American Journal of Computational Linguistics *Microfiche 35 : 70*

The Conceptual Description of Physical Activities

Norman Badler

Department of Computer and Information Science
The Moore School of Electrical Engineering
University of Pennsylvania
Philadelphia 19174

ABSTRACT

A system has been designed to translate connected sequences of visual images of physical activities into conceptual descriptions. The representation of such activities is based on a canonical verb of motion so that the conceptual description will be compatible with semantic networks in natural language understanding systems. A case structure is described which is derived from the kinds of information obtainable in image data. A possible solution is presented to the problem of segmenting the temporal information stream into linguistically and physically meaningful events. An example is given for a simple scenario, showing part of the derivation of the lowest level events. The results of applying certain condensations to these events show how details can be systematically eliminated to produce simpler, more general, and hence shorter, descriptions.

This research was primarily supported by Canadian Defense Research Board grant 9820-11, and partially by National Science Foundation grant ENG75-10535.

If we view a motion picture such as illustrated in Figure 1, we are able to give a description of the physical activities in the scenario. This description is linguistic in the sense that the words used express our recognition of objects and movements as conceptual entities. A system for performing a sizeable part of this transformation of visual data into conceptual descriptions has been designed. It is described in Badler (1975); here we will present one small part of the system which is concerned with the organization of abstracted data from successive images of the scenario.

We are interested in a possible solution to the following problem: Given that a conceptual description of a scenario is to be generated, how is it decided where one verb instance starts and another ends? In other words, we seek computational criteria which separate visual experience into discrete "chunks" or events. By organizing the representation of an event into a case structure for a canonical motion verb, events can be described in linguistic terms. Verbs of motion have been investigated directly or indirectly by Miller (1972), Hendrix et al.(1973a,1973b), Martin (1973), and Schank (1973); semantic databases using variants of case structure verb representations (Fillmore(1968)) include Winograd (1972), Rumelhart et al (1972), and Simmons (1973).

We are concerned with physical movements of rigid or jointed objects so that motions may be restricted to translations and rotations. Objects may appear or disappear and the observer is free to move about. The resulting activities are combinations of these where observer motions are factored out if at all possible. We assume that the scenarios contain recognizable objects exhibiting physically possible, and preferably natural, motions.

A particular activity might consist of a single event, a sequence of events, sets of event sequences, or hierarchic organizations of events. The concept of "walking" is a good example of the last. Events are the basic building blocks of the conceptual description, and our events indicate the motions of objects. The interpretation of motion in terms of causal relationships is generally

Figure 1. The moving car scenario

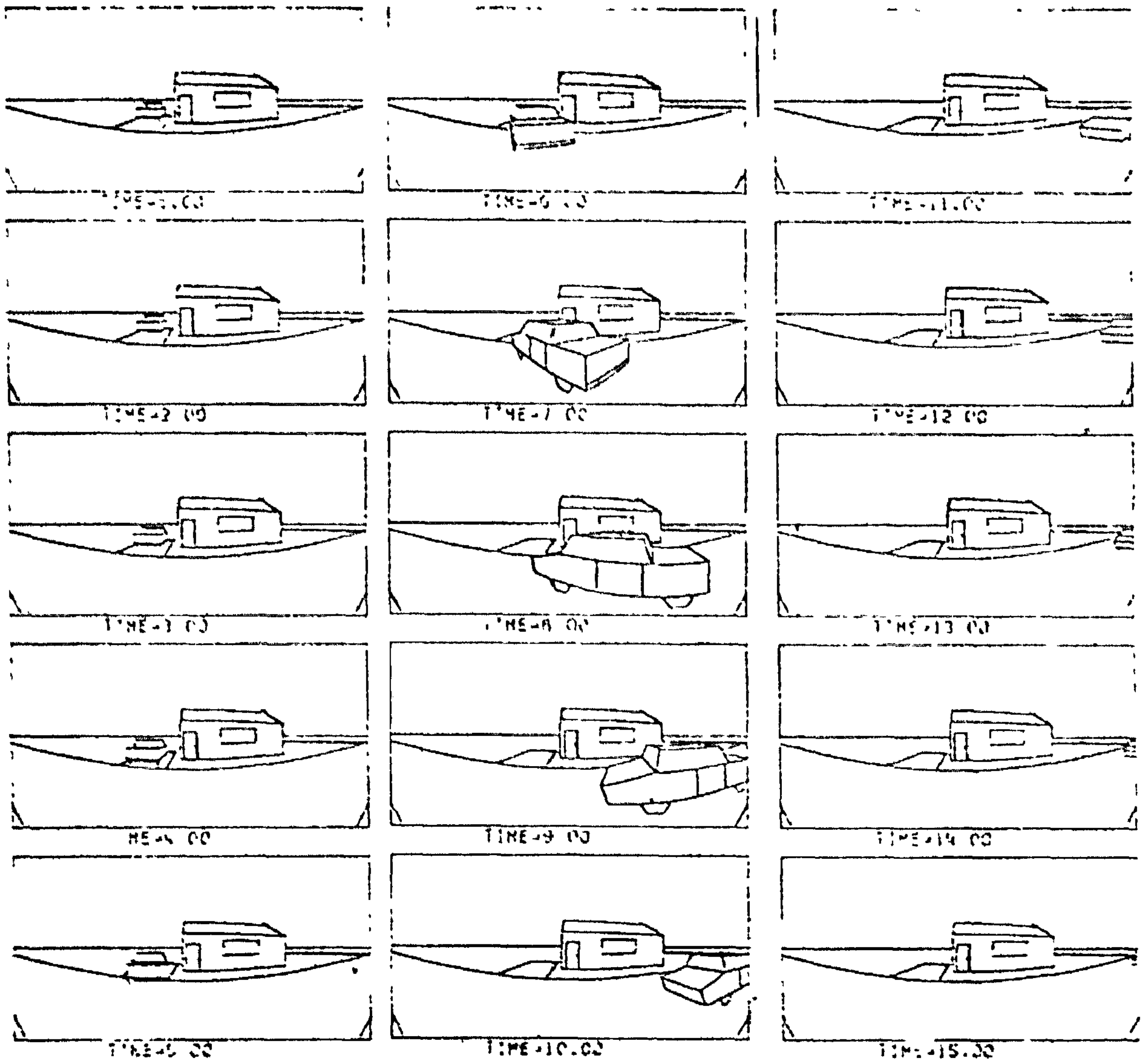

Table 1

Adverbials

Type	Relationships	Set of Concepts
1	between the orientation and trajectory or axis of an object	BACKWARD, FORWARD, SIDEWAYS AROUND, OVER,CLOCKWISE, COUNTERCLOCKWISE
2	between the trajectory of an object and fixed world directions -	DOWN(WARD),UP(WARD),NORTHWARD SOUTHWARD,EASTWARD,WESTWARD
3	changing between objects.	ACROSS,AGAINST,ALONG,APART, AROUND,AWAY,AWAY-FROM, BEHIND,BY,FROM,IN,INTO,OFF, OFF-OF,ON,ONTO,OUT,OUT-OF, OVER,THROUGH,TO,TOGETHER, UNDER
4	indicative of source and target	AWAY-FROM,IN-THE-DIRECTION-OF, IN(WARD),OUT(WARD),TOWARD
5	between the path of an object and other (moving) objects	AFTER, AHEAD-OF,ALONG,APART TOGETHER,WITH
6	between an event and a previous event	BACK-AND-FORTH,TO-AND-FRO, UP-AND-DOWN,BACK,THROUGH

beyond the scope of the current system, although a semantic inference component could be included. Our descriptions consist mostly of observation of motion in context rather than explanation of why motion occurred.

The general descriptive methodology is to keep only one static relational description of the scenario, that of the current image. Changes between it and the next sequential image are described by storing the names of changes in event nodes in a semantic network. In general, names of changes correspond to adverbs or prepositions (adverbials) describing directions or changing static relationships. Computational definitions for the set of adverbials in Table 1 appear in Badler (1975). We are only concerned with the senses of the adverbials pertaining to movement. Definitions are implemented as demons: procedures which are activated, the executed, by the successive appearance of certain assertions in the image description or current conceptual database. These demons are related to those of Charniak (1972), although our use of them, their numbers, and their organization are simplified and restricted. They are used to recognize or classify properties or changes and to generate the hierarchic descriptive structure. An essential feature of this methodology is that the descriptions are continually condensed by this change abstraction process; descriptions grow in depth rather than length.

The semantic information stored for each object in the scenario includes its TYPE, structural SUB-PARTs, VISIBILITY, MOBILITY, LOCATION ORIENTATION, and SIZE. Most of these properties are determined from the image sequence, but some are stored in object models (indexed by TYPE) in the semantic network.

The events are also nodes in the semantic network. Each object is potentially the SUBJECT of an event node. A sequence of event nodes forms a history of movement of an object; only the latest node in the sequence is active. The set of active event nodes describes the current events in the scenario seen so far. The cases of the event node along with their approximate definitions follow.

SUBJECT: An object which is exhibiting movement.
AGENT: A motile object which contacts the SUBJECT.
INSTRUMENT: A moving object which contacts the SUBJECT.
REFERENCE: A pair of object features (on a fixed object) which are used to fix absolute directions independent of the observer's position.
DIRECTION: A temporally-ordered list of adverbials and their associated objects which apply to this SUBJECT.
TRAJECTORY: The spatial direction of a location change of the SUBJECT.
VELOCITY: The approximate magnitude of the velocity of the SUBJECT along the TRAJECTORY; it includes a RATES list containing STARTS, STOPS and (optionally) INCREASES or DECREASES.
AXIS: The spatial direction of an axis of an orientation change (rotation) of the SUBJECT.
ANGULAR-VELOCITY: Similar to VELOCITY, except for rotation about the AXIS.
NEXT: The temporal successor event node having the same SUBJECT.
START-TIME: The time of the onset of the event.
END-TIME: The time of the termination of the event.
REPEAT-PATH: A list of event nodes which form a repeating sequence.

These cases differ from Miller's (1972) primarily in the lack of a "permissive" case and our separation of the TRAJECTORY and AXIS cases. REFERENCE is new; one of its uses is to resolve descriptions of the same event from different viewpoints. The explicit times could be replaced by temporal relations. Miller's reflexive/objective distinction is not needed as each moving object has its own event nodes, regardless of the AGENT.

A few necessary definitions follow before the presentation of the event generation algorithm.

A null event node has all its cases NIL or zero except START-TIME, END-TIME, and perhaps NEXT.

An event node is terminated when it has a non-NIL NEXT value.

The function CREATE-EVENT-NODE (property pairs) creates an event node with the indicated case values, returning the node as a result.

To compare successive values of numerical properties , a queue is associated with the case in current event nodes only. The front of the queue is represented by "*": the place where new information is stored. The queues have length three; the three positions will be referenced by prefixing the case name with either "NEW", "CURRENT", or "LAST". A function SHIFT manipulates property queues when they require updating:

```
LAST-property: = CURRENT-property;
CURRENT-property: = NEW-property;
NEW-property: = *
```

The time will be abbreviated by TN and TL. For a particular event node E:

```
TN: = NEW-END-TIME (E);
TC: = CURRENT-END-TIME (E);
```

Thus TN is always equal to the present image time.

Now we can present the algorithm for the demon which controls the construction of the entire event graph. It is executed once for each image when all lower level demons have finished; it creates,terminates, or updates each current event node.

A. 1. Creating event nodes.

A. 1. 1. An event node E is created when a mobile object first becomes visible and identifiable as an object.

```
E: = CREATE-EVENT-NODE((SUBJECT object-node)
                       (VELOCITY(* 0. 0.))
                       (ANGULAR-VELOCITY (* 0. 0.))
                       (START-TIME NIL)
                       (END-TIME (* TN TN))  ).
```

The NIL START-TIME has the interpretation that we do not know what was happening to this object prior to time TN.

A. 1. 2. An event node E is created when a jointed part of the parent object with current event node EP is first observed to move relative to the parent, for example, an arm relative to a person's body.

```
TC: = CURRENT-END-TIME(EP);
E: = CREATE-EVENT-NODE( (SUBJECT object-part-node)
                        (AGENT parent-object-node)
                        ( INSTRUMENT joint-node)
                        (REFERENCE ...)
                        ( DIRECTION  ...)
                        (TRAJECTORY ...)
                        (VELOCITY ...)
                        (AXIS ...)
                        ( ANGULAR-VELOCITY ...)
                        (START-TIME TC)
                        (END-TIME (TN TC TC)) ).
```

This is interpreted as the parent object moving the part using the joint as the "instrument". Any appropriate attributes are placed in the NEW-property positions. The node E is then immediately terminated (A.1.3).

A.1.3. An event node E2 is created whenever another event node E1 is terminated.

```
TC: = CURRENT-END-TIME(E1);
NEXT(E1): = CREATE-EVENT-NODE(
                (SUBJECT...)
                (AGENT...)
                (INSTRUMENT...)
                (REFERENCE...)
                (DIRECTION...)
                (TRAJECTORY SHIFT(TRAJECTORY(E1)))
                (VELOCITY SHIFT(VELOCITY(E1)))
                (AXIS SHIFT(AXIS(E1)))
                (ANGULAR-VELOCITY SHIFT(ANGULAR-
                        VELOCITY(E1)))
                (START-TIME TC)
                (END-TIME SHIFT(END-TIME(E1)));
E2: = NEXT(E1).
```

SUBJECT, AGENT, INSTRUMENT, REFERENCE, and DIRECTION are those which were present at termination of the previous node, subject to any additional conditions that changes in these may require.

A.2. Terminating event nodes. An event node E is terminated when there are significant changes in its properties. All queue structures are deleted.

```
END-TIME(E): = CURRENT-END-TIME(E);
TRAJECTORY(E): = CURRENT-TRAJECTORY(E);
AXIS(E): = CURRENT-AXIS(E);
VELOCITY(E): = (CURRENT-VELOCITY(E) RATES(VELOCITY(E)));
ANGULAR-VELOCITY(E): = (CURRENT-ANGULAR-VELOCITY(E)
                        RATES(ANGULAR-VELOCITY(E))).
```

The DIRECTION list is unaltered except that the terminating adverbial (s) may be added to DIRECTION(E) rather than to DIRECTION(NEXT(E)) (see A.2.5.).

A.2.1. Changes in SUBJECT. The assumptions of object rigidity and permanence preclude changes in an object.

A.2.2/3. Changes in AGENT and INSTRUMENT. These must be preceded by changes in CONTACT relations between objects and the SUBJECT. See A.2.5 on DIRECTION.

A.2.4. Changes in REFERENCE. A change in the REFERENCE features forces termination of every event node referencing those features, as such changes are usually caused by spatial or temporal discontinuities in the scenario.

A.2.5. Changes in DIRECTION.

Changes in type (1) adverbials must be preceded by changes in TRAJECTORY, VELOCITY, AXIS, or ANGULAR-VELOCITY, because a relationship between an orientation and a TRAJECTORY or AXIS cannot change without at least one of the four cases changing. Changes in BACKWARD, FORWARD, and SIDEWAYS cause termination; this may occur with no orientation change if the TRAJECTORY has a non-zero derivative. For example, move a box in a circle while keeping its orientation constant.

Changes in type (2) adverbials must be preceded by a change in TRAJECTORY, but some of these changes may be too slight to cause termination from the TRAJECTORY criteria. (A.2.6.). Changes from UP to DOWN or vice versa are the only ones in this group causing termination.

Changes in type (3) adverbials terminate event nodes if and only if there is a change in a CONTACT relation or a VISIBILITY property. If the CONTACT is made or the VISIBILITY established, the adverbial goes into the new node's DIRECTION list. If the CONTACT is broken or VISIBILITY lost, the adverbial remains on the front of the terminated node's DIRECTION list.

Since the type (4) adverbials are only indicators of current source and target, these do not change unless the path of the SUBJECT changes or the target object moves. Therefore no terminations arise from this group.

The type (5) adverbials relate paths of the SUBJECT to other objects. They cause termination when they come into effect, and terminate their own nodes when they cease to describe the path.

The type (6) adverbials include higher level events and the basic repetitions. These all terminate the current event node. The repeated events (for example, BACK-AND-FORTH) are terminated when the repetition appears to cease.

A.2.6. Changes in TRAJECTORY. The changes in TRAJECTORY that are mos t important are those which change its derivative significantly. A change in the derivative from or to zero can be used (the start or end of a turn), but only the start is actually used for termination. Once the turn is begun, how it ends is unimportant since the final (current) trajectory is always saved.

The other termination case watches for a momentarily large derivative which settles back to smaller values. This indicates a probable collision. It is of crucial importance in inferring CONTACT relations between objects when none were (or could be) directly observed.

A.2.7. Changes in VELOCITY. A change in VELOCITY from zero to a positive value (from a positive value to zero) terminates the current event node and enters STARTS (STOPS) in the new node's (old node's) VELOCITY RATES list.

A.2.8. Changes in AXIS. A reversal of rotation terminates the event node. This corresponds to a change in AXIS to the opposite direction, with no intermediate values.

A.2.9. Changes in ANGULAR-VELOCITY. A change in ANGULAR-VELOCITY from zero to a positive value (from a positive value to zero) terminate the current event node and enters STARTS (STOPS) in the new node s (old node's) ANGULAR-VELOCITY RATES list.

A.2.10. Changes in NEXT are not meaningful.

A.2.11/12. Changes in START-TIME and END-TIME are not meaningful.

A.2.13. Changes in REPEAT-PATH. When new data fails to match the appropriate sub-event node of a REPEAT-PATH event node E, E is terminated. The definition of "match" for the basic repetitions appears in Badler (1975). The problem, in general, remains open. See, for example, Becker (1973).

A.3. Maintaining event nodes. If the new assertions do not cause termination of the event node, the property queues are merely shifted:

```
TRAJECTORY(E): = SHIFT(TRAJECTORY(E));
VELOCITY(E): = SHIFT(VELOCITY(E));
AXIS(E): = SHIFT(AXIS(E));
ANGULAR-VELOCITY(E) : = SHIFT(ANGULAR-VELOCITY(E));
END-TIME(E): = SHIFT(END-TIME(E)).
```

What does an event mean? This algorithm motivates a theorem that the events generated are the finest meaningful partition of the movements in the image sequence into distinct activities. The hypothesis of the assertion is the natural environment being observed and the linguistically-based conceptual description desired. The conclusion is that an event node produced from this algorithm describes either the lack of motion or else an unimpeded, simple linear or smoothly curving (or rotating) motion of the SUBJECT with no CONTACT changes. In addition, the orientation of the SUBJECT does not change much with respect to the trajectory. The proof of this assertion follows directly from the choice of termination conditions.

We will apply this algorithm to data obtained from each of the images in Figure 1. The lower front edge of the house is arbitrarily chosen as the REFERENCE feature; NORTH is toward the right of each image. We will not discuss the computation of the static relations from each image, only list in Table 2 the changes in the static description from image-to-image. Trajectory and rotation data are omitted for simplicity, although changes of significance are indicated.

If we "write out" the event node sequence using the canonical motion verbs MOVES and TURNS with the adverbial phrases from the RATES and DIRECTION lists, we obtain the following lengthy, but accurate, description:

C.1 There is a CAR.

C.2 The CAR STARTS MOVING TOWARD the OBSERVER and EASTWARD, then ONTO the ROAD.

C.3 The CAR, while GOING FORWARD, STARTS TURNING, MOVES TOWARD the OBSERVER and EASTWARD, then NORTHWARD-AND-EASTWARD, then FROM the DRIVEWAY and OUT-OF the DRIVEWAY, then OFF-OF the DRIVEWAY.

Table 2

Selected assertions and changes involved in the description of Figure 1.

Time	Action	Static Assertion	Event Assertion	Result

1	ADD
	ADD
	ADD
	ADD
	ADD
	ADD
	ADD
	ADD
	ADD
	ADD
		IN-FRONT-OF(CAR OBSERVER)
		IN-BACK-OF(CAR HOUSE)
		RIGHT-OF(CAR HOUSE)
		NEAR-TO(CAR HOUSE)
		SURROUNDED-BY(CAR DRIVEWAY)
		LEFT-OF(CAR DRIVEWAY)
		IN-BACK-OF(CAR DRIVEWAY)
		RIGHT-OF(CAR DRIVEWAY)
		AT(CAR DRIVEWAY)
		SUPPORTED-BY(CAR DRIVEWAY)
				create C1

3	DELETE	IN-BACK-OF(CAR HOUSE)	VELOCITY (STARTS)	terminate C1 (A.2.7.)
			EASTWARD	--
			TOWARD OBSERVER	--

5	DELETE
	ADD
	ADD
		IN-BACK-OF(CAR DRIVEWAY)
		SUPPORTED-BY(CAR ROAD)
		IN-FRONT-OF(CAR DRIVEWAY)
			TRAJECTORY change	terminate C2 (A.2.6.)

			ONTO ROAD	terminate C2 (A.2.5.)
			ANGULAR-VELOCITY (STARTS)	terminate C2 (A.2.9.)
6	ADD	IN-FRONT-OF(CAR HOUSE)	NORTHWARD-AND-EASTWARD	--

7	DELETE
	DELETE
	DELETE
	ADD
		LEFT-OF(CAR DRIVEWAY)
		SURROUNDED-BY(CAR DRIVEWAY)
		AT(CAR DRIVEWAY)
		NEAR-TO(CAR DRIVEWAY)
			OUT-OF DRIVEWAY	--

			FROM DRIVEWAY	--
			FORWARD	--
8	DELETE	SUPPORTED-BY(CAR DRIVEWAY)	OFF-OF DRIVEWAY	terminate C3 (A.2.5.)
9			NORTHWARD	--

10	DELETE
	ADD
	ADD
		NEAR-TO(CAR DRIVEWAY)
		LEFT-OF(CAR HOUSE)
		FAR-FROM(CAR DRIVEWAY)
			AROUND HOUSE	--

			AWAY-FROM DRIVEWAY	--

12	DELETE
	ADD
		NEAR-TO(CAR HOUSE)
		FAR-FROM(CAR HOUSE)
			AWAY-FROM HOUSE	--

			ANGULAR-VELOCITY (STOPS)	terminate C4 (A.2.9.)
15	DELETE	VISIBILITY(CAR VISIBLE)	AWAY	terminate C5 (A.2.5.)

Notes: Relations with HOUSE use the house front orientation, not the observer's front.

Termination of Ci creates Ci+1 by A.1.3.

C.4 The CAR, while GOING FORWARD, MOVES NORTHWARD-AND-EASTWARD, then NORTHWARD, then AROUND the HOUSE and AWAY-FROM the DRIVEWAY, then AWAY-FROM the HOUSE and STOPS TURNING.

C.5 The CAR, while GOING FORWARD, MOVES NORTHWARD, then AWAY.

The canonical form follows easily from the case representation and the DIRECTION list orderings. The directional adverbials FORWARD, BACKWARD and SIDEWAYS are interpreted as lasting the duration of the event, hence are written as "while GOING..." clauses. STARTS is always interpreted at the beginning of the sentence, STOPS at the end. The termination conditions assure its correctness.

There is much redundancy in this description, but it is only the lowest level, after all, and many activities span several events. Two sets of condensations are applied by demons that watch over terminated event nodes. The first set is mostly concerned with interpreting certain null events caused by the image sampling rate and removing trajectory changes which prove to be insignificant. The second set of demons removes adverbials referring to directions in the support plane, removes RATES terms except STOPS, and generalizes redundant adverbials referring to the same object. The result of applying these condensations is:

C.2 The CAR MOVES TOWARD the OBSERVER, then ONTO the ROAD.

C.3 The CAR, while GOING FORWARD, MOVES TOWARD the OBSERVER, then FROM the DRIVEWAY.

C.4 The CAR, while GOING FORWARD, MOVES AROUND the HOUSE and AWAY-FROM the DRIVEWAY, then AWAY-FROM the HOUSE, then STOPS TURNING.

C.5 The CAR, while GOING FORWARD, MOVES AWAY.

Another condensation can be applied for the sake of less redundant output. It does not, however, permanently affect the database:

> The CAR MOVES TOWARD the OBSERVER, then ONTO the ROAD, while GOING FORWARD, then FROM the DRIVEWAY, then AROUND the HOUSE, then AWAY-FROM the HOUSE, then STOPS TURNING, then MOVES AWAY.

Note that FROM the DRIVEWAY follows ONTO the ROAD. This is due to the pictorial configuration: the car is on the road before it leaves the driveway. The position of the "while GOING FORWARD" phrase could be shifted backwards in time to the beginning of the translatory motion, but this may be risky in general. We will leave it where it is, since this is primarily a higher level linguistic matter.

By applying demons which recognize instances of specific motion verbs to the individual event nodes, then condensing as above, we get:

> The CAR APPROACHES, then MOVES ONTO the ROAD, then LEAVES the DRIVEWAY, then TURNS AROUND the HOUSE, then DRIVES AWAY-FROM the HOUSE, then STOPS TURNING, then DRIVES AWAY.

The major awkwardness with this last description is that it relates the car to every other object in the scene. Normally one object or another would be the focus of attention and statements would be made regarding its role. Such manipulations of the descriptions are yet unclear.

In conclusion, we have outlined a small part of a system designed to translate sequences of images into linguistic semantic structures. Space permitted us only one example, but the method also yields descriptions for scenarios containing observer movement and jointed objects (such as walking persons). The availability of low level data has significantly shaped the definitions of the adverbials and motion verbs. Further work on these definitions, especially motion verbs, is anticipated. We expect that the integration of vision and language systems will benefit both domains by sharing in the specification of representational structures and description processes.

References

Badler, N. (1975). "Temporal scene analysis: Conceptual descriptions of object movements." University of Toronto, Department of Computer Science, Technical Report No. 80, February 1975.

Becker, J. (1973). "A model for the encoding of experiential information." In Computer Models of Thought and Language, Schank, R. and Colby, K. (eds.), W.H. Freeman & Co., San Francisco, 1973, pp. 396-434.

Charniak, E. (1972). "Toward a model of children's story comprehension." MIT Artificial Intelligence Report TR-266, December 1972.

Fillmore, C. (1968). "The case for case." In Universals in Linguistic Theory, Bach, E. and Harms, R. (eds.), Holt, Rinehart, and Winston, Inc., Chicago, 1968.

Hendrix, G. (1973a.). "Modeling simultaneous actions and continuous processes." Artificial Intelligence 4, Winter 1973, pp. 145-180.

Hendrix, G., Thompson, C. and Slocum, J. (1973b). "Language processing via canonical verbs and semantic models." Third International Joint Conference on Artificial Intelligence, August 1973, pp. 262-269.

Martin, W. (1973). "The things that really matter - A Theory of prepositions, semantic cases, and semantic type checking." Automatic Programming Group, Internal Memo 13, MIT Project MAC, 1973.

Miller, G. (1972). "English verbs of motion: A case study in semantics and lexical memory." In Coding Processes and Human Memory, Melton, A. and Martin, E. (eds.), V.H. Winston & Sons, Washington, D.C., 1973, pp. 335-372.

Rumelhart, D., Lindsay, P. and Norman D. (1972). "A process model for long term memory." In Organization of Memory, Tulving, E. and Donaldson, W. (eds.), Academic Press, New York, 1972, pp. 197-246.

Schank, R. (1973). "The fourteen primitive actions and their inferences." Stanford A.I. Laboratory Memo AIM-183, 1973.

Simmons, R. (1973). "Semantic networks: Their computation and use in understanding English sentences." In Computer Models of Thought and Language, Schank, R. and Colby, K. (eds.), W.H. Freeman & Co., San Francisco, 1973, pp. 63-113.

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