EVENT MODELS FOR RECOGNITION AND NATURAL LANGUAGE DESCRIPTION OF EVENTS IN REAL-HORLD IMAGE SEQUENCES

Bernd Neumann and Hans-Joachim Novak

Fachbereich Informatik, Schluterstrafle 70
0-2000 Hamburg 13, W. Germany

ABSTRACT

For an adequate interpretation of image sequences it is not only necessary to recognize objects and object positions but also certain interesting temporal developments of the scene, called events. In this paper we discuss event models for traffic scenes as high-level conceptual structures which permit interfacing to an existing natural language dialogue system. Event models are declarative descriptions of classes of events organized around verbs of locomotion. They involve components which are directly related to the deep case structure of a corresponding natural language description. Event models may be used for bottom-up scene description as well as top-down question-answering. They may also incorporate expectations about a scene, thus providing an interface to experience and common sense.

1. Introduction

This paper deals with high-level image-sequence interpretation, i.e. with obtaining a meaningful description of time-varying visual data at a high level of abstraction. While in the traditional single-image paradigm of computer vision an interpretation in terms of object names and locations may be considered adequate, this is clearly not the case for an image sequence. Image sequences have much the same semantic potential as silent movies. Hence computer vision should ultimately be able to extract results comparable to human understanding of silent movies. While the silent-movie paradigm does not define "high-level interpretation", it suggests conceptual structures which are similar, if not partially identical, to meaning representation of natural language. As Miller and Johnson-Laird [1] put it: "Percepts and words are merely avenues into and out of this conceptual structure."

In this contribution we propose conceptual structures for a limited domain: motion in a traffic scene. Traffic scenes have been used in Hamburg for a long time both for vision and natural language research. On the vision side techniques have been developed to isolate moving objects and obtain 3D-shape descriptions [2]; the natural language group implemented a powerful dialogue system for simulated static scenes [3]. Recently efforts were started to connect both systems, with the ultimate goal of obtaining natural language descriptions of real-world image sequences [4].

Our approach to motion representation and recognition is strongly related to the pioneering work of Badler [5] and the more recent work of Tsotsos [6]. Both use a hierarchy of motion concepts to build a high-level description from low-level information. Motion primitives such as 'location-change' can be immediately retrieved from basic scene analysis data. Badler also generates simple verbal descriptions based on instantiated motion concepts corresponding to English verbs of motion and directional adverbials.

In section 2 a representational scheme will be presented which extends this work. We introduce "event models" which specify certain interesting subspaces of four- dimensional space-time continuum (much in accord with Webster's definition). Event models support bottom-up scene description as well as question answering.

In section 3 we discuss issues of evaluating event models.

2. Event models

In the preceding section an event has been loosely introduced as an interesting occurrence in a time-varying scene within a certain time interval and within certain spatial boundaries. In this section we shall define event models as a representation of classes of events and as a tool to recognize events in a given scene. Event models will be organized around verbs of change - currently restricted to verbs of locomotion - such that each event model describes events whose natural language description involves a particular verb.

It is convenient to sidestep scene analysis issues by assuming a certain standard representation of the analyzed scene. We use a "geometrical scene description" in terms of symbolic object names, object types, and 3D shape and position for equally spaced time slices (see [7] for details).

Event models specify predicates about the geometrical scene description in a relational form, e.g.

(ISA 08J1 VEHICLE)
(BEHIND 08J1 0BJ2)3TIME1

The first component of a tupel specifies the relation, all other components are either constants (VEHICLE) or unbound token variables (0BJ1, 0BJ2, TIMED. The A)-operator - introduced in [6] - denotes evaluation at a time instance or for all instances in an interval. Time

is considered discrete. Relations may be primitive (directly computable or retrievable from the scene data) or defined in terms of other relations.

A first example of an event model is given below. It specifies predicates which must be true for an event to be verbalized using "move".

```
(EVENT-MODEL E-MOVE
  (PARAMETERS OBJ1 TIME1 TIME2)
  (KERNEL
    (IN-MOTION OBJ1)A)(TIME1 TIME2))
  (C-FRAME C-M0VE1 C-M0VE2 ...))
```

An event model consists of a name and parameter section similar to a procedure declaration. The kernel specifies predicates with respect to the geometrical scene description. IN-MOTION is considered a primitive predicate. The event model is linked to possible case frames listed in the C-FRAMES statement. A case frame connects variables of an event model with the deep cases of a natural-language utterance. C-M0VE1 involves a source and a goal case, other case frames may involve other case combinations.

```
(CASE-FRAME C-M0VE1
  (CASES
    (VERB MOVE)
    (AGENT OBJ1)
    (START TIMED
    (END TIME2)
    (SOURCE OBJ1ATIME1)
    (GOAL OBJ1ATIME2))
  (E-MODEL E-MOVE))
```

In the case of question answering the parser uses the variables specified in a case frame to generate additional constraints supplementing the event kernel such that any instantiation of the event model is a positive instance for the question at hand. From the question

'Did the red lorry move from the drive-way
to the parking lot?"

we get the constraints

```
(ISA OBJ1 LORRY)
(COLOR OBJ1 RED)
(WITHIN (TIME1 TIME2) (PAST-TIME1 PAST-TIME2))
(ISA SOURCE 1 DRIVEWAY)
(MARK S0URCE1 0BJ1S)TIME1)
(IS A G0AL1 PARKING-LOT)
(MARK G0AL1 0BJ1ATIME2)
```

Conversely, if bottom-up scene description is required, case frames specify from which (instantiated) variables of an event verbalizations are to be generated. Many of the involved mechanisms, e.g. the generation of time bounds or MARK for relating two locations are available from previous work [8,4], but they must be remodelled to permit both bottom-up and top-down processing.

Event models may be partially defined by other event models, hence a hierarchy is induced. Model 1 is said to be a specialization of model2, if its kernel contains the kernel of model2 (after uni- fication of variables) or, in a logical interpretation, if it implies mode!2. In the following example it is assumed that APPROACH and RECEDE have already been defined, possibly using MOVE.

```
(EVENT-MODEL OVERTAKE
  (PARAMETERS OBJ1 OBJ2 TIME1 TIME2)
  (KERNEL
    (MOVE OBJ1)@(TIME1 TIME2)
    (MOVE OBJ2)@(TIME1 TIME2)
    (BEHIND OBJ1 OBJ2)@TIME1
    (BEHIND OBJ2 OBJ1)@TIME2
    (WITHIN (TIME3 TIME4) (TIME1 TIME2))
    (BESIDE OBJ1 OBJ2)@(TIME3 TIME4)
    (APPROACH OBJ1 OBJ2)@(TIME1 TIME3)
    (RECEDE OBJ1 OBJ2)@(TIME4 TIME2))
  (C-FRAMES ...))
```

As stated earlier many verbs of locomotion cannot be used without certain knowledge about what is typical and what is not. Consider the German verb "weitergehen" (appro*, "to carry on walking"). On the surface it may describe an uninterrupted walking event, but clearly it implies that halting had been expected. Similarly, 'turn off" is used when an object continues its path along a way which deviates from the natural extension of its previous path. Expectations may be specified in event models using the following notation:

(EXPECT <expectations> IN-VIEW-OF <premises>)

The EXPECT construct relates two partial scene descriptions, where the description termed <premises> provides the scene data on which the expectations are to be based. Usually the whole scene up to a certain time or except certain components constitutes the premises. In those cases it is convenient to use a shorthand THIS-SCENE BEFORE <time> or THIS-SCENE WITHOUT <objects>. In the following example we define the event model CARRY-0N-WALK1 using the EXPECT construct.

```
(EVENT-MODEL CARRY-ON-WALK
  (PARAMETERS OBJ1 TIME1 TIME2)
  (KERNEL
    (WALK OBJ1)@(TIME1 TIME2)
    (WITHIN TIME3 (TIME1 TIME2))
    (EXPECT (STAND OBJ1)@(TIME3 TIME2)
      IN-VIEW-OF THIS-SCENE BEFORE TIME3))
  (C-FRAMES ...))
```

The EXPECT notation is a first step towards interfacing event models with several knowledge sources other than scene data, prominently experience about past events and common sense. In our system we plan to generate expectations by matching the premises with encodings of typical events.

3. Event recognition

An event model may be viewed as a template which must match pertinent scene data, verbal information and background knowledge for an event to be true. From the examples it should be plausible that predicates are involved where the truth value can indeed be determined given certain arguments and the above mentioned knowledge sources. Some of the predicates are very simple (e.g. IN-MOTION), others are less so (e.g. EXPECT). Very little has been said, however, about the process of instantiating a complete event model, i.e. event recognition. In this section we shall discuss some features of the control structure. An

implementation is currently being prepared.

The general framework is a backtracking control regime induced by alternative predicate instantiations. Time variables, however, play a special part in event recognition and are treated differently. Note that many predicates involving time as well as durative events like MOVE are also satisfied for subintervals if they are satisfied for some interval at all. For these cases it is convenient to keep track of possible values using linear inequalities. Following the idea in [9], we can then apply linear programming methods to determine satisfiability each time a new constraint is added. Solutions may be obtained satisfying the inequalities and maximizing some criterion function, e.g. "the most recent instance"' or "the longest interval".

For an effective search it is necessary to control the order of alternative instantiations, the order of predicate evaluation and the order of invoking event models. A useful criterion for controlling instantiations is temporal and spatial proximity to the current focus of attention.

The order in which the predicates of an event model should be evaluated depends decisively on the constraints imposed by the natural language interface. Decision questions may propose specific events (e.g. "Did a yellow car turn off Schlueterstreet?") where event model instantiation is constrained by agent and source specif ications. On the other hand, event models must also be instantiated in a *free* verbalization context (e.g. "What happened?"). It is one of the advantages of the strictly declarative definition of event models that they can be employed for all cases ranging from recognition to verification. For an effective evaluation predicates may be dynamically ordered according to their degree of instantiation such that the branching rate can be kept low.

Finally, for bottom-up scene description one has to hypothesize one out of many event models. We presently consider ca. 70 verbs of locomotion (listed in [10]) and a corresponding number of event models. They are organized into a specialization hierarchy as discussed in the previous section. This makes it possible to proceed from general event hypotheses (e.g. MOVE) to increasingly special event hypotheses (e.g. CROSS) similar to the control structure used in [5], for that matter, in other hierarchical systems.

4. Conclusions

We have proposed a representation for event models which is designed to meet four main objectives.

(i) Event models provide a precise yet readable definition of a class of events. This is achieved using a relational notation with a clear logical interpretation.

(ii) Event models support event recognition in the framework of computer vision. A scene is assumed to be analyzed up to the level of object recognition before event recognition begins. Although the state of the art does not yet permit automatic object recognition in real-life traffic scenes, this *seems* to be an attainable goal.

(iii) Event models are linked to a corresponding natural language description. Although their conceptual structure is oriented towards visual data, they support both question-answering and free verbalization.

(iv) Event models may refer to expectations and hence interface to contextual knowledge, experience and common sense. This allows to model events in terms of deviations from the expected - a feature whose importance has long been recognized [1 1].

The evaluation strategy has been outlined. An implementation is underway.

References

[1] G.A. Miller and P.N. Johnson-Laird, Language and Perception. Cambridge University Press, Cambridge-London-Melbourne 1976

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

[3] W.v. Hahn, W. Hoeppner, A. Jameson. W. Wahlster, The Anatomy of the Natural Language System HAM-RPM. In: L. Bole (ed.), Natural Language Based Computer Systems. Munchen, Hanser-McMillan 1980,1 19-253

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

[5] N.I. Badler, Temporal Scene Analysis. Conceptual Descriptions of Object Movements. Report TR-80, Dept. of CS, University of Toronto, Toronto 1975

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

[7] B. Neumann, Towards Natural Language Description of Real-World Image Sequences. GI-12. Jahrestagung,Informatik Fachberichte 57, Springer 1982, 349-358

[8] W. Wahlster, A. Jameson, W. Hoeppner, Glancing, Referring and Explaining in the Dialogue-System HAM-RPM. AJCL 1978, Michrofiche 77, 53-67

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

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

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