying event recognition task. One such convention concerns the negation of decision questions. It is often desirable to provide more information than a simple "no". For example, if question (1) must be negated, several extended answers are conceivable (all assuming the same scene):

"No, the yellow VW drove through Schlueterstreet" (3)
"No, the yellow VW stopped in Hartungstreet" (4)
"No, the black Mercedes stopped in Schlueterstreet" (5)

Any one of these answers is generated by relaxing one or more of the constraints which caused the failing. It is clearly the task of the recognition process to instantiate these relaxed event models. It can also be observed that the proper answer - in this case probably (4) - is easily achieved if constraint evaluation proceeds in a particular order. Fortunately (and curiously) these order requirements seem to conform with those minimizing computational cost.

There are several more aspects to the communication situation which affect event recognition. For one, the desired level of ***detail*** of a scene description bears on the type of events which one should consider in the recognition phase. It is not yet clear, however, how "detailledness" of an event model could be determined. Similarly, dialogue ***focus*** restricts or favours event recognition of certain parts or certain aspects of a scene. Also, ***expectations*** raised in the course of the dialogue can cause instantiation of expectation-dependent events discussed earlier, e.g. "continue walking". These are, of course, subtleties compared with the central task of recognizing high-level vision concepts. We believe, however, that awareness of these subtleties can improve system architecture even if simplified goals have to be pursued for some time.

3. *Comparison With ALVEN*

In the previous section we have analyzed the task of NAOS: verbalizing events and answering questions for the domain of traffic scenes. Several problems and requirements arising from the natural language aspect have been discussed, this being the major difference to earlier work in high-level vision. We now turn to Tsotsos' system ALVEN which may be considered the most advanced system for high-level motion recognition existing today, and examine to which extent the formalisms and the framework of ALVEN can be used ir NAOS.

Overview of ALVEN

ALVEN is an expert vision system for left ventricular heart motion [TSOTSOS 80, TSOTSOS 81, TSOTSOS et al. 81]. Its input are X-ray image sequences which display heart motion in terms of displacements of markers implanted into a patient's heart wall. ALVEN encompasses a framework for general motion recognition and a knowledge base containing models for normal heart motion as well as several distinct anomalities. Models are embedded in a part-of and is-a hierarchy and hence are ultimately based on motion primitives which can be directly instantiated from the data. There are also exception links connecting models which differ with respect to a certain property.

The interpretation process is essentially bottom-up except of a certain amount of top-down guidance provided by expectations about marker positions. Hypothesis genera-

tion makes heavy use of the taxonomical relationships between motion concepts. Given evidence for a certain hypothesis A, various hypotheses conceptually close to A are also examined: specializations of A along is-a links, hypotheses which contain A as a part, hypotheses which follow in temporal order, and hypotheses which are accessed via exceptions in case A fails. Certainty values determine doom or eventually success of hypotheses.

We now compare ALVEN's representation and control flow with the requirements of NAOS in some more detail.

Representational Issues

ALVEN's motion concepts are represented using a frame notation based on PSN [LEVESQUE and MYLOPOULOS 78]. PSN is a general tool for knowledge representation, hence it is not surprising that the conceptual entities of NAOS can be represented using the same formalism. NAOS requires some frame types, however, which do not occur in this form in ALVEN. The following is a synopsis of the respective types of both systems (ignoring is-a specializations):

ALVEN	*NAOS*
PHYSICAL OBJECT (2D-centroids of markers)	PHYSICAL OBJECT (3D position and orientation, color and 3D shape)
MOTION (conceptual motion frame, constraints on object trajectories, taxonomical relations)	MOTION (same as left)
	INTERVAL EVENT (MOTION plus link to CASE-FRAME, common-sense interface)
	POINT-EVENT (sequence of two MOTIONs, otherwise same as INTERVAL-EVENT)
	CASE-FRAME (deep case structure of natural language utterance, link to EVENT frame)

The coexistence of motion concepts (MOTION) and event models (INTERVAL-EVENT and POINT-EVENT) in NAOS has been explained in section 2: motion concepts are taxonomically motivated conceptual units while event models owe their existence to corresponding verbs.

In ALVEN motion frames actually express more than constraints to be satisfied for instantiation. They also contain information directing the flow of control along various paths of the hierarchies. This aspect of the representational formalism cannot be carried over to NAOS, however. Commonalities end where matters of control flow begin.

Control Structure

ALVEN's control structure, i.e. the sequencing of computations to obtain an interpretation of the data, has characteristics unlike other major interpretation systems, e.g. HEARSAY-II or ACRONYM. The most prominent feature is hypothesis activation along is-a, part-of, temporal-next, and exception links between motion frames. This appears to be a very reasonable way for propagating and accumulating evidence in the absence of top-down constraints. It is not adapted, however, to highly restricted tasks like question answering.

Another characteristic is the strict progression along the time axis. At any instance of time, hypotheses are based only on data up to this time. Furthermore, future evidence cannot activate hypotheses retroactively. Hence a hypothesis pertaining to motion over a certain interval must be activated right at the beginning of the interval. The main advantages of this strategy seem to be, first, that backtracking can be completely avoided, second, that expectations can be generated in a unified way, and third, that the strategy works well for ALVEN's problem domain.

A look at the problem domain of NAOS, however, forecasts difficulties with ALVEN's "progressive" interpretation strategy. For many events in a traffic scene evidence may be very weak for a long period following the beginning of the event interval. For example, if a moving car is observed, a very large number of events may yet develop, e.g. "stop", "drive", "overtake", "turn-off", "arrive", etc. Evidence to the contrary may become available only very late: if "stop" ever fails, it fails at the end of the sequence. We do not suggest to hold back on hypothesis activation - Tsotsos' scheme seems to be very reasonable in this respect. Faced with many weak hypotheses, however, one should try to disprove them earlier. This can only be achieved by examining discriminating constraints regardless of the temporal order, thus abandoning progressive interpretation. For example, if an "overtake" hypothesis has been activated, the following constraints should be evaluated:

(LOCOMOVE OBJ1) @(TIME1 TIME2)
(LOCOMOVE OBJ2) @(TIME1 TIME2)
(BEHIND OBJ1 OBJ2) @TIME1
(BEHIND OBJ2 OBJ3) @TIME2
(WITHIN (TIME3 TIME4) (TIME1 TIME2))
(BESIDE OBJ1 OBJ2) @(TIME3 TIME4)
(APPROACH OBJ1 OBJ2Z) @(TIME1 TIME3)
(RECEDE OBJ1 OBJ2) @(TIME4 TIME2)

Start time and end time of the overtake event (as well as other time variables) are governed by linear inequalities arising from the constraints. They can be effectively determined using linear programming methods [MALIK and BINFORD 82].

A flexible order of evaluation for constraints within a hypothesis is certainly necessary if the control flow requirements discussed in section 2 are to be met. Due to the dynamic nature of constraints arising from questions there is no single best order which could be fixed once and for all. Note also the order requirements connected with answer generation for decision questions.

4. *Conclusions*

Language is a natural medium for communicating high-level concepts recognized in an image sequence. To this end representational tools and control strategies have to be developed which meet the demands of both, high-level vision and natural language communication. We have approached this task from the vision end, extending earlier

work in motion recognition. The main question of section 2 was: What requirements are placed by natural-language communication on representation and control in high-level vision? We found that verb-oriented concepts (called event models) are useful conceptual units which easily map into case frames of a natural-language system and also tend to cover the conceptual motion space investigated in earlier high-level vision work. Unfortunately, however, due to the unruliness of natural language not all useful motion concepts correspond to a verb and not all verbs correspond to observable motion. The former can be remedied by introducing non-verbal motion concepts along with verb-oriented event models. The latter calls for common-sense reasoning as part of the recognition process.

In the case of question answering the conceptual search space may be highly restricted. Questions can be viewed as adding dynamic constraints to those implied by the event models. As an important consequence, the control flow should be adaptable to the task at hand. Several aspects of the communication situation have also been shown to affect recognition control.

Comparing these requirements with the motion interpretation framework developed by Tsotsos for a task in the medical domain, we found that his representational formalism can be easily extended to include event models and case frames along with motion concepts. Also, his approach to hypothesis activation based on conceptual proximity seems to be a generally valid technique. Major changes would be necessary, however, when adapting to the control flow requirements arising from question answering. Tsotsos' progressive interpretation strategy does not leave room for dynamically ordered constraint evaluation.

5. *Acknowledgements*

We would like to thank the Department of Computer Science at the University of Toronto for inviting us and providing us with an excellent research environment. Special thanks, however, go to John Tsotsos who arranged the visit and helped to make it an equally pleasant and rewarding stay. We also gratefully acknowledge discussions with John Mylopoulos and Ray Perrault. Finally, we would like to thank the people of the vision group for their hospitality and timely help.

REFERENCES

[BADLER 75]
Temporal Scene Analysis: Conceptual Descriptions of Object Movements
N.I. Badler
Report TR 80, Department of Computer Science, University of Toronto, Toronto, Canada 1975

[LEVESQUE and MYLOPOULOS 78]
A Procedural Semantics for Semantic Networks
H. Levesque and J. Mylopoulos
in N.V. Findler (ed.), Associative Networks, Academic Press, 1978

[MALIK and BINFORD 82]
Representation of Time and Sequences of Events
J. Malik and T.O. Binford
in Proc. of a Workshop on Image Understanding, Palo Alto, CA, September 15-16, 1982

[MARBURGER and NOVAK 81]
Auswertung von natuerlichsprachlichen Entscheidungsfragen ueber Bewegungen in einer Strassenszene: Entwurf und Implementierungsansaetze
H. Marburger and H.-J. Novak
Diplomarbeit (Dezember 1981), Fachbereich Informatik der Universitaet Hamburg

[MARBURGER et al. 81]
Natural Language Dialogue about Moving Objects in an Automatically Analyzed Traffic Scene
H. Marburger, B. Neumann, H.-J. Novak
IJCAI-81, 49-51

[MILLER/JOHNSON-LAIRD 75]
Language and Perception
G.A. Miller and P.N. Johnson-Laird
Cambridge University Press, Cambridge-London-Melbourne 1975

[NOVAK 82]
On the Selection of Verbs for Natural Language Description of Traffic Scenes
H.-J. Novak
in: W. Wahlster (ed.), GWAI-82, Informatik Fachberichte 58, Springer 1982, 22-31

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

[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 and S.W. Zucker
IEEE Trans. Pattern Analysis and Machine Intelligence PAMI-2 (1980), 563-573

[TSOTSOS 81]
On Classifying Time-Varying Events
J.K. Tsotsos
IEEE PRIP-81, 193-199