(Mis?)-Using DRT for Generation of Natural Language Text from 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

Abstract. The abundance of geometric results from image sequence evaluation which is expected to shortly become available creates a new problem: how to present this material to a user without inundating him with unwanted details? A system design which attempts to cope not only with image sequence evaluation, but in addition with an increasing number of abstraction steps required for efficient presentation and inspection of results, appears to become necessary. The system-user interaction of a Computer Vision system should thus be designed as a natural language dialogue, assigned within the overall system at what we call the 'Natural Language Level'. Such a decision requires to construct a series of abstraction steps from geometric evaluation results to natural language text describing the contents of an image sequence. We suggest to use Discourse Representation Theory as developed by [14] in order to design the system-internal representation of knowledge and results at the Natural Language Level. A first implementation of this approach and results obtained applying it to image sequences recorded from real world traffic scenes are described.

1 Introduction

Creating a link between Computer Vision and Natural Language Processing becomes a research area of growing significance. This development – although unexpected at first sight – becomes more plausible once it is realized that methods for the automatic detection, initialisation, and tracking of moving objects in image sequences have matured significantly over the past years. As a consequence, results become available which are no longer restricted to tracking a single object through a short image sequence obtained from recording a controlled scene. It is now possible to evaluate extended image sequences recorded with a minimum of constraints from real world scenes, for example traffic scenes such as the one illustrated by Fig. 1. Results such as those illustrated by this figure will quickly exhaust the willingness and ability of researchers and users to scan through lists of data with scene coordinates of vehicles as a function of the frame number, i. e. of time. Even looking at the graphical representation of trajectories will soon loose its visual appeal which should not be underestimated

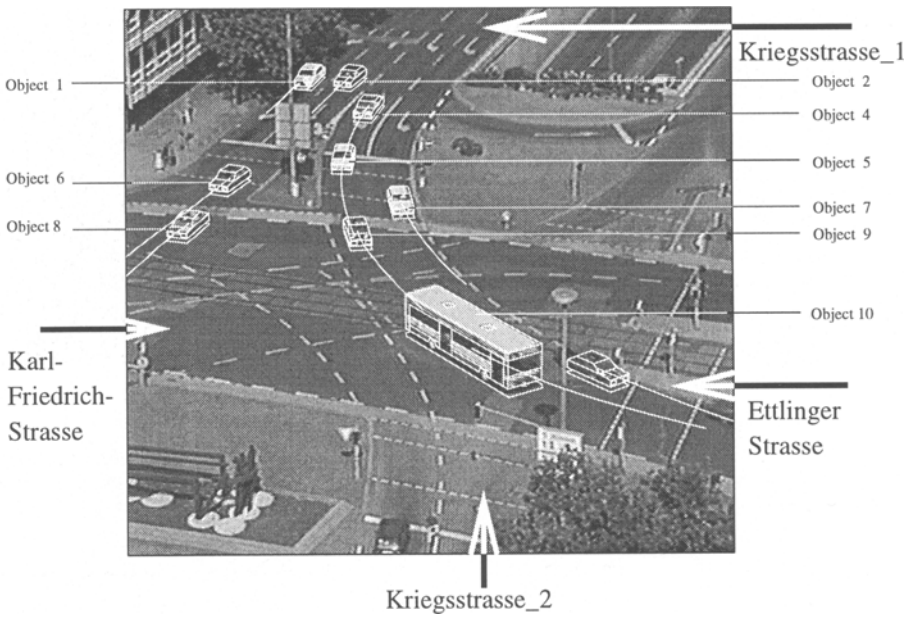


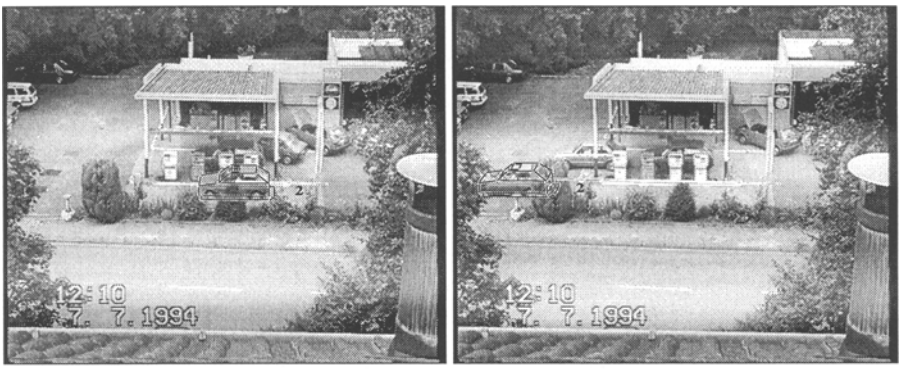
Fig. 1. Single frame of a real-world traffic scene recorded at an intersection. Vehicle trajectories obtained by automatic detection and model-based tracking are shown, together with the vehicle models used.

for a trained eye. Eventually, conceptual abstractions of relevant aspects covered by an image sequence will be desired, with the added complexity to facilitate specification of the desired aspect not at the time when the individual images of a sequence are processed, but at the time of inspection – possibly even depending on previous presentations of partial results from the same sequence.

We expect, therefore, that the geometric results from image sequence evaluation will be considered as a kind of data base which needs to be interrogated in the most flexible manner. Once one accepts such a scenario, it becomes obvious that the formulation of questions about a recorded scene in natural language terms and the expectation to obtain answers in the form of natural language text – at least as some kind of supplementary option – appear natural.

Our contribution should be understood as a step in this direction. We shall first outline our experimental approach in order to set the frame for a subsequent survey of the – not yet abundant – literature about links between computer vision and the generation of natural language text. Problems encountered in an attempt to generate natural language descriptions of different aspects captured about the temporal development within a scene by a video image sequence will then be presented in more detail.

A characteristic of our approach consists in that we do not develop ad-hoc solutions for the problems encountered, but rather investigate the possibility of adapting an established framework from computational linguistics. For this



(a) (b)

Fig. 2. Frame #700 (a) and #1320 (b) of a gas station sequence, including the automatically derived trajectory for the moving vehicle 'object_2'.

purpose, *Discourse Representation Theory (DRT)*, see [14], is of particular interest to us since it discusses algorithms for the transformation of coherent natural language text into a computer-internal representation, based – to the extent possible – on First Order Predicate Logic (FOPL). This formalism thus offers an immediate link to a representation which, on the one hand, incorporates logic for the manipulation of semantically relevant components of a description and, on the other hand, provides a framework which links logic-oriented representations to natural language text.

2 Outline of the System Structure of our Approach

Our image sequence *interpretation* process determines natural language descriptions of real-world vehicle traffic scenes. The structure of this interpretation process can be described as a Three-Layers-Architecture (see Fig. 3). The first layer comprises image evaluation steps like object detection and model-based vehicle tracking. Geometric results obtained in this *Signal Layer (SL)* comprise automatically generated trajectory data and the instantiation of various pieces of model knowledge underlying the model-based image evaluation steps. The first layer incorporates, too, the association of trajectory data with elementary motion verbs (see, e. g., [15]) and other conceptual primitives such as spatial or occlusion relations. These primitives provide the interface to the second layer, the so-called *Conceptual Layer (CL)*. The third layer, called *Natural Language Layer (NL)*, transforms results obtained in the CL into 'Discourse Representation Structures (DRS)' ([14]) which facilitate the derivation of natural language text.

In this contribution, we emphasize a discussion of the Natural Language Layer (NL). It will be shown that the generation of acceptable textual descriptions from a set of 'facts' obtained automatically by image sequence evaluation requires basic design decisions which justify to treat the NL as a separate compo-

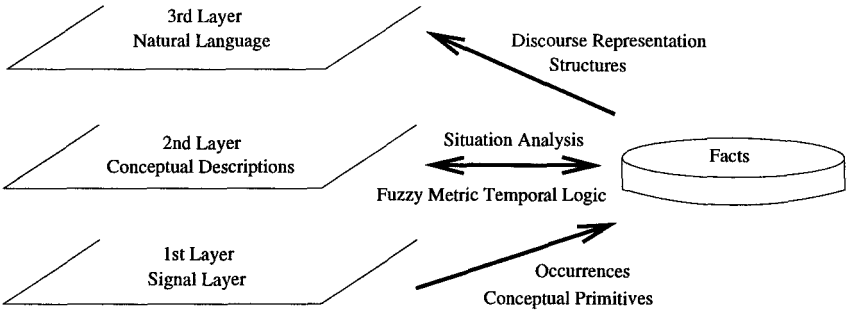


Fig. 3. The Three-Layers-Architecture of our Image Sequence Interpretation Process.

nent within an overall system concept. In doing so, we proceed on the assumption that sufficiently rich and reliable 'facts' about the temporal development within a traffic scene are provided by a robust SL, as described in [16] or [10]. The performance of the CL has been significantly improved in comparison with the state reported in [8, 9] by a transition from processing based on a procedural approach towards one based on 'Fuzzy Metric Temporal Logic' as described by [20].

This transition not only simplified a link to the logic-oriented representation of Discourse Representation Theory, but in addition facilitated to incorporate – in a most natural manner – a number of boundary conditions for natural language text generation from image sequence evaluation results. Initial results for such a logic-based intermediate conceptual representation have recently been published by [11].

3 Survey of Related Research

A comprehensive survey of advanced visual surveillance systems which interpret behavior of moving objects in image sequences can be found in [4]. In the sequel, we assume that the reader is familiar with the state described by [4], and concentrate mostly on more recent publications.

[18] suggested a 'Geometrical Scene Description (GSD)' as a suitable interface between the Computer Vision task and a conceptual representation level, discussed in the context of road traffic scenes. The natural language descriptions derived in NAOS ([18]) from a GSD, however, are based exclusively on synthetically generated trajectories in contradistinction to the work reported here.

As part of the project VITRA ([12]), the system SOCCER ([2]) and its subsystems derive natural language descriptions from soccer scenes. The authors use real-world image data as background, but have to interactively provide (synthetic) trajectories in order to study the performance of their approach. [4], too, are concerned with conceptual descriptions of traffic scenes, but do not derive natural language descriptions. These authors use Bayesian Belief Networks to

perform the required inference steps for the extraction of a conceptual representation deemed suitable for the tasks envisaged by them.

[1] present a system which derives natural language descriptions of the location of renal stones found in radiographs. Since their system is based on the evaluation of single frames only, [1] do not take any kind of temporal relationships between the objects in the scene into account.

In contrast to the above–mentioned systems which derive conceptual descriptions from image sequences, the system PICTION ([22]) addresses the inverse problem of interpreting single images by exploiting accompanying linguistic information about the images, such as captions of newspaper photographs. In this case, the problem consists not in the generation of natural language descriptions, but in converting a conceptual description into an internal representation in order to use it for identifying objects in images.

In this context, two additional approaches should be mentioned which are concerned with an association between the output of some kind of image evaluation and the algorithmic generation of natural language text. [7] uses special so-called 'Lexical Conceptual Structures (LCS)' in order to represent the output expected from a computer vision system for language generation. His considerations, however, have not yet been applied to results obtained from real image data.

[3] describe a system for selecting excerpts from text documents treated as a digitized image. The goal consists in the extraction of summary sentences using algorithms based on the statistical characterization of word appearances in a document. Images are segmented to localize text regions, text lines, and words. Sentence and paragraph boundaries are identified. A set of word equivalence classes is computed. Word frequencies, word locations, and word image widths are used to rank equivalence classes as to their likelihood of being significant words. Subsequently, each sentence is scored based on the statistical appearance of significant words. The n best-scoring sentences are selected as an excerpt of the text. Although [3] generate natural language text based on image processing, these authors do not build up a system-internal representation for the *semantics* of the analysed text paragraphs. As a consequence, they are unable to generate independently formulated sentences.

A basically analogous task is pursued by [21] who developed an algorithm for *skimming* videos in order to automatically identify sections which comprise either 'important' audio or video information. Relative importance of each scene in the video is determined based on objects appearing in a segment, on associated words, and on the structure of the scene. Language analysis identifies 'important' audio regions by considering their frequency of occurrence. Words, for which the ratio of occurrence frequency divided by the corresponding frequency in a standard text exceeds a fixed threshold, are selected as keywords. A first level of analysis creates a reduced audio track, based on keywords. In order to increase comprehensibility of the audio track, 'keyphrases' are used which are obtained by starting with a keyword and extending segment boundaries to areas of silence or neighboring keywords. Although such a procedure provides a kind of 'semantic

sieve', it does not allow to capture the semantics proper, for example to facilitate paraphrasing.

A similar problem has been taken up by [19] who developed 'Name–It', a system which associates faces and names in news videos. This system can either infer the name related to a given face, or it can locate faces in news videos based on a name obtained from caption text enclosed with the video image or from the related audio track. A natural language processing step utilizes a dictionary, a thesaurus, and a parser for lexical/grammatical analysis as well as knowledge about the structure of a typical news video. 'Name–It' calculates a normalized score for each word. A high score corresponds to a word which is assumed to correspond to a face. So, similar to [22], [19] address the inverse problem of interpreting image sequences by exploiting accompanying linguistic information.

Similar to [4], [13] use dynamic belief networks in order to reason about traffic events which have been recognized by tracking contours of vehicle images in highway traffic scenes. Associated with each moving object is a belief network which contains nodes corresponding to sensor values and nodes corresponding to descriptions derived from the sensor values. The knowledge about the domain of discourse is encoded in probabilities associated with each node. These probabilities can be learned provided enough examples are available. But in this case, the driving behaviour which has been modelled by these parameters is not represented in an explicitly intuitive way. The inspection of such probabilistic models is more difficult than in our hierarchical approach. Moreover, the underlying image data of [13] (highway scenes) does not contain as complex maneuvers as our image sequences. Though [13] operate in the image domain, they are able to treat certain occlusion relations by exploiting knowledge about the special camera pose.

[5,6] describe their concurrent, hierarchical system SOO–PIN for high level interpretation of image sequences. These authors use a procedural programming language for geometric processing and an object–oriented logic programming language for symbolic reasoning. The domain of discourse is modelled by means of a network of concept–frames. Uncertainty is propagated using the Dempster–Shafer–Theory. While this system is well structured and the extracted high–level–descriptions are rich and facilitate consistency checks on a rather abstract level, the image processing level appears more brittle. Object recognition is performed by background subtraction and an object's speed is estimated by matching object candidates in only three consecutive frames. The resulting geometric data provides only a snapshot of the scene and, like [13], only refers to the *image domain.* In contradistinction, our system has been fed by automatically estimated trajectories of moving objects which have been tracked uninterruptedly over thousands of half frames in the *scene domain.* This enables us to precisely evaluate complex spatial relations, in particular occlusion relations. The large volume as well as the relatively high precision and reliability provided by this kind of image sequence evaluation stimulates extensive research into the use of abstract knowledge representation for the transformation of such data into natural language text.

The situation graph formalism for explicitly representing (traffic) situations is based on [17], which use conceptual graphs in order to represent state and action descriptions of situations. In this approach, the uncertainty of the input data was not exploited, and a procedural language was used for the implementation of the graph traversal algorithm. Now, we employ logic predicates for knowledge representation and an extended logic programming language developed by [20] in order to capture *fuzzy metric temporal predicates.* This facilitates the investigation of different graph traversal strategies by simply modifying some rules of the logic program.

4 Algorithmic Generation of Natural Language Text

The CL obtains descriptions regarding elementary facts derived by the Signal Layer in the form of *fuzzy metric temporal* predicates. Each such predicate comprises a relation identifier together with identifiers denoting the relation arguments, a degree_of_certainty in the form of a real number between 0.0 and 1.0, and a tuple consisting of the initial and final half-frame number defining the associated validity interval. Among the conceptual primitives, we distinguish motion-primitives, road-primitives, occlusion-primitives, and spatial relations. Motion-primitives characterize vehicle movements with respect to *mode* (such as forwards, stationary, backwards) and velocity (very slow, slow, normal, fast, very fast). The set of admissible movements comprises, too, the special 'movement' *to stand.*

Road-related primitives characterize details of the road model, in particular characteristics of lanes. The overall road model is structured hierarchically according to a type-hierarchy. Examples are *left_turning_lane(fobj_5)* which indicates that the object reference *fobj_5* refers to part of a left-turning lane. A predicate *kriegsstrasse(fobj_5)* indicates that *fobj_5* is part of a street named 'Kriegsstrasse'.

Occlusion primitives characterize in a deictic manner occlusion relations between road components, vehicles, and other objects.

4.1 Descriptions provided by the Conceptual Layer

The CL depends on what we call 'situation analysis' in order to construct abstractions from the facts provided by the SL. Trees of situation-graphs comprise the schematic knowledge about elementary 'occurrences' denoted by motion-primitives and their composition to more abstract concepts, for example characterizing sequences of vehicle manouvers, each of which is specified by a motion primitive. In this context, we speak about a *situation node* in a graph as comprising both a schematic *state description* of the agent within its current environment and an associated *action description* which specifies the action to be performed by an agent in the given state. The state description specifies all conditions which must be satisfied in order to enable the agent to perform the associated action.

```
GRAPH gr_get_petrol : get_petrol
{
    START SIT drive_to_pump
        : drive_to_pump(Lane),
          fill_in_petrol(Lane)
        {
            on(Agent, Lane);
            passing_lane(Lane);
            approaching(Agent);
        }
        { drive(Agent); }
    SIT fill_in_petrol
        : fill_in_petrol,
          leave_pump
        {
            on(Agent, filling_place: Fplace);
            velocity(Agent, zero);
        }
        { remain_standing(Agent); }
    FINAL SIT leave_pump
        : leave_pump(Lane)
        { on(Agent, Lane); passing_lane(Lane); }
        { leave(Agent); }
}
```

Fig. 4. Small part of a Situation-Graph in *SIT++*-notation. The subgraph *gr_get_petrol* includes situations which specialize the more common situation *get_petrol*. The defining part of a situation is introduced by the denotation 'SIT'. It is followed by the situation name and a list of possible successor situations sorted by decreasing priority. The first parenthesis includes the state scheme, the second parenthesis includes the action scheme associated with a situation.

The hierarchical arrangement of situation graphs enables the specification of subgraphs characterizing particular configurations of situations which occur sufficiently frequently to treat them as a conceptual unit. Several tree traversal strategies have been implemented for a traversal of this schematic knowledge base in the course of attempts to instantiate a particular sequence of situation nodes which is compatible with the stream of facts supplied by the SL. The instantiation process attempts to find at each half-frame time point the most specialized situation node which is compatible with the facts observed at that time point. Situation Trees are specified in a formal language illustrated in Fig. 4. Such specifications are translated automatically into metric temporal logic programs which can be evaluated by the metric temporal logic system *F-Limette* ([20]). F-Limette has generalized a tableau calculus by operations to treat metric temporal and fuzzy attributes.

Results at the level of situation analysis in our system are derived by the CL in the following manner. Traversal of the Situation Tree generates 'factual

data' or 'facts' by instantiating situation descriptions which are compatible with geometric results provided by the SL. We distinguish between several types of facts. Those which characterize the *actual motion description* (i. e. corresponding to the current half-frame number) can be derived from the state-schema component of a situation-node schema. Facts which characterize *intentions* of an agent can be extracted from a-priori knowledge encoded in admissible paths through sequences of Situation-Graph nodes linked by edges of this graph: an intention is simply considered to be represented by a possible continuation of a partially instantiated path through the Situation-Graph. Facts related to actions terminating in some situation are characterized by the path through the Situation-Graph instantiated so far. All facts are considered to be true, i. e. we have not yet implemented the fuzzy logic equivalent of linguistic hedges.

In contradistinction to earlier, procedural approaches, the approach sketched here supports modelling several alternative 'views' on an image subsequence, for example differentiating with respect to the level of detail presented or by optional suppression of occlusion relations between selected objects in the scene.

4.2 Transformation of primitive descriptions into natural language text

[14] describe algorithms which transform a natural language English input text into a system-internal, logic-oriented representation called *Discourse Representation Structure (DRS)*. In analogy to the *Discourse Representation Theory (DRT)* of [14], we treat the stream of fuzzy metric temporal facts provided by the CL as a string of symbols, i. e. as a text, and convert it into DRSs. The grammar specifying the exchange of a symbol string between the CL and the NL is simpler than even the restricted English grammar used by [14]. The important point, however, consists in the conversion of symbol strings into a semantic representation which can be manipulated by standard logic operations. We thus had to devise conversion rules and an algorithm which evaluates these rules in order to convert the input stream of (abstracted) facts provided by the CL into suitable DRSs.

The syntax rules of the grammar specifying the 'language' used to communicate between the CL and the NL are used to first convert the stream of symbols into a syntax tree which is subsequently analysed according to rules corresponding to DRT-rules as described by [14]. The advantage of such a procedure consists in generating a system-internal representation which can – at least in principle – be manipulated by any program that is able to evaluate DRSs for the semantic content of a natural language text according to [14].

The DRS construction rules used in the current version of our system implementation differ somewhat from those described in [14]. Our language fragment differs from the fragment of the English language used by [14]. This pertains also to the syntax trees. In addition, [14] have designed their system to cope with texts generated outside their system, whereas the evaluation of 'text' transmitted within our system from the CL to the NL can exploit many presuppositions. The

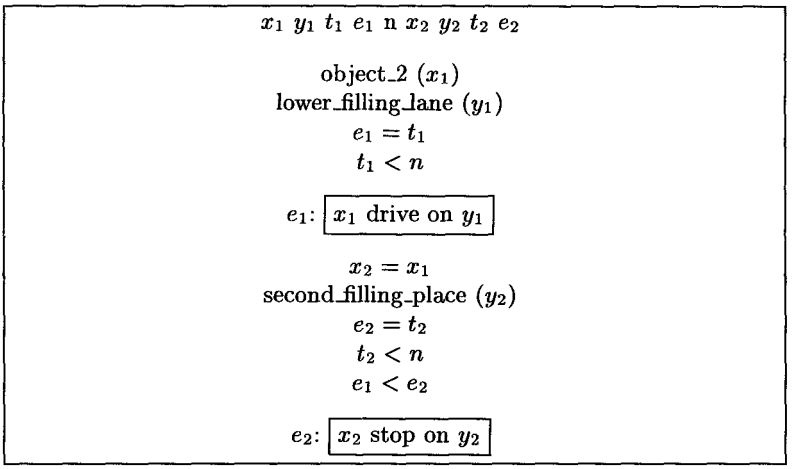


Fig. 5. Simple Example of a DRS.

current version of our NL implementation thus does not have to cope with ambiguities. This implies, of course, that a particular semantics has to be encoded by the CL by selection of appropriate instances of facts.

Figure 5 illustrates a simple Discourse Representation Structure (DRS) which consists of the following two facts:

5 : 107 ! drive_on(object_2, lower_filling_lane).
108 : 144 ! stop_on(object_2, second_filling_place).

The first of these represents the statement valid between half-frame number 5 and 107, expressing that the object 'object_2' drives on the lower filling_lane. The second statement expresses the fact that this object stops during the half-frame interval 108 through 144 on the second filling_place. The first 'discourse referent (DR)' denotes the agent and is given by the first argument. The second argument denotes a reference object. Facts themselves are converted to event structures denoted by e_i which represent the duration of such an event. Since the second event of this example immediately succeeds the first one, we have that $e_1 < e_2$. Discourse referents t_i refer to temporal intervals comprising events. Since no such comprising intervals have been specified in our example, we have $e_i = t_i$. Since all facts are treated in retrospect, all intervalls in our example belong to the past, i. e. $t_i < n$ where n denotes the time at which the sentences are formulated, i. e. 'now'.

The transformation of DRS into natural language text does not belong to Computer Vision, but to Computational Linguistics. We nevertheless found it useful to implement a simple conversion from DRSs to natural language text in order to facilitate inspection of discourse representations created automatically by our system. This 'casual' implementation comprises a number of heuristics. It is our hope that the system approach will not be judged entirely on the

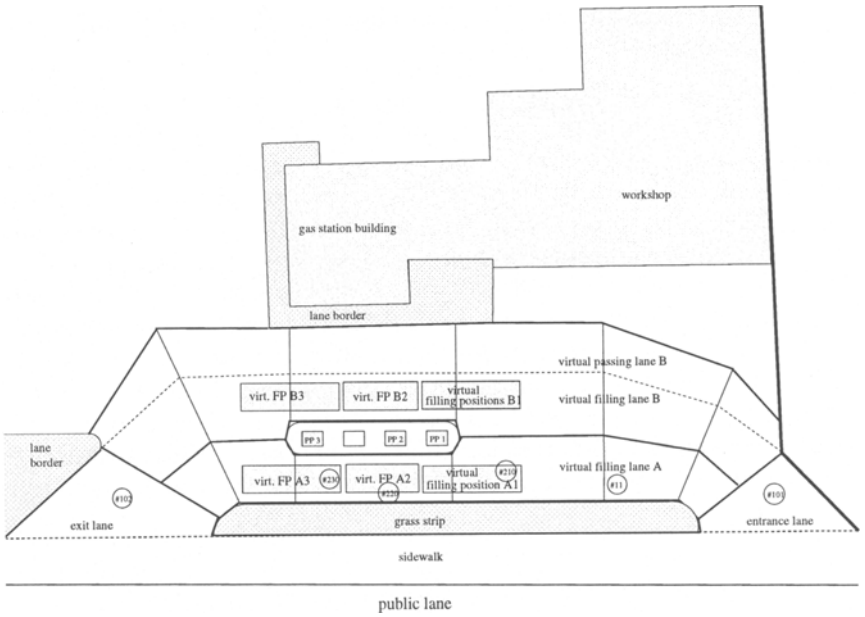


Fig. 6. Groundplane schematic map of the gas station.

insufficient quality of the natural language 'surface text'. We hope to eventually leave this transformation to professional linguists.

5 Experimental Results

The geometric evaluation process interpolates between fields of a digitized video sequence in order to obtain full-frame resolution with a sampling rate corresponding to a 20 msec interval between the resulting half-frames. The image sequence illustrated by Fig. 2 comprises about 8000 half-frames or almost three minutes of recording time. It thus becomes possible to test natural language text generation based on data sets which approach realistic proportions. This allows to compute several different natural language descriptions in order to study the formulation of texts which emphasize *different aspects* about the *same vehicle* as being relevant for the reader.

In order to simplify reference to details within images from this sequence, Fig. 6 shows a groundplane schematic map of the gas station in Fig. 2.

In the automatically generated texts of subsequent examples, the term *lower filling lane* refers to the 'virtual filling_lane A' from this map, whereas the term *upper filling lane* refers to 'virtual filling_lane B'. The term *filling_place* refers to a location which is situated on a filling lane, but next to a petrol pump (denoted as 'virtual filling_position' in Fig. 6). The text generation assumes that a vehicle obtains gas when it stops and remains standing for a while on a filling_place.

5.1 Stepwise introduction of knowledge about the relation between actions and locations

Our first example refers to 'object_2' which can be seen in the foreground of Fig. 2. The automatic generation of natural language text presented in Fig. 7 emphasizes actions of the agent (i. e. the vehicle 'object_2') as they can be observed for a particular time (i. e. half-frame number) interval. The verbphrase in a sentence thus exclusively comprises motion-verbs. The CL analyses *spatial relations* in this example *independently from actions* and, therefore, does not relate location and action with each other. In this case, action and location are only combined at the NL. It thus is difficult to extract statements about any *intention* of the agent from the information provided by the CL to the NL. This deficiency is compensated for by the fact that no a-priori knowledge about a particular scene is used. In this case, the system may generate reasonable statements about the vehicle even for situations where this agent does not behave according to some a-priori expectations, i. e. knowledge specific to a situation in this particular scene.

```
  5 :  119 ! drive_to(obj_2, dispensing_pump).
  5 : 1340 ! on(obj_2, lower_filling_lane).
  7 :  104 ! drive_slowly(obj_2).
105 : 1164 ! remain_standing(obj_2).
120 : 1222 ! take_in(obj_2, petrol).
1165 : 1366 ! drive_slowly(obj_2).
1223 : 1366 ! leave(obj_2, filling_station).
1341 : 1353 ! on(obj_2, upper_filling_lane).
```

Obj_2 slowly drove on the lower filling lane to the dispensing pump. Then it remained standing in order to take in petrol. After that, it slowly drove on the upper filling lane in order to leave the filling station.

Fig. 7. Natural language text automatically derived from facts generated by the 'Conceptual Layer (CL)'. The text refers to 'object_2', the vehicle shown in the foreground of Fig. 2. In this example, the 'Natural Language Layer (NL)' itself, i. e. without an explicit inference process by the CL, could relate an *action* ('remain standing') to an *intention* ('to take in petrol'), since this relation has been coded directly into the knowledge base of our system. This, in turn, is due to the fact that it has not yet been possible to automatically determine the activity 'to take in petrol' based on image sequence evaluation. This would require the analysis not only of moving *rigid,* but in addition of *nonrigid, jointed* bodies such as a human.
The last sentence refers to driving on the 'upper' filling lane since the *final* part of the vehicle trajectory overlapped also the 'virtual filling lane B' shown in Fig. 6.

In contradistinction to the example illustrated in Fig. 7, the text generation shown in the example of Fig. 8 exploits scene-specific knowledge provided by a situation-analysis based on a situation-graph. It thus becomes possible to identify 'intentions' of the agent and express this fact by introducing formulations like

5 :	107 !	want_to_get_to(obj_2, free_dispensing_pump).
5 :	107 !	drive_on(obj_2, lower filling lane).
108 :	144 !	stop_on(obj_2, second filling place).
145 :	1157 !	take_in(obj_2, petrol).
1158 :	1164 !	start(obj_2).
1158 :	1353 !	finish_getting(obj_2, gas).
1158 :	1353 !	want_to_leave(obj_2, filling_station).
1165 :	1353 !	drive_forward(obj_2).
1354 :	1366 !	get(obj_2, petrol).

Obj_2 wanted to get to the free dispensing pump. So it drove on the lower filling lane. Later it stopped on the second filling place. Then it took in petrol. Then it started because it finished getting gas. Now it wanted to leave the filling station. It, therefore, drove forward.

Fig. 8. Second example of natural language text automatically derived from facts generated by the Conceptual Layer.

want to. Analogously, inferences about partial goals which have been reached by the agent will be formulated using the verb to *finish*.

In addition, the text generation attempts to take into consideration to which extent knowledge about the relation between the vehicle and stationary scene components influences the formulation of agent behavior, in particular regarding the actions performed as well as regarding potential intentions underlying these actions. These attempts result in the formulation of intermediate goals such as, for example, to reach a free filling_place.

5.2 Incorporation of quantified statements into the text

Other sequences are shorter – for example from the intersection scene illustrated by Fig. 1. They are used in order to demonstrate additional aspects of our approach. The third experiment documented in the following Fig. 9 demonstrates how statements about groups of vehicles are incorporated into the formulations. On the one hand, sentences refer to single vehicle behavior; on the other hand, the NL text generation uses quantorized statements, too. Such formulations are based on a situation analysis comprising *several* vehicles in order to associate their compound behavior with an appropriate natural language quantifier. The road names used in this text have been indicated in Fig. 1.

6 Discussion of Related Research and Conclusion

An increasing algorithmic ability to fast and reliably evaluate extended image sequences will result in voluminous sets of geometric data. A potential user is thus challenged to interactively interrogate this material according to a variety of aspects which in general can not be prespecified. As a consequence, the user is likely to formulate his question in natural language and will expect to obtain

Obj_4 turned left on Kriegsstrasse into Ettlinger Strasse. It followed obj_5. It fell behind of obj_5 and obj_9.
Most vehicles drove from Kriegsstrasse into Ettlinger Strasse. Some vehicles drove straight ahead on Kriegsstrasse. No vehicle turned right from Kriegsstrasse into Karl-Friedrich-Strasse. A few vehicles drove from Karl-Friedrich-Strasse into Ettlinger Strasse. No vehicle drove from Karl-Friedrich-Strasse into Kriegsstrasse. No vehicle turned left from Kriegsstrasse into Karl-Friedrich-Strasse. No vehicle drove from Ettlinger Strasse into Karl-Friedrich-Strasse. No vehicle drove from Ettlinger Strasse into Kriegsstrasse.

Fig. 9. Two examples of automatically generated natural language descriptions refering to the intersection sequence illustrated by Fig. 1.

either a natural language text answering his question or some segment of the video illustrating a possible system response. In order to analyse the desired question, the system will require the ability to analyse first the natural language question posed by the user. This analysis then provides the boundary condition for the extraction of facts from data obtained by image sequence evaluation and – equally important – for the manner in which this material will be presented to the user.

We thus expect that the abstraction process from digitized video to natural language text will be much more involved than a search for simple 'features' – either pictorial ones, acoustical ones in the accompanying audio track, or a combination of both. Only detailed knowledge about the scenes captured by the video and about the intentions of agents in the scene as well as about possible intentions of a user will enable a system to extract and present the desired information in a suitable manner.

A system living up to such a challenge will have to build a dynamic internal representation of both the discourse context of the user-system interaction and of the contents of video segments which have already been evaluated. In order to retain the flexibility expected, it appears best to convert both the natural language text analysis of the user-system interaction as well as the abstractions from the image sequence evaluation into a compatible structure which will facilitate a joint logic manipulation necessitated in the course of the interaction.

Our experience with the evaluation of extended real-world image sequences has shown that the gap between geometrical descriptions, which have been extracted by image evaluation steps, and appropriate natural language descriptions is too large in order to be bridged by a heuristically extended Geometric Scene Description (GSD). We, therefore, decided to use a much more formal approach in order to be able to *systematically study all intermediate steps.* It turned out to be advantageous to subdivide the abstraction process from geometric descriptions to a level expressible as natural language text into two additional layers: one is devoted to a logical manipulation of material (the CL), and the other to the selection of material to be presented to the user at a particular moment during an interaction. In addition, the latter step requires the ability to suitably

package the selected material for presentation to the user. Selection and 'packaging' of evaluation results for the user at the appropriate level of abstraction is considered the task of the NL which is the topic of this contribution.

Although one might accuse us to 'misuse' Discourse Representation Theory (DRT) as developped by [14], DRT appears to us to offer an interesting framework for realizing a Natural Language Layer (NL). Looked at from this point of view, the NL as suggested in this contribution becomes a constituent part of an advanced Computer Vision System for the evaluation of image sequences. The examples given above illustrate that such an approach is at least feasible. All in all, data currently available to us comprise about fifty vehicles which have been detected and tracked automatically through subsequences extending sometimes through thousands of video frames. This allows to compute several different descriptions per vehicle in order to study the formulation of texts which emphasize *different aspects* about the *same vehicle* as being relevant for the reader.

Our initial attempts in the direction outlined in this contribution obviously still exhibit many limitations. We nevertheless consider the results obtained so far as encouraging future research.

Acknowledgement

We thank M. Haag for making image sequence evaluation data available for our experiments.

References

[1] A. Abella and J.R. Kender: *Description Generation of Abnormal Densities Found in Radiographs.* Proc. Workshop on Conceptual Descriptions from Images, Cambridge/UK, 19 April 1996, H. Buxton (Ed.), pp. 97-111.

[2] E. Andrè, G. Herzog, and T. Rist: *The System Soccer.* Proc. of the 8th European Conference on Artificial Intelligence, Munich/Germany, 1-5 August 1988, pp. 449-454.

[3] D.S. Bloomberg and F.R. Chen: *Document Image Summarization without OCR.* Proc. IEEE International Conference on Image Processing (ICIP '96), Lausanne/CH, 16-19 September 1996, Vol. II, pp. 229-232.

[4] H. Buxton and S. Gong: *Visual Surveillance in a Dynamic and Uncertain World.* Artificial Intelligence **78** (1995) 431-459.

[5] S. Dance, T. Caelli, and Z.-Q. Liu: *Picture Interpretation: A Symbolic Approach.* Series in Machine Perception and Artificial Intelligence Vol. **20**, World Scientific, Singapore a. o. 1995.

[6] S. Dance, T. Caelli, and Z.-Q. Liu: *A Concurrent, Hierarchical Approach to Symbolic Scene Interpretation.* Pattern Recognition **29**:11 (1996) 1891–1903.

[7] L. Friedman: *From Images to Language.* Proc. Workshop on Conceptual Descriptions from Images, Cambridge/UK, 19 April 1996, H. Buxton (Ed.), pp. 70-81.

[8] R. Gerber and H.-H. Nagel: *Berechnung natürlichsprachlicher Beschreibungen von Straßenverkehrsszenen aus Bildfolgen unter Verwendung von Geschehens- und Verdeckungsmodellierung.* In B. Jähne, P. Geißler, H. Haußecker und F. Hering

(Hrsg.), Mustererkennung 1996; 18. DAGM-Symposium, Heidelberg/Germany, 11.-13. September 1996, pp. 601-608 (in German).

[9] R. Gerber and H.-H. Nagel: *Knowledge Representation for the Generation of Quantified Natural Language Descriptions of Vehicle Traffic in Image Sequences.* Proc. IEEE International Conference on Image Processing (ICIP '96), Lausanne/CH, 16-19 September 1996, Vol. II, pp. 805-808.

[10] M. Haag, H.–H. Nagel: *Beginning a Transition from a Local to a More Global Point of View in Model–Based Vehicle Tracking.* H Burkhardt, B. Neumann (Eds.): Proc. European Conference on Computer Vision 1998 (ECCV '98), Freiburg/Germany, 2-6 June 1998.

[11] M. Haag, W. Theilmann, K.H. Schäfer, and H.-H. Nagel: *Integration of Image Sequence Evaluation and Fuzzy Metric Temporal Logic Programming.* KI-97: Advances in Artificial Intelligence, Proc. 21st Annual German Conference on Artificial Intelligence, Freiburg/Germany, 9-12 September 1997; G. Brewka, C. Habel, and B. Nebel (Eds.): Lecture Notes in Artificial Intelligence vol. **1303**, Springer-Verlag Berlin, Heidelberg, New York 1997, pp. 301-312.

[12] G. Herzog and P. Wazinski: *VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions.* Artificial Intelligence Review Journal **8** (1994) 175-187.

[13] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. Rao, S. Russell, and J. Weber: *Automatic Symbolic Traffic Scene Analysis Using Belief Networks.* Proc. 12th National Conference on Artificial Intelligence, Seattle/WA, 31 July – 4 August 1994, pp. 966–972.

[14] H. Kamp and U. Reyle: *From Discourse to Logic.* Kluwer Academic Publishers, Dordrecht/NL, Boston/MA, London/UK 1993.

[15] H. Kollnig und H.–H. Nagel: *Ermittlung von begrifflichen Beschreibungen von Geschehen in Straßenverkehrsszenen mit Hilfe unscharfer Mengen.* Informatik – Forschung und Entwicklung **8** (1993) 186-196 (in German).

[16] H. Kollnig and H.-H. Nagel: *3D Pose Estimation by Directly Matching Polyhedral Models to Gray Value Gradients.* International Journal of Computer Vision **23**:3 (1997) 283–302.

[17] H.–H. Nagel, H. Kollnig, M. Haag, and H. Damm: *The Association of Situation Graphs with Temporal Variations in Image Sequences.* Working Notes AAAI–95 Fall Symposium Series 'Computational Models for Integrating Language and Vision', R.K. Srihari (ed.), Cambridge/MA, 10–12 November 1995, pp. 1–8.

[18] B. Neumann und H.-J. Novak: *NAOS: Ein System zur natürlichsprachlichen Beschreibung zeitveränderlicher Szenen.* Informatik – Forschung Entwicklung **1** (1986) 83-92 (in German).

[19] S. Satoh, Y. Nakamura, and T. Kanade: *Name-It: Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing.* Proc. 15th International Joint Conference on Artificial Intelligence (IJCAI '97), 23-29 August 1997, Nagoya/Japan, Vol. II, pp. 1488-1493.

[20] K.H. Schäfer: *Unscharfe zeitlogische Modellierung von Situationen und Handlungen in Bildfolgenauswertung und Robotik.* Dissertation, Fakultät für Informatik der Universität Karlsruhe (TH), Juli 1996. Published in: Dissertationen zur Künstlichen Intelligenz (DISKI), Band **135**, infix–Verlag St. Augustin 1996 (in German).

[21] M.A. Smith and T. Kanade: *Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques.* Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97), 17-19 June 1997, San Juan, Puerto Rico, pp. 775-781.

[22] R.K. Srihari: *Linguistic Context in Vision.* Proc. IEEE Workshop on Context-Based Vision, Cambridge/MA, 19 June 1995, pp. 100–110.