AI Magazine Volume 25 Number 2 (2004) (© AAAI)



# Steps toward a Cognitive Vision System

*Hans-Hellmut Nagel*

■ An adequate natural language description of developments in a real-world scene can be taken as proof of "understanding what is going on." An algorithmic system that generates natural language descriptions from video recordings of road traffic scenes can be said to "understand" its input to the extent that algorithmically generated text is acceptable to the humans judging it. A fuzzy metric-temporal Horn logic (FMTHL) provides a formalism for representing both schematic and instantiated conceptual knowledge about the depicted scene and its temporal development. The resulting conceptual representation mediates in a systematic manner between the spatiotemporal geometric descriptions extracted from video input and a module that generates natural language text. This article outlines a 30-year effort to create such a cognitive vision system, indicates its current status, summarizes lessons learned along the way, and discusses open problems against this background.

For ages, students have been asked to repeat a previously given explanation in their own words: An experienced teacher can infer the degree of understanding—or the lack of it—from the manner in which an explanation has been paraphrased. The ability to present a "variant formulation" without distorting the essential parts of the original message is taken as a cue that these essentials have been "understood." During art lessons, in particular those concerned with classical or ecclesiastic paintings, students are initially invited to merely describe what they see. Frequently, considerable a priori knowledge about ancient mythology or biblical traditions is required to succinctly characterize the depicted scene. Lack of the corresponding knowledge about other cultures can make it difficult for someone with only a European education to really understand and describe in an appropriate manner a painting by, for example, a Far East classic artist.

Familiar human experiences mentioned in the preceding paragraph will now be "morphed" into a scientific challenge: to design and implement an algorithmic engine that generates an appropriate textual description of essential developments in a video sequence recorded from a real-world scene. Such an algorithmic engine will serve as one example of a cognitive vision system (CVS), which leaves room, as the experienced reader has noticed, for there to be more than one way to introduce the concept of a CVS. An alternative clearly consists in coupling a computer vision system with a robotic system of some kind and assessing the reactions of such a compound system. To whomever accepts the formulation, "one of the actions available to an agent is to produce language. This is called a *speech act*. Russell and Norvig (1995)" is unlikely to consider the two variants of a CVS alluded to previously as being fundamentally different.

With regard to the first CVS version in particular, the following remarks are submitted for consideration: Obviously, we avoid a precise definition of *understanding* in favor of having humans compare the reaction of an algorithmic engine to that expected from a human. This fuzzy approach toward the circumscription of a CVS opens the road to constructive criticism—that is, to incremental system improvement—by pinpointing aspects of an output text that are not yet considered satisfactory. One might ask, moreover, whether unsatisfactory results are the result of an in-

*Figure 1. Frame 10 from the Hamburg Taxi Sequence.*

It was recorded more than a quarter of a century ago from our laboratory window at the Universität Hamburg.

ability of a CVS to exploit principally accessible knowledge or the result of the fact that the CVS does not command the a priori knowledge necessary to generate an appropriate formulation. Such a question focuses on the system-internal representation and exploitation of knowledge.

Readers familiar with the history of AI will note that the proposed CVS cannot (easily) pretend understanding based on ELIZA-type syntactic manipulations. The price for this advantage has to be paid in the form of heavy computational expenses for machine vision processes. Students who are introduced to image processing with currently available facilities—to record an image sequence using a notebook is a routine activity today—can scarcely imagine the effort required in the early 1970s to merely digitize a short video sequence and transfer it onto a laboratory computer. It took about four years for the research group I established in 1971 at the Universität Hamburg to acquire the facilities to record a sequence such as the well-known Hamburg taxi sequence (figure 1).

There is another subtlety associated with the CVS to be discussed here: The postulate to describe essential developments in a scene provides a built-in focus on changes, in particular, on movements. The number of verbs, for example, in the German language (about 9200) is much smaller than the number of words (about 140,000) available to denote abstract or concrete entities (nouns for living creatures, inanimate objects, abstract concepts) and their attributes (adjectives). Obviously, the threshold regarding computational expenses is much higher for the evaluation of entire image sequences than for a single image. Once one is able to pay the price involved (either by spending money or waiting patiently until time-consuming computations have been finished), image sequences recorded by a stationary camera offer an inherent focus of attention: what moves are relevant, at least on a short-term basis. In the more general case of animals, something moving has to be inspected for whether it is a thread, a prey, or a mating partner (or possibly a playmate in the case of young animals).

The discourse domain constitutes another important ingredient for CVS research. In the case to be discussed here, inner-city vehicular road traffic was chosen. It allows one to study a rich variety of vehicle maneuvers and spatiotemporal vehicle configurations. However, the amount of background knowledge to be provided to the system remains manageable. Vehicle maneuvers can be represented by a comparatively small number (≈ 100) of parameterized concepts. In addition, the lane structure of inner-city roads and road intersections can be extracted from images and provides a useful reference for both the prediction of vehicle movements and the formulation of textual descriptions.

The choice of a varied but well-structured discourse domain raised a considerable barrier against quick and dirty approaches. In hindsight, it was an advantage that some subproblems were amenable to early isolated treatment. Limitations quickly became apparent, however, when solutions to certain subproblems could serve as building blocks in the construction of a more encompassing system (figure 2). Two aspects deserve to be mentioned already at this stage. First, the necessity to iterate several times through the design-implementation-test cycle implied that some components had to be reconceived and reimplemented more than once to remove increasingly discernible bottlenecks of the overall system. Second, geometric and conceptual processing gradually separated with the consequence that an increasingly rigorous approach became feasible for conceptual processing, based on a fuzzy metric-temporal (Horn) logic, which includes a clear-cut interface to system components for signal and geometric processing.

Subsequent sections outline selected approaches, milestones, and results during the development of the CVS sketched in figure 2 in an attempt to condense the accumulated experience into insights about CVSs and their potential future development.

## The Core Computer Vision Subsystem

The discussion concentrates initially on the four layers in the lower left half of figure 2. Consecutive layers are connected by bidirectional links to the conceptual primitives level, which comprises the interface between geometric and conceptual processing. Subsequently mentioned examples will mostly refer to image sequences recorded by a single stationary fixed-lens camera such that the downward flow of information toward the sensor-actuator level for control of camera parameters will be of no concern here. Signal-related image transformations in the image- signal level such as low-pass filter operations are not treated.

A clear distinction between the picture and scene domain level (Kanade 1978) helps to organize the knowledge representation: The picture domain refers to the representation of spatiotemporal geometric structures restricted to the image plane—such as regions in a segmented gray-value image or optical flow field—whereas structures related to the depicted three-dimensional (3D) scene are treated in the scene domain.

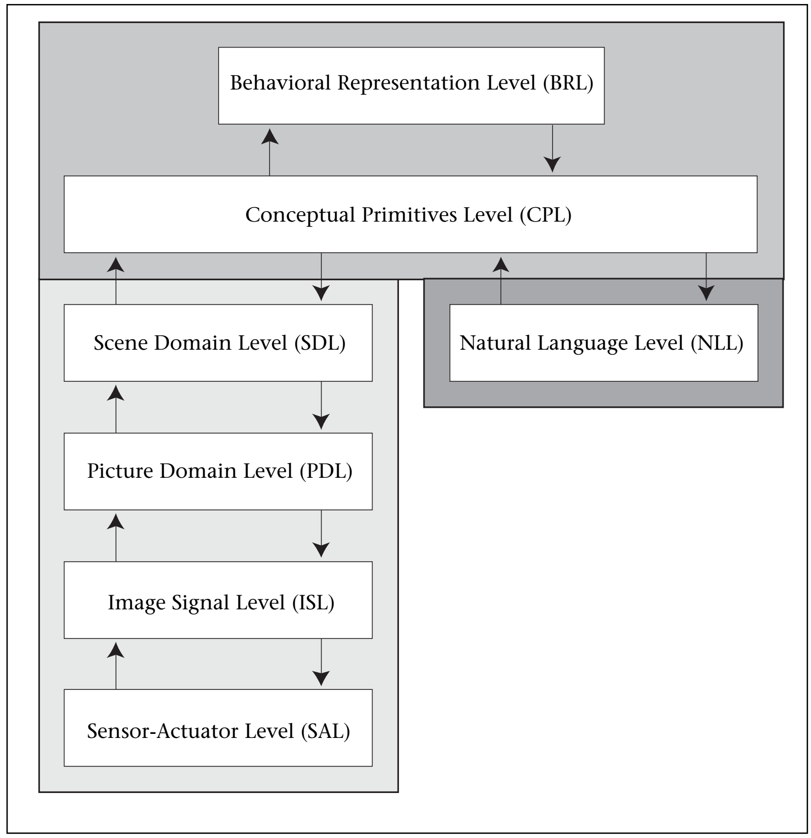


*Figure 2. Coarse-Layer Structure of the Overall System.*

The layers underlaid in light gray in a black-and-white printout constitute the core computer vision subsystem for the extraction of a geometric three-dimensional scene representation. The conceptual representation subsystem is underlaid in medium gray, the text generation is incorporated into the natural language level underlaid in dark gray. (From Nagel [2000] where a more detailed explanation can be found; © 2000 IEEE, reproduced with permission).

### Extraction of Vehicle Image Candidates

Given an image sequence of a road traffic scene recorded by a stationary camera, a first processing step has to detect images of vehicles and estimate their parameters, such as their position within an image frame. Because of limited computing power, we started by thresholding gray-value differences between consecutive image frames. The unreliability of such an approach led to efforts to develop more robust stochastic tests, but we eventually abandoned these efforts as well. Change cues can be the result of many reasons (motion but, in addition, illumination changes including time-varying reflections) and, thus, are difficult to interpret. It appeared more advantageous to directly estimate the frame-to-frame shift of identifiable gray-value structures (for example, Zimmermann and Kories [1984] and Sung and Zimmermann [1986]), or feature-based optical flow (figure 3). Thresholding the norm of such optical flow estimates and clustering the surviving neighboring optical flow vectors of (approximately) the same length and orientation directly extracted regions that exhibit in general a much higher correlation with images of moving vehicles than change regions (figure 4).

In principle, each cluster can be tracked from frame to frame, yielding an image-plane vehicle trajectory such as the ones illustrated in figure 5. Image-plane vehicle trajectory data obtained in this manner have successfully been associated with motion verbs in exploratory experiments (Koller, Heinze, and Nagel 1991) based on considerations developed in Nagel (1988).

Experiences during attempts to increase the robustness of this approach resulted eventually in the decision to drastically redesign the vehicle-detection and -tracking components at the picture-domain level and the interface to the conceptual primitive level (compare figure 2). In particular for small vehicle images, extraction and interframe linkage of gray-value features can vary considerably from frame to frame, which, in turn, influences the clustering of resulting optical flow vectors with the clearly visible effect that trajectories appear ragged (figure 5). Our research group did resist the temptation to fight this effect by smoothing operations in the image plane and decided to counteract its root cause by improving the tracking process.

*Figure 3. Feature-Based Optical Flow Results (Right Panel) Estimated from Image Regions (Left and Center Panels) That Have Been Cropped from Two Frames of a Sequence.*

Left and right panels: from Koller et al. (1991) © 1991 IEEE, reproduced with permission; center panel from Koller (1992) © 1992 infix, reproduced with permission.


*Figure 4. A Frame from a Sequence Recorded at a Busy Karlsruhe Intersection.*

Left panel, from Kollnig and Nagel (1997), © 1997 Kluwer, reproduced with permission. The center panel shows rectangles enclosing clusters of optical flow vectors obtained by a feature-based–estimation approach illustrated in the right panel of figure 3. The two triangles inscribed on each rectangle indicate the image-motion direction obtained from the optical flow estimates. The right panel shows an enlargement around the three vehicle image candidates, 5, 7, and 9, in the center panel. One can recognize that only a small number of feature-based optical flow estimates contribute to each cluster (center and right panels from Kollnig [1995], © 1995 infix, reproduced with permission.)

## Switching to a Model-Based Scene-Domain Tracking Process

Rather than tracking a cluster of optical flow vectors directly, such a cluster can serve merely to initialize a 3D–model-based tracking process, building on ideas reported earlier by Lowe (1991). A polyhedric vehicle model is tentatively placed in the scene at a location estimated by back projection of optical flow vectors within a cluster onto a plane somewhat above and parallel to the road plane. This approach assumes that the optical flow vectors originate from features painted onto a planar facet hovering about half the vehicle's height above the road plane. A back projection of optical flow vectors onto this facet provides an initial estimate for orientation and speed of the vehicle (figure 6). This initial model pose allows one to associate visible model segments with edge segments extracted from the image frame to improve the pose estimate (Koller 1992; Koller, Daniilidis, and Nagel 1993). The resulting improved vehicle pose constitutes the starting point of a 3D vehicle trajectory obtained through a succession of prediction/update cycles realized by a Kalman filter.

It turned out that shadows could create havoc during such a gradient descent pose improvement, as illustrated by figure 7. The lower contour segments of the car body have been fitted mostly to data segments associated with the car's shadow because the contrast between lower parts of the car's body and the shadowed road surface is smaller than the contrast between the shadow and the illuminated part of the dark road surface. Inclusion of the vehicle shadow in the model projection alleviates this problem (Koller 1992; Koller, Daniilidis, and Nagel 1993).

## Improved Three-Dimensional Pose Initializations

Numerous experiments with the approach outlined in the preceding subsection gradually convinced our research group to replace the feature-based optical flow estimates with gradient-based ones (Otte 1994; Otte and Nagel 1995); these estimates provide a much denser optical-flow vector field and enable a more robust initialization (figure 8).

A second significant modification abandoned the data-driven aggregation of edge elements into data segments that subsequently were tested for association with model segment projections computed on the basis of the cur-


*Figure 5. Vehicle Trajectories Obtained by Tracking Clusters of Feature-Based Optical Flow Estimates. (From Otto [1990]).*

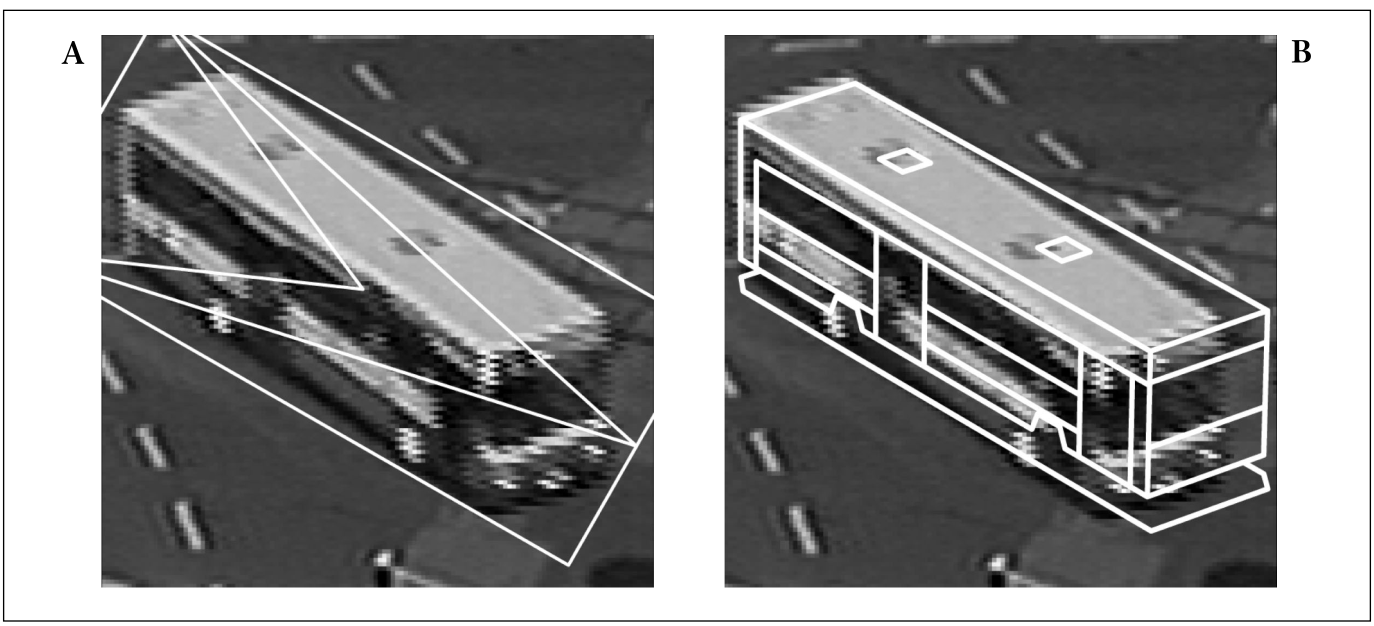


*Figure 6. Initialization of a Bus Model.*

A. This panel shows the vehicle image candidate represented by a rectangle oriented along the bus overlaid on the image of a bus taken from the third frame of the sequence used in figure 4. The two triangles within the rectangle indicate the driving direction estimated from the optical flow vectors incorporated into the cluster that provided the basis for this initialization. B. The resulting initial model instantiation, projected onto the image plane with hidden lines removed (see also figure 9). (From Kollnig and Nagel [1997] © 1997 Kluwer, reproduced with permission).

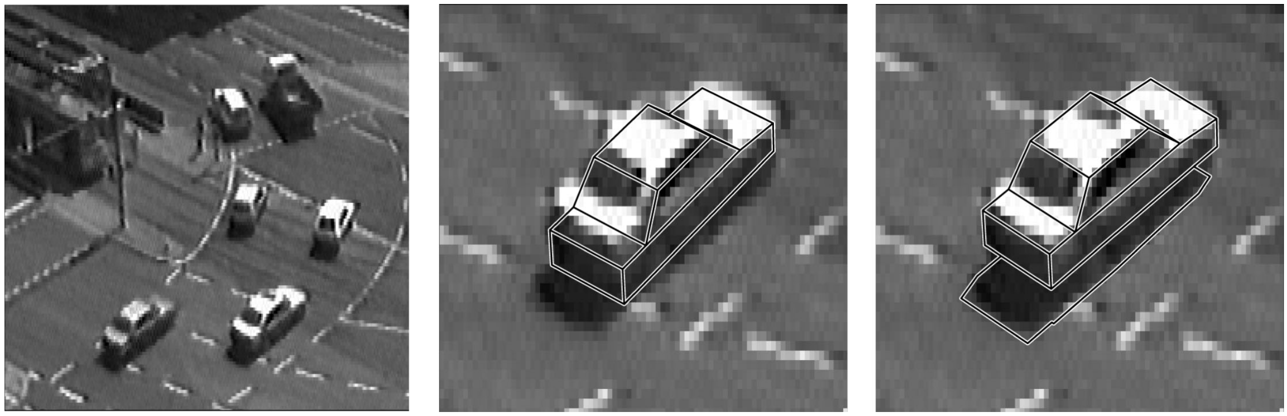

*Figure 7. The Advantage of Taking Shadows into Account.*

The left panel shows an image frame from an early sequence recorded at the Durlacher-Tor-Platz in Karlsruhe (from Koller [1992] © 1992 infix, reproduced with permission). An initialization step for three-dimensional model-based tracking exploited information about vehicle image candidates obtained from the segmentation of feature-based optical flow fields such as those illustrated in figures 3 and 4. The center panel shows a polyhedric model for a sedan, superimposed on a window cropped around the vehicle in the lower right part of the left panel, following a gradient descent fit of data line segments (extracted from the image) to model segments. If the shadow of the vehicle is included in the model projection—as illustrated by the right panel—the overall fit greatly improves (center and right panels from Koller, Daniilidis, and Nagel [1993] © 1993 Kluwer, reproduced with permission).

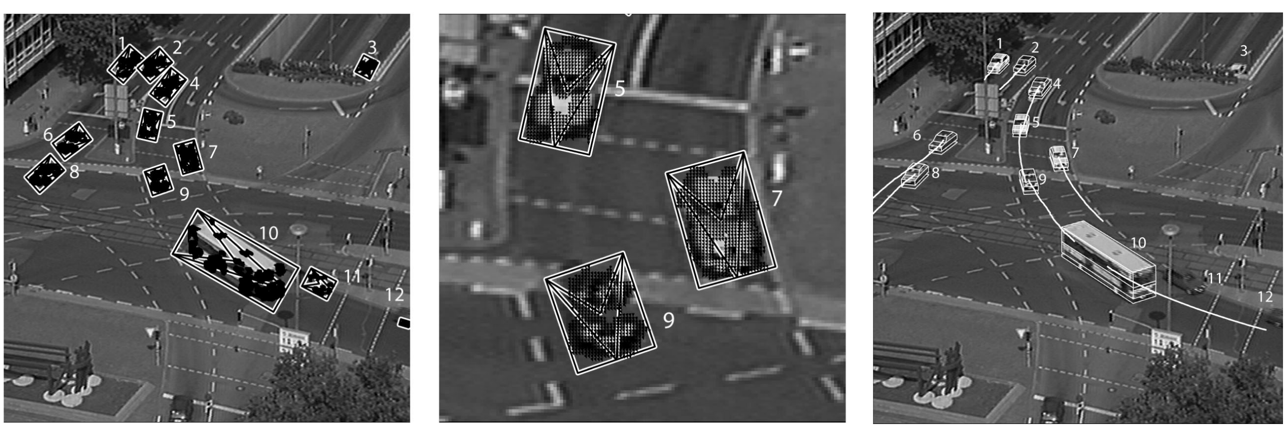


*Figure 8. Improved Initialization Exploiting the Segmentation of Gradient-Based Optical Flow Field Estimation.*

The left and center panels are analogous to the center and right panel of figure 4, but here the clustering algorithm is applied to a much denser optical flow field derived by a gradient-based approach described in Otte and Nagel (1995). The right panel illustrates results of a model-based tracking approach (from Kollnig, Nagel, and Otte [1994] © 1994 Springer-Verlag, reproduced with permission).

rent vehicle pose estimate. As illustrated by figure 9, edge elements extracted from the current image frame are individually tested for association with visible model segments, thereby exploiting the knowledge provided by the predicted vehicle pose to select only edge elements in the vicinity around visible model segments. This modification not only avoided a time-consuming edge-element aggregation process in areas where it would not matter anyway, but it also reduced the danger that such a "blind" edge-element aggregation process would result in an incorrect aggregation and, subsequently, in unwanted matches between data and model segments. Such mismatches could distort the pose estimate and thereby increased the risk of tracking failures.

A third important modification exploits a priori knowledge about the position of lanes. Optical flow vectors associated with the images of vehicles, which drive close to each other in neighboring lanes at about the same speed, can

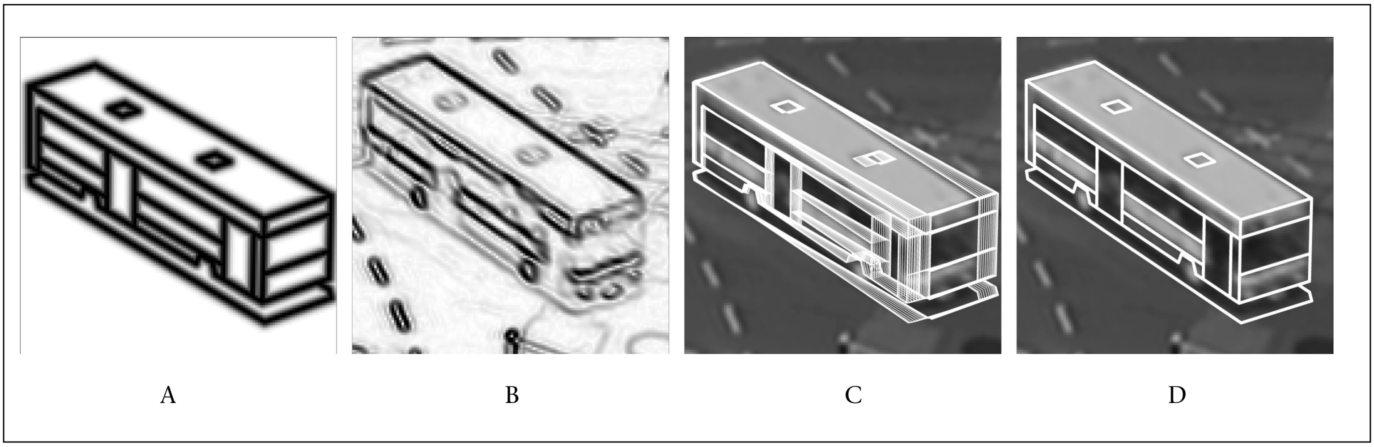


*Figure 9. Fitting Fuzzified Model Segments Directly to Gray-Level Gradient Magnitude.*

A. This panel shows the visible model segments of the initial model instantiation (see figure 6) "fuzzified" (that is, extended into their image plane environment) by convolution with a two-dimensional Gaussian filter. B. Edge elements extracted from the same image frame as used for figure 6 are given, where darker values indicate a higher-gradient norm. C. This panel illustrates the succession of fits obtained by the update step of an iterated extended Kalman filter. D. The final result is overlaid on the bus image (from Kollnig and Nagel [1997] © 1997 Kluwer, reproduced with permission).

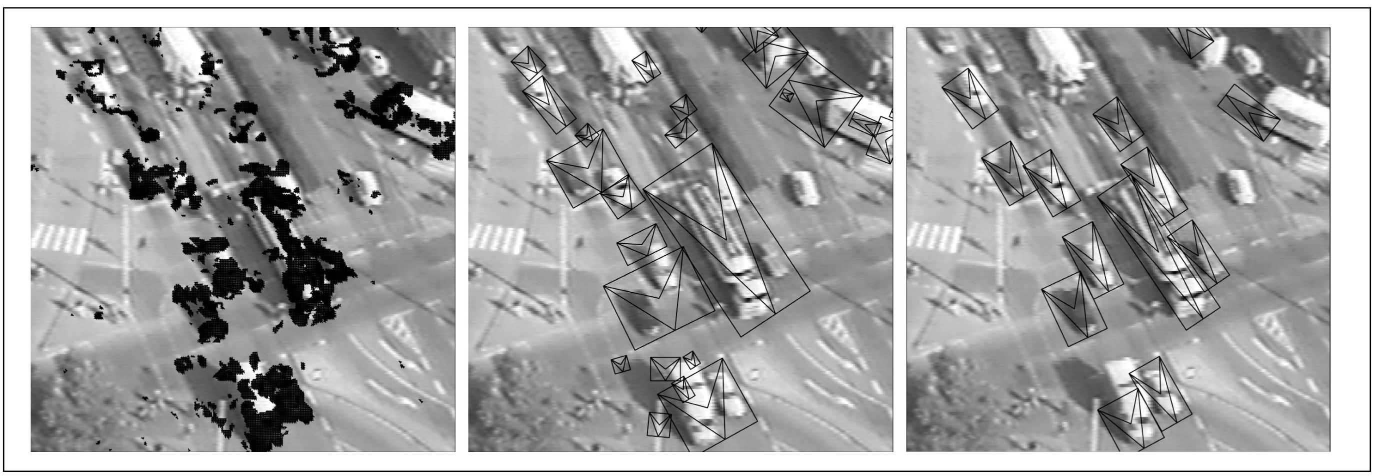

*Figure 10. Exploiting Knowledge about the Lane Structure.*

The left panel shows a window cropped from a video sequence recorded at a road intersection in Frankfurt, Germany, overlaid by optical flow vectors whose norm exceeds a threshold. Rectangles corresponding to clusters of optical flow vectors are overlaid the same image area in the center panel. Vehicle image candidates generated in this manner can cover the image of more than a single vehicle or none at all (for example because of pedestrians). If knowledge about the lane structure is available, optical flow vector blobs covering neighboring lanes can be split based on the hypothesis that they are the result of two separate vehicles driving side by side (see right panel). In addition, clusters have been suppressed if their size did not exceed a minimum area threshold. (From Kollnig, Leuck, and Nagel [1995] © 1995 Springer-Verlag, reproduced with permission).

be collected together into a single large cluster. The interactive provision of a polygonal representation of the lane structure for the road plane in the scene enables a heuristic that splits a single blob of optical flow vectors covering two neighboring lanes into two subblobs. Each of these subblobs can then be used to generate a vehicle image candidate, as illustrated in figure 10. The image corresponds to a time shortly after the traffic lights had switched to green for vehicles coming from the top. Heavier vehicles or those in the rear parts of the queues are just beginning to accelerate with the consequence that their speed—and, thus, the associated optical flow vectors—were still rather small. Because of the dense optical flow field in image areas corresponding to (parts of) moving vehicles, the overlaid optical flow vectors appear as dark blobs.

## Improvement of Tracking Capabilities

Once the initialization phase had considerably been improved compared to earlier system versions, less obvious problems in the tracking phase were attacked. A combination of three major modifications enabled a kind of "quantum jump" for tracking robustness: (1) transition to half-frame tracking for interlaced video sequences, (2) exploitation of the direction information associated with edge elements, and (3) incorporation of optical flow estimates into the state update with an iterative extended Kalman filter.

Based on a judicious discretization of partial derivatives of a Gaussian low-pass filter, it became possible to estimate image gradients and optical flow vectors with full-frame resolution at each half-frame time point: The operator masks incorporated a suitable interpolation between odd and even half frames (fields) of interlaced video digitizations (Otte 1994; Otte and Nagel 1995), allowing a cut in the prediction period by a factor of two, thereby reducing the extrapolation error to one-fourth compared to full-frame prediction. The implied doubling of the prediction/update frequency has to be paid for by a doubling of computational expenses. This effect has been compensated, however, by the increase in computing power within eighteen months: It turned out to be more advantageous to lag behind the highest tracking speed attainable at any one time than to trace down complicated tracking failures that could build up over a long period by accumulating very small residual discrepancies of the fitting process.

Previously, only the distance between an edge-element location and the model segment had been taken into account by the state-update phase of the Kalman filter. The second improvement also incorporated the orientation difference between the gradient direction and the normal from the location of an edge element to the model segment. This modification allowed the exclusion of edge elements from being tentatively associated with a model segment if the orientation difference turned out to be too large. In addition, edge elements that are better aligned with the current model segment contribute more to the state parameter update than those that are less well aligned although still within the orientation tolerance.

The third major modification extended the residual function by including the difference between the displacement rate—determined for each visible surface picture element on the basis of the current state estimate—and the optical flow vector estimated at the corresponding image location. Because of the recording conditions that in our case require a large field of view of the stationary camera to follow vehicles during significant maneuvers, vehicle images are usually small. The number of pixels that can be exploited for edge-element extraction thus tends to be much smaller than that accessible to optical flow estimation. Incorporation of optical flow estimates improves the velocity estimation and thereby significantly stabilizes the tracking process, in particular during partial occlusion of a vehicle.

As a result of these major improvements—and a number of other ones that cannot be treated here because of space limitations—the rate of successfully tracked vehicle images increased significantly, as documented in Haag (1998) and Haag and Nagel (1999) (figure 11).

## Association of Maneuver Concepts with Vehicle Trajectories

Given the ability to track road vehicles under realistic boundary conditions, a next step toward a CVS associates concepts for *recognizable movement primitives* with segments of estimated vehicle trajectories. Such an association imports geometric results from the CVS subsystem—see the previous section—across the interface between the scene domain level and the conceptual primitives level into the conceptual representation subsystem introduced in figure 2.

Recognizable movement primitives can be considered elementary maneuvers that on the one hand can be performed by a vehicle and on the other hand can be described by simple verb phrases. To emphasize the distinction between the system-internal representation of such an elementary activity and its linguistic expression, the abstract term *occurrence* is used for the internal representation.

Table 1 contains a small subset of occurrences for which system-internal representations have been constructed, here in particular for verb phrases involving the vehicle as agent and a location (for details, see Gerber and Nagel [2002]). Each occurrence can be characterized uniquely by a conjunction of predicates. These, in turn, consist of a conjunction of as many as three (sub)predicates, namely (1) a precondition (PREC) that has to be satisfied before the occurrence in question could be considered to represent a valid description of the temporal development in which the agent is involved; (2) a monotonicity condition (MONC or MC), indicating the type of admissible monotonous change that might take place when the occurrence represents a valid description; (3) a postcondition (POSTC) that becomes true once the occurrence in question no longer constitutes an adequate description of the temporal development in which the agent is involved.


*Figure 11. Typical Results Obtained by Vehicle Tracking at the Scene Domain Level.*

The polyhedral vehicle models used and the resulting trajectories are overlaid in a vehicle-candidate–specific shade, with a vehicle candidate number plotted next to each vehicle image in the same shade. The video image frame corresponds to a time shortly after traffic lights had switched to red for vehicles coming from the right. Those vehicles that had turned left into the two lanes ending at the lower border of the field of view had to slow down because they could not yet proceed further. The traffic lights had already changed to green for vehicles coming from the top left corner; the first of these vehicles had just begun to cross the intersection. (Courtesy M. Haag.)

Fuzzy membership functions such as those illustrated in figure 12 encode the (principally vague) a priori knowledge about the relation between the (3D) speed estimate for the vehicle in question, as obtained by the geometric tracking process, and the conceptual values used to describe qualitatively this numerically given speed. Similar membership functions have been defined for the conceptual values that can be assumed by the predicates *has_course_toward_loc* and *has_distance_to_loc*. This information is used to convert the quantitatively given results obtained by the geometric tracking process to a *degree of validity* (a real number between 0.0 and 1.0) to the "fact" that the predicate has the corresponding qualitative conceptual value at a particular (half-frame) time point. These degrees of validity are evaluated by an inference engine (see Schäfer [1996]) that combines the conjunction of subpredicates—evaluated for each occurrence as a function of time according to a separately specified acceptance automaton—to obtain a degree of validity for the association of such an occurrence with the vehicle trajectory at a particular point in time. Figure 13 visualizes these associations for a small part of the trajectory of the bus (vehicle candidate 10), shown in the right panel of figure 8.

## Representation of Behavior

To this point, only individual actions (maneuvers) of an agent vehicle have been treated at the conceptual level. Associated occurrences correspond to verb phrases that can be com-

| Occurrence | has_speed(Agent) | | | has_course (Agent,Location) | | | has_distance (Agent,Location) | | |
|---|---|---|---|---|---|---|---|---|---|
| | PREC | MONC | POSTC | PREC | MONC | POSTC | PREC | MONC | POSTC |
| approach loc | moving | — | moving | approaching | — | approaching | not_zero | > | small |
| reach loc | moving | — | moving | — | — | — | small | > | zero |
| drive across loc | moving | — | moving | — | — | — | zero | — | zero |
| drive away from loc | moving | — | moving | leaving | — | leaving | small | < | not_zero |

*Table 1. Time-Dependent Predicates Defining Occurrences That Refer to Both the Agent and a Location.*

The symbol > indicates a decreasing slope for the value subject to the monotonicity condition MONC; the symbol < correspondingly indicates an increasing slope. The term *has_course* denotes the abbreviation of the predicate *has_course_toward_loc* with the conceptual values *approaching* and *leaving*. Similarly, *has_distance* stands for the predicate *has_distance_ to_loc*. See Gerber and Nagel (2002).

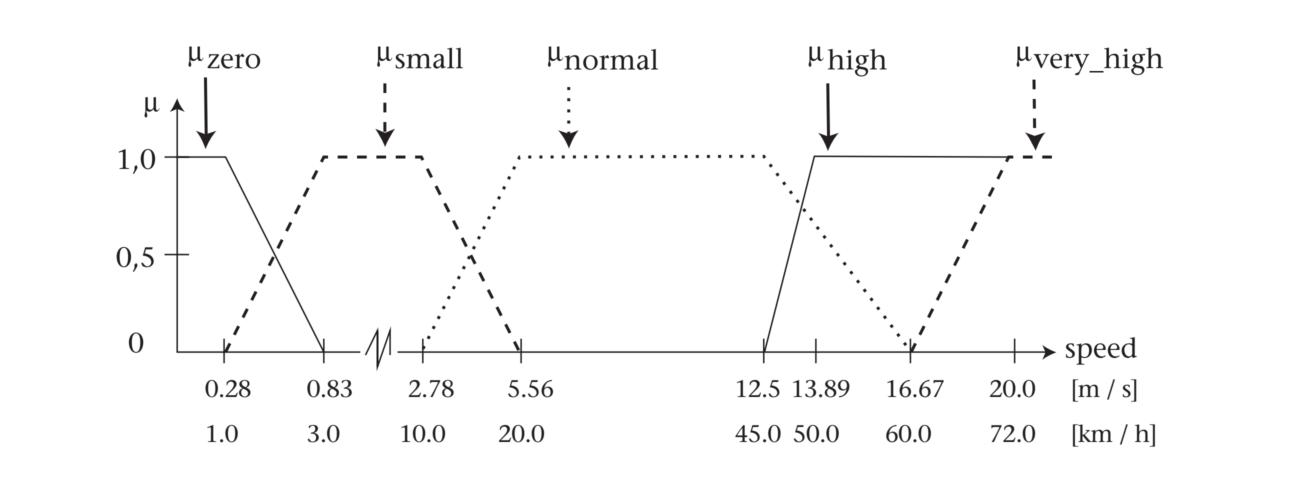


*Figure 12. Discretization of Continuous Speed Values into a Set of Intervals.*

The graph shows the fuzzy membership functions $\mu_{speed\ value}$ for the subset {zero, small, normal, high, very_high} of discrete conceptual speed values.

bined with a noun phrase referring to the agent vehicle—in the simplest case, just an identifier that is treated as a proper name—to construct a single sentence in isolation. A natural next step consists of an attempt to treat such actions within their mutual context, namely, to concatenate individual maneuvers in a manner compatible with experience to study the behavior of vehicular agents. Such a step corresponds to a progression from the conceptual primitives level in figure 2 to the behavior representation level.

## Situation Graphs and Situation Graph Trees

The system has to incorporate, therefore, a priori knowledge about which vehicle maneuvers can be concatenated—and under which conditions—into admissible sequences of occurrences. Such knowledge about vehicular behavior is represented internally as a situation graph formed by situation nodes connected by prediction edges; see figure 14. A *situation node* combines a state representation scheme—expressed as a conjunction of fuzzy metric-temporal logic predicates—and an action scheme. The action scheme indicates the action open to the agent provided the state scheme can be instantiated from observations related to this agent. In other words, the (time-indexed) results imported from the core computer vision subsystem are converted into a model-theoretic set of individuals that are used to interpret the logic formulas representing the a priori knowledge about temporal developments in the depicted scene.

At each consecutive point in time (that is, for each half-frame), the inference engine activated for the interpretation task selects the highest-prioritized prediction link to attempt to interpret the state representation scheme of the successor node. If such an attempt fails, it is repeated iteratively following successively lower prioritized links until either a state representation scheme of a successor situation can be instantiated, or the list of possible successor situations has been exhausted.

The next steps performed by the inference engine depend on the position of the last successfully instantiated situation node. In principle, the number of predicates to be checked during an instantiation attempt can become rather large, with a high probability that most predicates remain true at the (frame)time succeeding the last point in time with a successfully instantiated situation node. Thus, it appears advantageous to organize situation nodes not only according to their temporal concatenation but also according to a degree of conceptual refinement. A more abstract situation node can be refined into a less abstract representation by either the addition of new predicates to the state representation scheme (specialization) or the temporal decomposition into a subsequence of situation nodes referring to a more detailed state representation scheme. The (most) abstract situation node *cross* (for "cross an intersection") in figure 14 is refined into a subgraph constituted by a concatenation of three situation nodes, namely, (1) drive_to_intersection, (2) drive_on_intersection, and (3) leave_intersection. Such a refinement can take place recursively, as illustrated by figure 14. A *subordinate situation node*, that is, a situation node in a graph that refines a more abstract situation node, inherits all predicates from its superordinate situation nodes. These predicates are included in the set of logic formulas constituting the state representation scheme of the subordinate situation node. This hierarchical organization of a situation graph greatly simplifies the design and maintenance of more complex behavior representations: A situation graph is turned into a special case of a directed hypergraph, namely, into a situation graph tree.

The tree property is important for the situation graph tree traversation rule followed by the inference engine. If a situation node has successfully been instantiated, it is attempted next to instantiate the entry node of its subordinate situation graph (if there is one): This rule aims at reaching the most detailed situation node compatible with the currently prevailing facts. If an attempt to instantiate a successor node in a subgraph fails at some later point in time, the situation graph tree traversation algorithm returns to the uniquely specified more abstract situation node and attempts to continue from there. The unsatisfiability of a more refined state representation scheme does not exclude that a more abstract scheme can still be satisfied by current observations. Because of this rule, a more general (for example, emergency) reaction can still be possible even if the originally anticipated detailed sequence of actions must be ruled out because their conditions—namely, the satisfaction of all predicates required by the state representation scheme of the more detailed situation nodes—can no longer be confirmed.

A path through a directed situation graph tree implies that the agent executes the actions specified in the most detailed situation node reached at each point in time during traversal; that is, such a path implies the behavior associated with the concatenation of actions encountered along such a path.

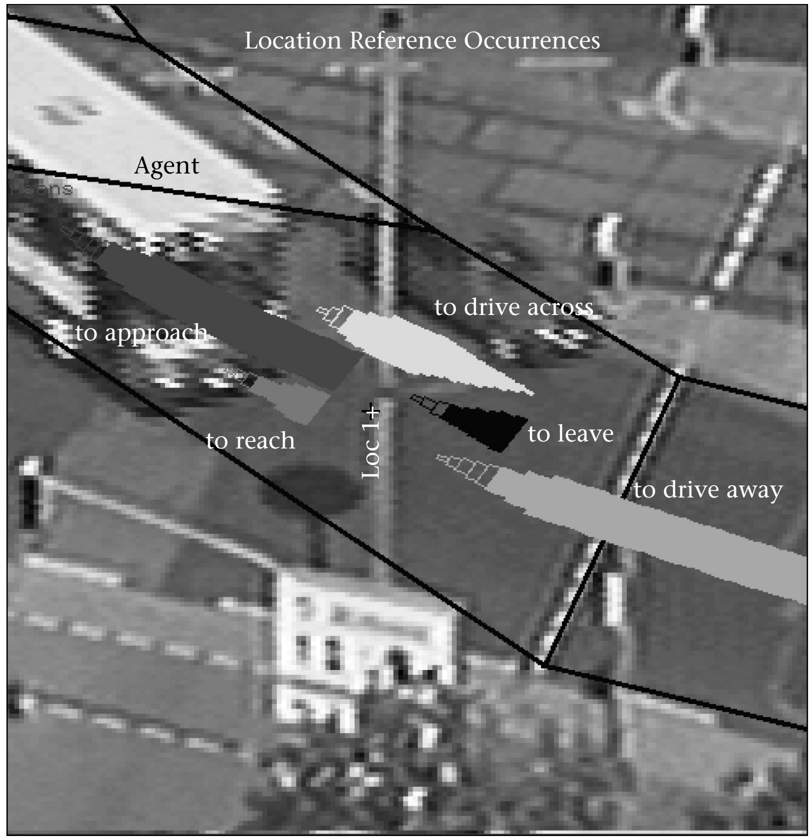


*Figure 13. Part of the Bus Trajectory from Figure 8, Reproduced as a Set of Shaded Ribbons.*

The width of each ribbon reflects the degree of validity with which an occurrence (indicated next to each ribbon) from table 1 describes a segment of the bus trajectory. The + sign in the image center indicates the location on the road plane to which the occurrence definitions refer. Some ribbons have been shifted sideways relative to the original bus trajectory to avoid overlaps. A conceptual description has to be associated consecutively for a minimum number of frames (indicated by hollow ribbon sections) until it definitely becomes accepted. Note how the different associations either terminate abruptly or peter out as time goes on, depending on the particular occurrence definition.

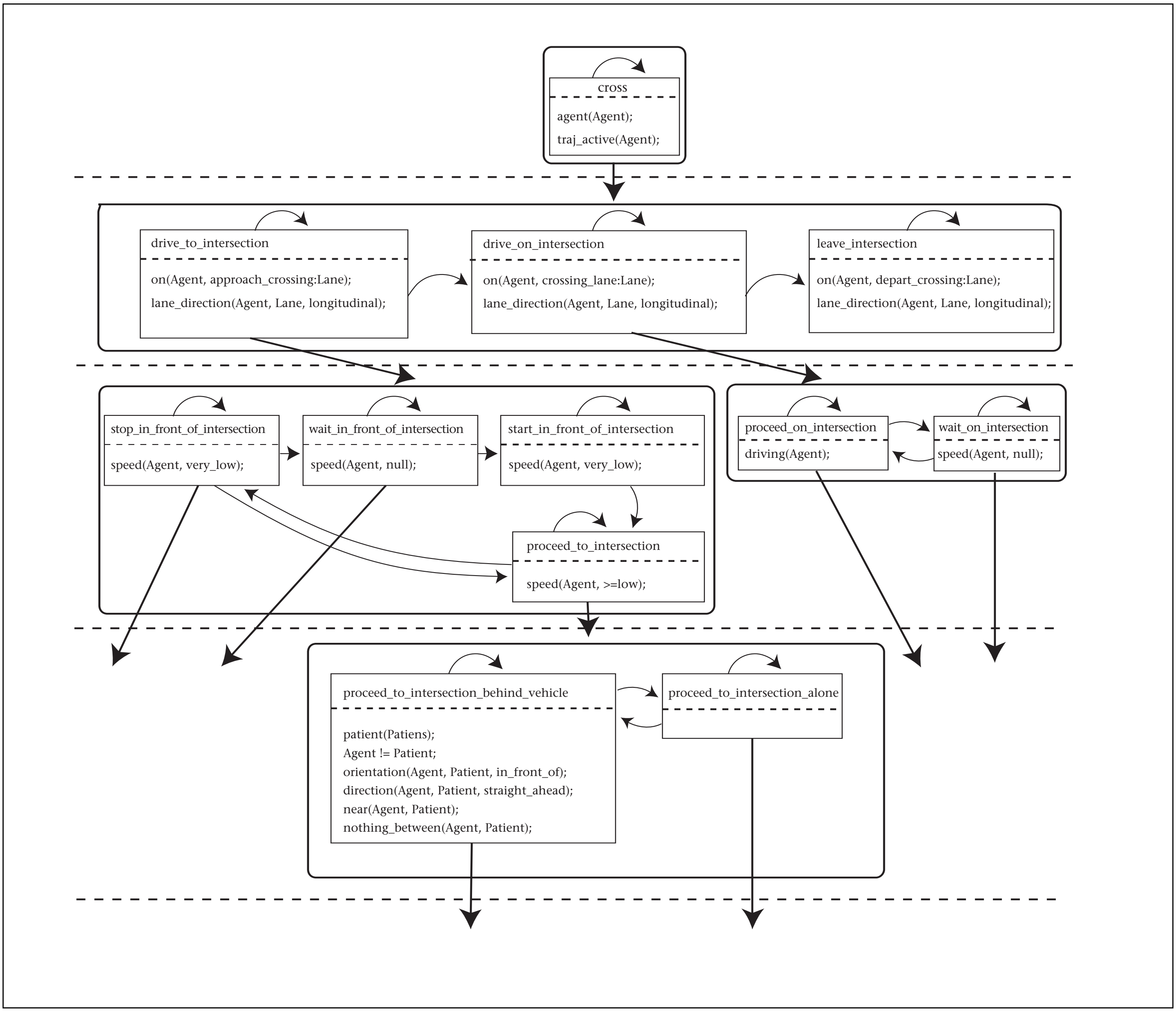


*Figure 14. Top Four Levels from a Situation Graph Tree.*

The top four levels represent a situation *cross* (an intersection) and its refinement into a subordinate situation graph constituted by the concatenation of situation nodes for drive_to_intersection, drive_on_intersection, and leave_intersection. The first two situation nodes within this subordinate situation graph have been refined further. The action part of situation nodes has been omitted for simplicity. (Adapted from Haag and Nagel [2000] © 2000 Elsevier, reproduced with permission).

## Feedback for Tracking Via the Behavioral Representation Level

The exploitation of a priori knowledge incorporated into a situation graph tree will be illustrated by an approach to cope with the behavior of vehicles which changes while they are occluded. The right panel of figure 15 shows (interactively generated) 3D polyhedral models of a tree, several masts, and a large traffic sign, the latter in the upper left quadrant of the scene depicted by the left panel. This allows, for example, to determine the degree of occlusion—see figure 16—of vehicles while they pass behind this traffic sign.

Figure 17 shows enlargements of the image area cropped within the square window indicated in the upper left quadrant of figure 15 for three different frames of a subsequence. During the initial part of this subsequence, the


*Figure 15. Three-Dimensional Models for Stationary Objects in the Scene.*

A frame from a sequence recorded at another Karlsruhe intersection is shown in the left panel (courtesy M. Haag). The small square in the left upper quadrant indicates where a window has been cropped that is used after enlargement in figure 17. The right panel represents a somewhat enlarged section from the left panel, overlaid by polyhedral models for a road sign mounted at a separately modeled mast, for a tree, and for various other masts carrying traffic signs, traffic lights, or lamps. (From Haag and Nagel [1999] © 1999 Kluwer, reproduced with permission).

larger bright van passed behind the traffic sign—see the left (dashed) occlusion curve in figure 16—and then slowed down in front of the red traffic light until it had come to a complete stop. The smaller vehicle following the van, a fastback, was occluded somewhat later by the same sign (right occlusion curve in figure 16). This fastback began to slow down immediately prior to occlusion and came to a full stop shortly afterward when it was completely occluded. It only began to move again after the van had started driving when the traffic light in front of it had switched to green.

As soon as the degree of occlusion exceeds a threshold of about 70 percent of the projected model area, the state update occurs no longer on the basis of edge element and optical flow data but instead relies on numeric input derived from the behavior predicted on the basis of the situation graph. Thus, researchers can take into account that the fastback in figure 17 brakes and comes to a full stop to avoid crashing into the van in front of it in the same lane. The fastback will begin to accelerate again only after the preceding van starts driving, and a safety distance has built up that allows the fastback to follow without danger. To our knowledge, this is the first example of a Kalman filter–based vehicle-tracking process being temporarily controlled not by data but by a fuzzy metric-temporal logic inference engine (Haag [1998]; Haag and Nagel [1998]).

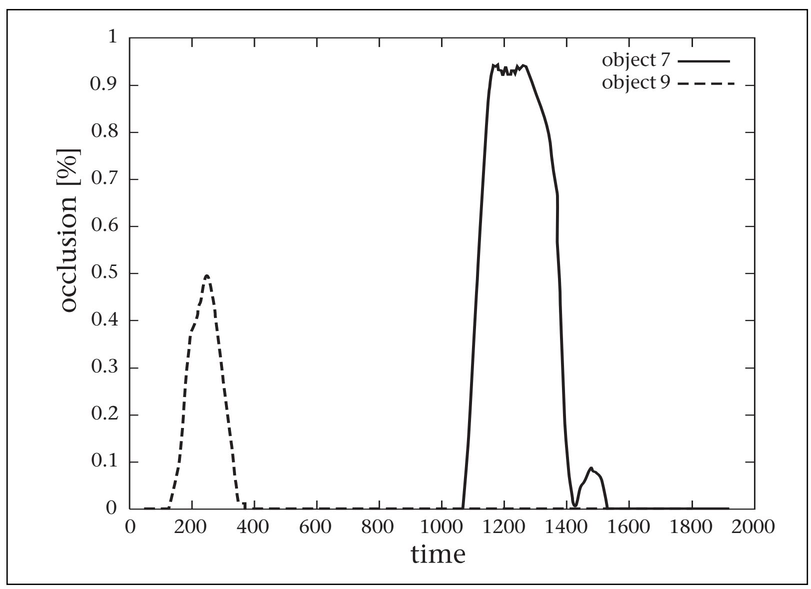


*Figure 16. Degree of Occlusion of Vehicles by Stationary Scene Objects.*

Degree of occlusion of the vehicles marked in figure 17 by their models overlaid in a dashed line (the van, passing behind the traffic sign first) and a solid line (the fastback, following shortly thereafter). Note that the van spent less time being occluded by the traffic sign because it still drove, whereas the fastback came to a full stop behind the traffic sign to avoid crashing into the van in front of it in the same lane. The occlusion of the fastback began to diminish when it emerged from behind the traffic sign once the van had started to drive again. The small occlusion extremum immediately following the large solid one is the result of pose corrections (see the kick in the fastback trajectory superimposed on the lowest panel in figure 17) caused by the tracking process once at least about 30 percent of the fastback had emerged from the occlusion according to occlusion reasoning based on an evaluation of the relative 3D geometric relations between camera, traffic sign, and fastback. (Courtesy M. Haag, adapted from Haag and Nagel [1998] © 1998 Springer-Verlag, reproduced with permission.)

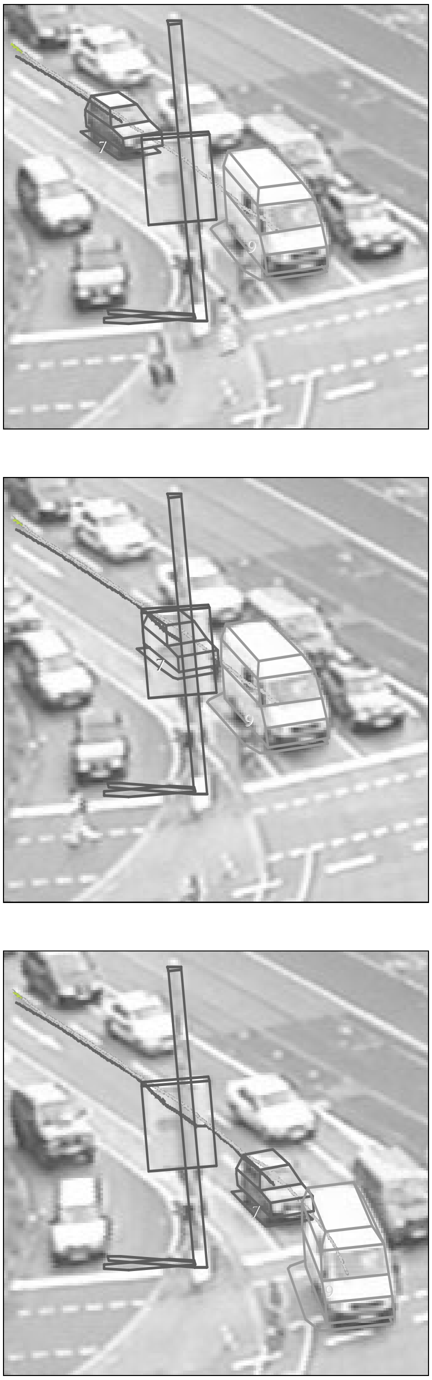


*Figure 17. Feedback to the Geometric Tracking Process Using the Behavioral Representation Level.*

The left column of panels illustrates three particular states from a subsequence during which the bright van, followed by the small fastback, passes behind the large traffic sign. (Courtesy M. Haag, adapted from Haag and Nagel [1998] © 1998 Springer-Verlag, reproduced with permission).

The center column shows the state representation and the action scheme for the situation nodes *approach_preceding_vehicle, wait_in_front_of_intersection_behind_vehicle,* and *start_up_behind_preceding_vehicle.* The right column shows the instantiation of the corresponding schemes in the center column, now giving the identifier for the individuals that the inference engine substituted for the logic variables during the interpretation of the predicates shown in the state representation schemes in the center column. The identifiers *obj_7* and *obj_9* are reproduced in the same shade as the vehicles in the left column to which they refer. Note that predicates shown explicitly as part of the state representation schemes in the center column form only a fraction of the set of logic formulas constituting the entire state representation scheme to be satisfied: Predicates belonging to the state representation schemes of superordinate situation nodes have been suppressed here for clarity (see, too, Haag [1998]).


*Figure 18. Two Representative Image Frames (Left: 400; Right: 875) from a Road Traffic Intersection Sequence That Makes Up a Total of 2320 Frames.*

The frame rate of this sequence is equivalent to a sampling rate of 50 frames a second; that is, the entire image sequence covers slightly more than 45 seconds of road traffic. (Adapted from Gerber, Nagel, and Schreiber [2002] © 2002 IOS Press, reproduced with permission).

## Text Generation from Conceptual Representations

Given a conceptual representation of temporal developments in the form of a time-attributed set of logic formulas, it appears natural to look for a method that relates natural language text to logic. If it should become possible to invert such a method to "turn the meaning" of logic formulas by algorithmic means into a natural language text, a road would open along which the content of a video sequence could be converted into a natural language textual description—a systematic realization of genuine multimedia.

The *discourse representation theory* (see, for example, Kamp and Reyle [1993]) treats a method that converts a natural language text into an internal representation—the so-called *discourse representation structure*—which is closely oriented toward predicate logic. In fact, the authors discuss conditions and algorithmic means by which a discourse representation structure can be transformed into a set of first-order–predicate logic formulas. We thus studied this formalism to obtain a systematic approach toward transforming the fuzzy metric-temporal logic representation of vehicular behavior extracted from video sequences into natural language descriptions (see, for example, Gerber and Nagel [1998]). Space does not permit me to go into details. It turned out that no modules were readily available yet that inverted the text-to-logic branch. To bridge this gap, Gerber developed a module with the aim of providing at least a rudimentary text-generation component based on partial instantiations of situation graph trees (Gerber 2000). This module has primarily been used with data pertaining to the behavior of single vehicles.

The systematic basis of this approach enabled us to extend it to generate simple descriptions of the formation and dissolution of vehicle queues (Gerber, Nagel, and Schreiber 2002). Figure 18 shows two representative image frames from a long video sequence illustrating a traffic queue that built up in front of a red traffic light (left panel) and began to dissolve shortly after the traffic light switched to green.

In this case, the polygonal lane structure referred to earlier has been used to select all vehicles approaching and crossing the depicted intersection on the left lane marked in the right panel of figure 19. Conceptual concatenation of the different lane segments marked in figure 19 by heavy boundary lines has been performed by the same inference engine that evaluates the situation graph trees. A sample of the algorithmically generated—rather simple—textual description is given in figure 20.

It will be instructive to reflect on the reasons for the increasingly objectionable monotonous formulations found in figure 20. First, each vehicle has been mentioned to provide a check for vehicles that might have been lost somewhere along the processing chain. In addition, the system still lacks suitable linguistic abstraction facilities that would allow replacing the detailed recounting by a phrase characterizing


*Figure 19. Lane Structure Exploited for Generation of Traffic Queue Descriptions.*

The left panel shows an image frame from the sequence illustrated in figure 18, overlaid by the lane model for this intersection. One lane segment of the incoming lane entering the field of view from the top left corner is marked by heavy boundary lines. The right panel exhibits the same image, but this time, the intersection crossing segment and the outgoing segment of this same lane are also marked by heavy boundary lines. In addition, the lane identifiers for the three segments of this lane are given. (Adapted from Gerber, Nagel, and Schreiber [2002] © 2002 IOS Press, reproduced with permission.)

an entire set of repetitive developments. Another stylistic tool used by humans in such a case, namely, a variation in the manner by which the different vehicles are referenced, is also not yet available; it necessitates providing the required identifying properties for individual vehicles (such as their color or shape) or spatiotemporal relations (such as *beside, slightly to the left of,* and *the one after it*). These last examples are particularly interesting because they illustrate the close interaction between fundamental competences in the geometric and conceptual subsystems.

## Discussion and Conclusions

The idea of having an algorithm convert a video sequence into natural language text has been pursued in our group for several decades now (Nagel 1988, 1977). As the exposition in the preceding sections illustrated, at least the semblance of a solution becomes amenable to systematic investigations. It might be interesting to ponder for a moment why this development took so long.

### The Interaction between Available Computing Resources and Algorithm Development

The development sketched here suggests a combination of causes. It appears that mere change detection provides intermediate results that are too brittle for a long and complicated sequence of additional evaluation steps. Obviously, computing resources needed to become available at sufficiently low prices that university laboratories could afford to test computationally more expensive algorithms, in particular, ones based on optical flow estimation.

It took some time until the switch from picture domain tracking to model-based 3D scene domain tracking became a seriously investigated alternative. The systematic introduction of a priori knowledge in the form of 3D body and motion models, in combination with the adaptation of Kalman filtering to this kind of tracking task, provided a quantum jump in robustness. It allowed researchers to experiment with the evaluation of longer video sequences such that it became possible to study not only vehicle motion but also vehicle maneuvers.

A next big step forward became possible when early feature-based optical flow estimates could be replaced by gradient-based approaches that provided a much denser optical flow field to work with. The combination of edge element and optical flow matching during the state-update cycle then increased the tracking precision and stability to a point where one could begin to think about the investigation not only of single maneuvers but also of entire maneuver sequences and, thus, about the investigation of vehicle behavior.

At about this point in the development, ear-

"**Obj_2** entered the lane. Later **obj_6** entered the lane. The vehicles formed a pair.
Later **obj_8** entered the lane. In the meantime the vehicles formed a queue. **Obj_8** was the last vehicle of the queue. O**bj_2** was the head of the queue.
In the meantime **obj_9** entered the lane. It was the last vehicle of the queue.
In the meantime **obj_12** entered the lane. It was the last vehicle of the queue.
It left the queue. In the meantime **obj_9** was the last vehicle of the queue.
In the meantime **obj_15** entered the lane. It was the last vehicle of the queue.
In the meantime **obj_8** left the queue.
In the meantime **obj_25** entered the lane. It was the last vehicle of the queue.
In the meantime **obj_27** entered the lane. It was the last vehicle of the queue.
In the meantime **obj_2** left the queue.
In the meantime **obj_6** was the head of the queue. It left the queue.
In the meantime **obj_9** was the head of the queue.
In the meantime **obj_29** entered the lane. It was the last vehicle of the queue.
In the meantime **obj_9** left the queue.
In the meantime **obj_15** was the head of the queue. It left the queue.
In the meantime **obj_25** was the head of the queue. The remaining vehicles formed a pair.
**Obj_25** left the lane.
Later **obj_27** left the lane. In the meantime **obj_29** remained as single vehicle."

*Figure 20. Output Text Generated for the Vehicle Queues Illustrated in Figure 18.*

(Adapted from Gerber, Nagel, and Schreiber [2002] © 2002 IOS Press, reproduced with permission.)

ly attempts to associate conceptual descriptions with geometric results could be tested systematically enough to lay open less frequently occurring deficiencies. As a consequence, a more systematic approach based on formal methods became imperative unless one runs the risk of being swamped by difficult to analyze deficiencies of ad hoc approaches.

Chaining all required processing steps from video recording through to the algorithmic generation of natural language textual descriptions now offers the chance to systematically assess an overall approach to detect and remove the most disturbing bottlenecks.

## On Exercises and Research Problems[1]

The desire to analyze more complex temporal developments necessitates the ability to process long image sequences without gross failures. Even rarely occurring failures can interrupt the provision of correct geometric results to the inference processes involved and thereby prevent the generation of an appropriate description of the spatiotemporal development in the scene.

According to our current experience, some of these bottlenecks still seem related to early processing stages, in particular, to the detection of vehicles to be tracked and to a robust initialization of a model-based tracking process. Because this is a highly nonlinear process, a problematical initialization can have repercussions much later on, both during geometric tracking and during subsequent treatment of tracking results at the conceptual level. It can be difficult to trace back the root causes for such problems.

Apart from tracing down erroneously conceived or implemented algorithmic details, parameter tuning and provision of appropriate models can turn into a potential bottleneck. There have been efforts already to continuously estimate whether the current illumination is *directed* (in our case, bright sunshine) or *diffuse* (the sky being covered by clouds) (see Leuck and Nagel 2001). Another effort addressed the estimation of lane structures from image sequences (Mück 2000; Mück, Nagel, and Middendorf 2000). Further details about these and other related questions were reported earlier.

A third problem area concerns the provision of vehicle models. To this point, we have used mostly standard models for vehicles that can be observed most frequently at inner-city intersections, namely, sedans, fastbacks, and station wagons. The problems in these cases are more related to automatically estimating the appropriate length, width, and height parameters. This is a kind of hen-and-egg problem: Unless vehicles can be tracked reliably, parameter estimation becomes unreliable, but reliable vehicle parameters are essential for tracking a vehicle through difficult traffic situations (such as diminished contrast with respect to foreground and background, nontrivial occlusion by stationary components of the scene or by other vehicles). Busses for inner-city public transport have mostly been standardized in Germany with available 3D model data, so this did not generate great difficulties. All other vehicle types have had to be modeled interactively.

Readers might have noticed that pedestrians and bicycle riders have been excluded thus far from the discourse domain. Results about the detection, tracking, and description of the behavior of persons have been reported by others, for example, Remagnino, Tan, and Baker (1998) and Rota and Thonnat (2000). Given the gamut of problems hinted at in the preceding sections, it appears important to emphasize robustness in a somewhat restricted discourse domain over attempts to admit developments in a more broadly defined domain.

## Conclusions

This contribution outlined an overall system concept regarding a—nonexclusive—understanding of what constitutes a cognitive vision system. It aimed first to indicate that such experimental approaches have become feasible. An equally important aim was to illustrate where problem areas developed; why they became hot spots; and which methodological approach helped to defuse them, at least for a time.

The particulars presented do not imply that the system approach outlined here is the only or the best one. Alternative approaches toward tracking and describing road traffic have been pursued increasingly over the past decade; see, for example, Buxton and Gong (1995); Chella, Frixione, and Gaglio (2000); Dance, Caelli, and Liu (1995); Howarth and Buxton (2000); Intille and Bobick (1999); Kojima, Tamura, and Fukunaga (2002); Neumann (1989); and Pece and Worrall 2002. It appears too early to decide which (combination of) approach(es) offers the greatest promise, given well-defined boundary conditions regarding specifics of the discourse domain (required success and false alarm rates, illumination conditions, admissible vehicle types, field of view to be covered, and so on) and the task. A more principled discussion could concentrate on whether one should use a fuzzy metric-temporal logic or Bayesian (belief) networks. Based on researchers' experience, much larger experimental series than those used to date will likely have to be evaluated to achieve reliable results.

A similar problem is likely to come up in the future if details of natural language text generation have to be judged. Support for this hypothesis can be found in Sparck Jones and Galliers (1995), where input from video sequences had not even been considered!

References to literature beyond the computer vision discipline can be followed up with other links,[2] for example, to spatial reasoning. As the short discussion in connection with figure 20 illustrated, areas in AI that have developed largely without intensive contact with computer vision increasingly gain interest for CVSs. Clearly, the evaluation of image sequences has reached a degree of maturity that allows the study of the conversion of geometric tracking results into conceptual representations and beyond. The hints regarding remaining difficulties can be looked on as bad or good news, depending on the age and stamina of the reader.

### Acknowledgments

My thanks go to Henrik I. Christensen whose patient prodding prevented me repeatedly from giving up on finishing this article. During the many years devoted to the research outlined here, I had the pleasure of cooperating with a large number of students who stimulated me again and again with their interest and results. Unfortunately, I could not mention all who contributed. Among those not explicitly quoted, I just want to mention W. Krüger for his work on early versions of situation formalism and K. Daniilidis for extensive discussions regarding Kalman filtering and optical flow estimation. The research mentioned here would not have been possible without repeated support by the Deutsche Forschungsgemeinschaft and the European Union (IST-2000-29404).

### Note

1. Richard Bellman is said to have appended a section entitled "Exercises and Research Problems" to one of his books on dynamic programming. When a colleague remarked to him that he had forgotten to indicate which problems were exercises and which ones were research problems, Bellman reportedly answered, "If you can solve it, it was an exercise; otherwise it's a research problem." Unfortunately, I cannot give an exact reference for this definition.

2. Nagel, H.-H. 2001. Towards a Cognitive Vision Sys-

tem. http://kogs. iaks. uni-karlsruhe. de/CogViSys/ kogs_CogViSys_homepage_V22_lin. pdf.

## References

Burkhardt, H., and Neumann, B., eds. 1998. *Computer Vision—ECCV98.* Lecture Notes in Computer Science 1407. Berlin: Springer-Verlag.

Buxton, H., and Gong, S. 1995. Visual Surveillance in a Dynamic and Uncertain World. *Artificial Intelligence* 78(1–2): 431–459.

Chella, A.; Frixione, M.; and Gaglio, S. 2000. Understanding Dynamic Scenes. *Artificial Intelligence* 123(1–2): 89–132.

Dance, S.; Caelli, T.; and Liu, Z.-Q. 1995. *Picture Interpretation–A Symbolic Approach.* Machine Perception and Artificial Intelligence Volume 20. Singapore: World Scientific.

Gerber, R. 2000. Natürlichsprachliche Beschreibungen von Strassenverkehrsszenen durch Bildfolgeauswertung (Natural Language Description of Road Traffic Scenes Based on Image Sequence Evaluation). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH). www.ubka.uni-karlsruhe.de/cgibin/psview?document=2000/informatik/8.

Gerber, R., and Nagel, H.-H. 1998. (Mis-?)Using DRT for Generation of Natural Language Text from Image Sequences. Paper presented at the Fifth European Conference on Computer Vision, 2–6 June, Freiburg, Germany.

Gerber, R., and Nagel, H.-H. 2002. "Occurrence" Extraction from Image Sequences of Road Traffic Scenes. Paper presented at the Workshop on Cognitive Vision, 19–20 September, Zurich, Switzerland.

Gerber, R.; Nagel, H.-H.; and Schreiber, H. 2002. Deriving Textual Descriptions of Road Traffic Queues from Video Sequences. In *Proceedings of the Fifteenth European Conference on Artificial Intelligence,* 736–740, ed. F. van Harmelen. Amsterdam, The Netherlands: IOS.

Haag, M. 1998. Bildfolgenauswertung zur Erkennung der Absichten von Strassenverkehrsteilnehmern (Evaluation of Image Sequences for the Recognition of Intentions of Road Traffic Participants). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH).

Haag, M., and Nagel, H.-H. 1998. "Begriffliche Rückkopplung" zur Behandlung von Verdeckungssituationen in der Bildfolgenauswertung von Strassenverkehrsszenen (Conceptual Feedback for the Treatment of Occlusion Situations in Road Traffic Scenes by Evaluation of Image Sequences). In *Informatik '98–Informatik zwischen Bild und Sprache,* eds. J. Dassow and R. Kruse, 13–22. GI Informatik aktuell, Berlin: Springer-Verlag.

Haag, M., and Nagel, H.-H. 1999. Combination of Edge Element and Optical Flow Estimates for 3D-Model-Based Vehicle Tracking in Traffic Image Sequences. *International Journal of Computer Vision* 35(3): 295–319.

Haag, M., and Nagel, H.-H. 2000. Incremental Recognition of Traffic Situations from Video Image Sequences. *Image and Vision Computing* 18(2): 137–153.

Howarth, R. J., and Buxton, H. 2000. Conceptual Descriptions from Monitoring and Watching Image Sequences. *Image and Vision Computing* 18(2): 105–135.

Intille, St., and Bobick, A. 1999. Visual Recognition of Multi-Agent Action Using Binary Temporal Relations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Volume 1, 56–62. Washington, D.C.: IEEE Computer Society.

Kamp, H., and Reyle, U. 1993. *From Discourse to Logic.* Dordrecht, The Netherlands: Kluwer Academic.

Kanade, T. 1978. Region Segmentation: Signal versus Semantics. Paper presented at the Fourth International Joint Conference on Pattern Recognition, 7–10 November, Kyoto, Japan.

Kojima, A.; Tamura, T.; and Fukunaga, K. 2002. Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. *International Journal of Computer Vision* 50(2): 171–184.

Koller, D. 1992. Detektion, Verfolgung und Klassifikation bewegter Objekte in monokularen Bildfolgen am Beispiel von Strassenverkehrsszenen (Detection, Tracking, and Classification of Moving Objects in Monocular Image Sequences Using Road Traffic Scenes as Examples). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH).

Koller, D.; Heinze, N.; and Nagel, H.-H. 1991. Algorithmic Characterization of Vehicle Trajectories from Image Sequences by Motion Verbs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 90–95. Washington, D.C.: IEEE Computer Society.

Koller, D.; Daniilidis, K.; and Nagel, H.-H. 1993. Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes. *International Journal of Computer Vision* 10(3): 257–281.

Kollnig, H. 1995. Ermittlung von Verkehrsgeschehen durch Bildfolgenauswertung (Determination of Traffic Occurrences Based on Image Sequence Evaluation). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH).

Kollnig, H.; Nagel, H.-H.; and Otte, M. 1994. Association of Motion Verbs with Vehicle Movements Extracted from Dense Optical Flow Fields. In *Proceedings of the Third European Conference on Computer Vision,* 338–347. Lecture Notes in Computer Science 801. Berlin: Springer-Verlag.

Kollnig, H.; Leuck, H.; and Nagel, H.-H. 1995. Detektion und Verfolgung von Fahrzeugen in Strassenverkehrsszenen: Systematische Bewertung und Steigerung der Systemleistung (Detection and Tracking of Vehicles in Road Traffic Scenes—Systematic Assessment and Improvement of System Performance). In *DAGM Symposium Bielefeld,* eds. G. Sagerer, S. Posch, and F. Kummert, 303–310. Berlin: Springer-Verlag.

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

Leuck, H., and Nagel, H.-H. 2001. Model-Based Initialization of Vehicle Tracking: Dependency on Illumination. In *Proceedings of the Eighth International*

*Conference on Computer Vision, Volume 1,* 309–314. Washington, D.C.: IEEE Computer Society.

Lowe, D. G. 1991. Fitting Parameterized Three-Dimensional Models to Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(5): 441–450.

Mück, K. 2000. Rechnergestützte Erkennung und Beschreibung innerstädtischer Strassenkreuzungen (Computer-Based Recognition and Description of Inner-City Road Intersections). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH). www.ubka.uni-karlsruhec.de/cgibin/psview?-document=2000/informatik/8.

Mück, K.; Nagel, H.-H.; and Middendorf, M. 2000. Data-Driven Extraction of Curved Intersection Lanemarks from Road Traffic Image Sequences. In *Proceedings of the Sixth European Conference on Computer Vision, Volume 2,* ed. D. Vernon, 411–427. Lecture Notes in Computer Science 1843. Berlin: Springer-Verlag.

Nagel, H.-H. 2000. Image Sequence Evaluation: 30 Years and Still Going Strong. In *Proceedings of the Fifteenth International Conference on Pattern Recognition, Volume 1,* eds. A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alquézar, J. -O. Eklundh, and Y. Aloimonos, 149–158. Washington, D.C.: IEEE Computer Society.

Nagel, H.-H. 1988. From Image Sequences toward Conceptual Descriptions. *Image and Vision Computing* 6(2): 59–74.

Nagel, H.-H. 1977. Analyzing Sequences of TV Frames: System Design Considerations. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 626. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Neumann, B. 1989. Natural Language Description of Time-Varying Scenes. In *Semantic Structures–Advances in Natural Language Processing,* ed. D. Waltz, 167–206. Hillsdale, N.J.: Lawrence Erlbaum.

Otte, M. 1994. Extraktion von linienförmigen Merkmalen und Ermittlung des optischen Flusses mit seinen Ableitungen aus Voll- und Halbbildfolgen (Extraction of Linelike Features and Determination of Optical Flow with Its Derivatives on Full-Frame and Half-Frame Image Sequences). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH).

Otte, M., and Nagel, H.-H. 1995. Estimation of Optical Flow Based on Higher-Order Spatiotemporal Derivatives in Interlaced and Non-Interlaced Image Sequences. *Artificial Intelligence* 78(1–2): 5–43.

Otto, P. 1990. Integration von Systemkomponenten zur Berechnung von Bewegungsverben aus Bildfolgen (Integration of System Components for the Computation of Motion Verbs from Image Sequences). Diplomarbeit, Fakultät für Informatik der Universität Karlsruhe (TH).

Pece, A. E., and Worrall, A. D. 2002. Tracking with the EM Contour Algorithm. In *Proceedings of the Seventh European Conference on Computer Vision 2002 (ECCV2002),* 3–17. Lecture Notes in Computer Science 2350. Berlin: Springer-Verlag.

Remagnino, P.; Tan, T.; and Baker, K. 1998. Agent-Oriented Annotation in Model-Based Visual Surveillance. In *Proceedings of the Sixth International Conference on Computer Vision,* 857–862. New Delhi, India: Narosa Publishing.

Rota, N. A., and Thonnat, M. 2000. Activity Recognition from Video Sequences Using Declarative Models. Paper presented at the European Conference on Artificial Intelligence, 20–25 August, Berlin, Germany.

Russell, S., and Norvig, P. 1995. *Artificial Intelligence–A Modern Approach.* Upper Saddle River, N.J.: Prentice-Hall.

Schäfer, K. H. 1996. Unscharfe zeitlogische Modellierung von Situationen und Handlungen in Bildfolgenauswertung und Robotik (Fuzzy Temporal Logic Models of Situations and Actions in Image Sequence Evaluation and Robotics). Doctoral diss., Fakultät für Informatik der Universität Karlsruhe (TH).

Sparck Jones, K., and Galliers, J. R. 1995. *Evaluating Natural Language Processing Systems.* Lecture Notes in Artificial Intelligence 1083. Berlin: Springer-Verlag.

Sung, C. -K., and Zimmermann, G. 1986. Detektion und Verfolgung mehrerer Objekte in Bildfolgen (Detection and Tracking of Several Objects in Image Sequences). In *Proceedings of the DAGM Symposium Paderborn,* ed. G. Hartmann, 181–184. Informatik-Fachberichte Volume 125. Berlin: Springer-Verlag.

Zimmermann, G., and Kories, R. 1984. Eine Familie von Bildmerkmalen für die Bewegungsbestimmung in Bildfolgen (A Family of Image Features for Motion Determination in Image Sequences). In *Tagungsband DAGM/OAGM Symposium Graz,* ed. W. Kropatsch, 147–153. Informatik-Fachberichte Volume 87. Berlin: Springer-Verlag.

**Hans-Hellmut Nagel** obtained his doctoral degree in physics from the Universität Bonn, Germany, in 1964. Following a postdoctoral period at the Massachusetts Institute of Technology, he moved from automatic bubble-chamber film measurement into image sequence evaluation when he was appointed o. Professor für Informatik at the Universität Hamburg, Germany, in 1971. In 1983, he became codirector of the Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB) at Karlsruhe, Germany, in a joint appointment as o. Professor für Informatik at the Universität Karlsruhe (TH). He served as chairman of the International Conference on Computer Vision 1993 in Berlin, Germany. In 2000, he retired from the IITB and became professor emeritus in April 2003. He is a member of AAAI, a fellow of IEEE, a fellow of the International Association of Pattern Recognition, and a fellow of the European Coordinating Committee on Artificial Intelligence. His e-mail address is nagel@iaks.uni-karlsruhe.de.