KNOWLEDGE REPRESENTATION FOR THE GENERATION OF QUANTIFIED NATURAL LANGUAGE DESCRIPTIONS OF VEHICLE TRAFFIC IN IMAGE SEQUENCES

Ralf Gerber[1] *and Hans-Hellmut Nagel*[1,2]

[1]Institut für Algorithmen und Kognitive Systeme
Fakultät für Informatik der Universität Karlsruhe (TH)
Postfach 6980, D-76128 Karlsruhe, Germany

[2]Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB),
Fraunhoferstr. 1, D-76131 Karlsruhe, Germany
Telephone +49 (721) 6091-210 (Fax -413), E-Mail hhn@iitb.fhg.de

ABSTRACT

Our image sequence interpretation process, which generates conceptual descriptions of the behaviour of vehicles in real-world traffic scenes, essentially treated only a single vehicle (the *agent*) so far. Simultaneous behaviours of other vehicles in the scene have been formulated only relative to the agent. The approach discussed in this contribution allows us to quantify occurrences and thus to generate more global conceptual descriptions of behaviour by using natural language quantifiers. The semantics of such quantified occurrences are represented by special logic structures. Natural language descriptions are derived from these internal knowledge representations.

INTRODUCTION

Although both Natural Language Processing and Computer Vision have been active research areas separately and despite the fact that a link between these areas has been continuously pursued for about 20 years now ([1]), results from other sources about linking these areas have become available only recently. [2] discuss the usefulness of formal knowledge representation and introduce some extensions to increase the applicability of terminological systems to image interpretation tasks. [3] uses special Lexical Conceptual Structures (LCS) to represent the outputs from visual systems to language. His considerations have not been applied on real image data. [4] present a system to generate natural language descriptions of the location of renal stones found in radiographs. These authors do not take any kind of temporal relationships of the objects in the scene into account, since their system is based on the evaluation of only a single frame. [5] use Bayesian Believe Networks and conceptual knowledge to interpret the behaviour of moving objects in a real-world traffic scene. Natural language descriptions are not derived from their knowledge representation structures. We want to generate natural language descriptions from recorded video sequences of real-world traffic scenes. For a satisfactory description it is not sufficient to inspect the behaviour of each moving object separately. Rather we have to determine the behaviour of ensembles of various objects. Such ensembles will be described by quantified occurrences, modelled by means of fuzzy sets. The resulting representations can be transformed into natural language descriptions.

QUANTIFIED OCCURRENCES

Our image evaluation system XTRACK (see [6], [7], [8] and [9]) determines occurrences that describe the behaviour of an examined road vehicle. Occurrences comprise information about the examined vehicle, the motion verb and the validation time. Occurrences can be categorized into four different classes: occurrences which refer only to the agent itself (*agent reference*), the ones which additionally refer to the road (*road reference*), to another moving object (*object reference*), or to a selected location (*location reference*). Occurrences associated with the agent reference category can in turn be categorized into five subclasses, so-called *velocity occurrences* (like *drive_slowly*), *direction occurrences* (such as *drive_forward* and *back_up*), *branching occurrences* (e.g. *turn_left*), *acceleration occurrences* (e.g. *brake, accelerate*) and *terminative occurrences* (*stop, drive_off*). Using this classification, we can

Figure 1: Real-world road traffic scene at the Durlacher–Tor in Karlsruhe. The numbers specify the internal reference of the detected objects.

derive the following three postulates:

(P1) A vehicle either moves or is standing.

(P2) Each moving object is related to one occurrence of each class at each instant of time.

(P3) At each instant, the state of an object is defined by at most one occurrence of each class.

Figure 1 shows one image of a traffic scene recorded at the Durlacher Tor in Karlsruhe. This image sequence consists of nearly one hundred half-frames, the scan rate amounts to fifty half-frames per second. For each half-frame time, agent reference occurrences were determined for each trajectory in the Durlacher–Tor–Sequence. Using the three postulates we can define transition (time) points between different occurrences of the same class which are overlapping since the occurences have been modelled by means of fuzzy sets. As an example, Figure 2 lists the generated agent reference occurrences

7	:	93	!	drive_at_regular_speed(obj_7).
7	:	93	!	drive_forward(obj_7).
7	:	18	!	drive_straight_ahead(obj_7).
19	:	93	!	turn_left(obj_7).

Figure 2: Generated agent reference occurrences of 'object_7' of the scene depicted in Figure 1.

of 'object_7' of the regarded scene. In order to derive quantified occurrences, we examine at first how many and which vehicles perform the same movements at the same time. These quantities are related to a total number of objects moving in the scene. This total number can be the sum of all vehicles ever localized in the visual field of the camera or the sum of moving objects localized only during a particular period. Alternatively, it could be any other subset of moving objects which relate to the regarded objects in a particular way (e.g. all objects moving on the same street). Numerical relations derived in this manner are associated with natural language quantifiers like *a few*, *many*, *some*, *nearly all*, *all*. Such associations are mediated by means of fuzzy sets. Finally, quantified occurrences are determined for each time instant and each agent reference occurrence. Each quantified occurrence description consists of its quantifier, its validation time, the motion verb, the names of the involved objects and a degree of confidence. In order to facilitate the generation of the natural language description, we select the quantifier which is valid for the longest continuous period.

SEMANTIC REPRESENTATION

The data structures, on which the generation of quantified occurrences is performed, are based on the Discourse Representation Theorie (DRT) developed by [10]. In their research, these authors define special logic structures (Discourse Representation Structures, DRSs), in order to represent the semantics of natural language, and a set of construction rules to automatically transform natural language sentences into DRSs. A DRS consists of a set of Discourse Referents (DR) and a set of DRS-conditions. There is a close affinity between DRSs and First Order Predicate Calculus. For details on DRT, see [10]. We use the DRT to transform the generated conceptual descriptions into logic structures representing their natural language semantics. This is possible because the generated occurrences are very similar to natural language. The agent can be conceived as the subject of a 'virtual' sentence, the motion verb represents the verb of this sentence and the validation time acts as an adverbial element of time. As an example, Figure 3 shows such a DRS for an agent-reference occurrence. Figure 4 shows a DRS for a quantified occurrence. The generation of DRSs representing quantified occurrences from a set of DRSs representing single agent occurrences is performed by a set of derivation rules. These deriva-

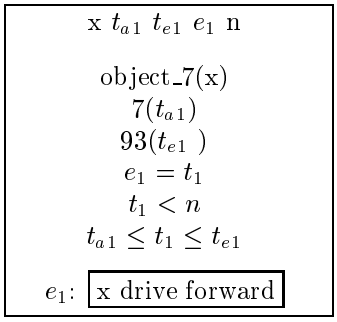

Figure 3: DRS representing the single agent occurrence *7 : 93 ! drive_forward(obj_7)*, which is conceived as the natural language assertion *Object_7 drove forward (from frametime 7 to frametime 93)*. The DR x is identified with the agent, the described event e_1 with the verb phrase. The DR n ('now') represents the *utterance time*. The event is located in the past ($t_1 < n$) in the specified time interval.

tion rules are based on the methods described in the preceding section. Due to limited space we are unable to desribe the used derivation rules in detail.

EXPERIMENTAL RESULTS

Using additional priority lists for quantified occurrences and introducing iterative adaptation of the derived quantified occurrences, we obtain the following sample description for vehicular traffic on the Durlacher–Tor image sequence:

Most vehicles turned left. A few vehicles turned right. Some vehicles drove straight ahead. All vehicles drove at regular speed. All vehicles drove forward.

Our method for the derivation of quantified occurrences has been successfully applied to two other video sequences of real-world road traffic scenes. Figure 5 shows one image of a real-world traffic scene recorded at the Nibelungenplatz in Frankfurt. Figure 6 lists the derived global natural language description.

CONCLUSION

In this contribution we introduced an approach to generate more global natural language descriptions of the behaviour of moving objects in real-world traffic scenes by using the Discourse Representation Theory. In order to generate quantified occurrences from single agent occurrences, we developped derivation rules based on behavioural knowledge defined in three postulates. Our approach has been successfully applied to three recorded video

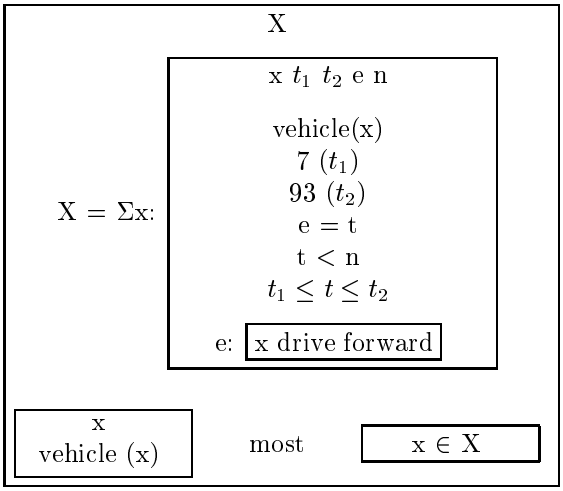

Figure 4: DRS for the quantified occurrence *Most vehicles drove forward (from frametime 7 to frametime 93)*. The first inner DRS is similar to the DRS shown in Figure 3. All Objects x according to the inner DRS are collected into a set X. The next DRS-condition represents the declaration that of all vehicles (left inner DRS) ever localized, the most are as well elements of X (right inner DRS).

Many vehicles drove straight ahead.
Some vehicles turned left.
Some vehicles turned right.
Most vehicles drove at regular speed.
A few vehicles drove slowly.
All vehicles drove forward.

Figure 6: Global natural language description of the behaviour of the objects moving in the Nibelungenplatz sequence (see Figure 5).

sequences of real-world traffic scenes. The natural language descriptions generated consist of rather simple constructed sentences including simple temporal relations (e.g. *first ..., then ...)* so far. Our approach can be improved with respect to several points. For example, when a human being is confronted with a statement like *some vehicles turned left*, he usually implies that all of these vehicles turn left from one and the same street into one and the same other street. So far, our approach does not perform this distinction. We perceive it as a challenge to separate behaviour of objects moving on different lanes by relating occurrences to the road segments on which they are performed. As well as generating natural language descriptions of quantified occurrences, our system is able to describe

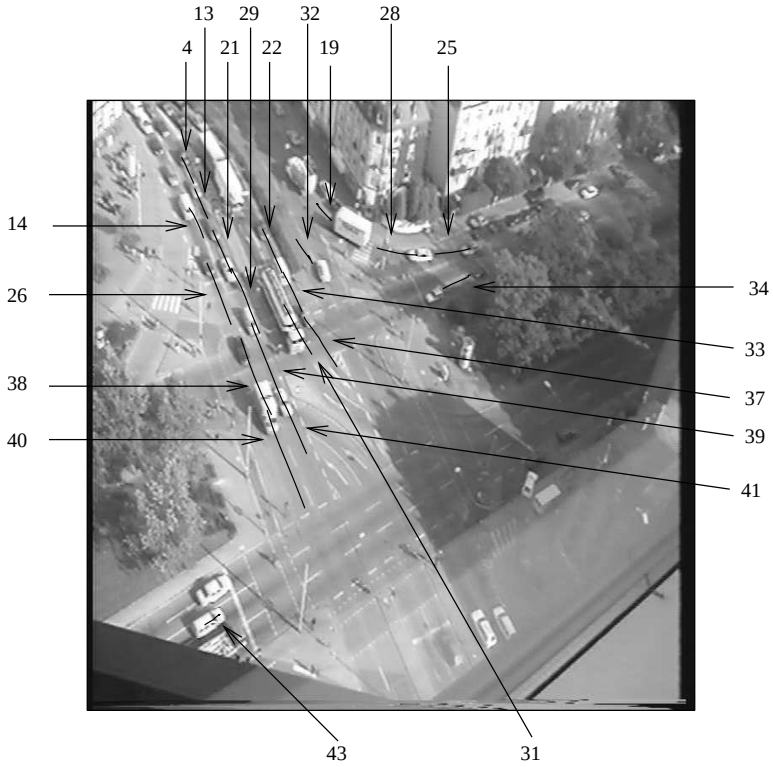

Figure 5: Real-world road traffic scene at the Nibelungenplatz in Frankfurt. The numbers specify the internal reference of the detected objects.

occlusions between stationary or moving objects in the scene ([11]).

References

[1] H.-H. Nagel: *Analysing Sequences of TV-Frames: System Design Considerations.* Proc. 5th Int. Joint Conf. on AI (IJCAI '77), August 1977, Cambridge/MA, p. 626

[2] B. Neumann, C. Schröder: *How useful is Formal Knowledge Representation for Image Interpretation?* Proc. Workshop on Conc. Descript. from Images, Cambridge/UK, 19 April 1996, pp. 58-69.

[3] L. Friedman: *From Images to Language.* Proc. Workshop on Conc. Descript. from Images, Cambridge/UK, 19 April 1996, pp. 70-81.

[4] A. Abella, J.R. Kender: *Description Generation of Abnormal Densities Found in Radiographs.* Proc. Workshop on Conc. Descript. from Images, Cambridge/UK, 19 April 1996, pp. 97-111.

[5] H. Buxton, S. Gong: *Visual Surveillance in a Dynamic and Uncertain World.* Artif. Intell. **78** (1995) pp. 431-459.

[6] D. Koller, K. Daniilidis, H.-H. Nagel: *Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes.* Int. Journal of Comp. Vis. 10 (1993) 257-281.

[7] H. Kollnig, H.-H. Nagel, M. Otte: *Association of Motion Verbs with Vehicle Movements Extracted from Dense Optical Flow Fields.* Proc. 3rd Europ. Conf. on Comp. Vis. (ECCV '94), Vol. II, Stockholm/S, 2-6 May 1994, Lect. Notes in Comp. Vis. **801**, Springer-Verl., Berlin a.o. 1994, pp. 338-347.

[8] H. Kollnig, H.-H. Nagel: *3D Pose Estimation by Fitting Image Gradients Directly to Polyhedral Models.* Proc. 5th Int. Conf. on Comp. Vis. ICCV '95, Cambridge/MA, 20-23 June 1995, pp. 569-574.

[9] H. Kollnig, H.-H. Nagel: *Matching Objects to Segments from an Optical Flow Field.* Proc. 4th Europ. Conf. on Comp. Vis. (ECCV '96), Vol. II, Cambridge/UK, 15-18 April 1996, Lect. Notes in Comp. Vis. **1065**, Springer-Verlag, Berlin a.o. 1996, pp. 388-399.

[10] H. Kamp, U. Reyle, *From Discourse to Logic.* Kluwer Academic Publishers, Dordrecht, Boston/MA, London 1993.

[11] T. Frank, M. Haag, H. Kollnig, H.-H. Nagel, *Tracking of Occluded Vehicles in Traffic Scenes.* Proc. 4th Europ. Conf. on Comp. Vis. (ECCV '96), Vol. II, Cambridge/UK, 15-18 April 1996, Lect. Notes in Comp. Vis. **1065**, Springer-Verlag, Berlin a.o. 1996, pp. 485-494.