Tracking of Occluded Vehicles in Traffic Scenes

Thomas Frank[1], Michael Haag[1], Henner Kollnig[1], and Hans-Hellmut Nagel[1,2]

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

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

Abstract. Vehicles on downtown roads can be occluded by other vehicles or by stationary scene components such as traffic lights or road signs. After having recorded such a scene by a video camera, we noticed that the occlusion may disturb the detection and tracking of vehicles by previous versions of our computer vision approach. In this contribution we demonstrate how our image sequence analysis system can be improved by an explicit model–based recognition of 3D *occlusion* situations. Results obtained from real world image sequences recording gas station traffic as well as inner-city intersection traffic are presented.

1 Introduction

Occlusion causes truncated image features and, therefore, may mislead the object recognition process. In the sensitivity analysis of [Du *et al.* 93] occlusion is, therefore, treated as a source of error. [Du *et al.* 93] compare truncated image features with features which are contaminated by image noise or clutter. But unlike for the two latter effects, the analysis of the image sensing process is not adequate in order to cope with occlusion, nor is a pure 2D picture domain analysis. We model occlusion, therefore, in the 3D scene domain.

Our investigations are illustrated by an image sequence recording gas station traffic where intrinsically many occlusions occur, and a second one showing the traffic at a much frequented inner–city intersection.

2 Related Work

Related publications by other groups have been surveyed in [Sullivan 92; Koller *et al.* 93; Sullivan *et al.* 95; Cédras & Shah 95]. Inner-city traffic scenes are evaluated by [Koller *et al.* 93; Meyer & Bouthemy 94; Sullivan *et al.* 95]. Rather than repeating the bulk of this material, we concentrate on a small set of selected publications. In our previously published system (see [Koller *et al.* 93]), initial model hypotheses are generated using information extracted from optical flow fields in order to initialize a Kalman Filter tracking process. By projecting a hypothesized 3D polyhedral vehicle model into the image plane, 2D model edge segments are obtained which are matched to straight-line edge segments, so called data segments, extracted from the image. In [Koller *et al.* 94] the camera is mounted above the road on a bridge, looking down along the driving direction of the vehicles. Exploiting this special camera pose, [Koller *et al.* 94] can decide about partial occlusion of vehicles by comparing the vertical image coordinates of the candidate regions for moving vehicles.

[Meyer & Bouthemy 94] present a purely picture domain approach, without requiring three-dimensional models for the vehicles or for their motion. By comparing a predicted image region with a measured image region, [Meyer & Bouthemy 94] detect occlusions. In contrast to this geometric approach, [Toal & Buxton 92] used spatio-temporal reasoning to analyze occlusion behavior. In this latter approach, temporarily occluded vehicles are correctly relabeled after re-emerging from behind the occluding object rather than being treated as completely independent vehicles. So far it has been assumed by these authors, however, that vehicles had already been tracked and classified prior to the onset of occlusion.

[Sullivan 92] documents an approach which tracks one vehicle that is occluded by a lamp post — which is part of his scene model — as well as by another car which is tracked simultaneously. In contrast to his approach, we do not match edge segments but match synthetic model gradients to gray value gradients. The investigations of Baker, Sullivan, and coworkers are discussed in more detail in [Sullivan *et al.* 95] who report recent improvements of an important sub-step in the Reading approach, namely refinement of vehicle pose hypotheses.

[Dubuisson & Jain 95] present a 2D approach, where vehicles are detected by means of combining image subtraction and color segmentation techniques. In distinction to our approach, their 2D vehicle model requires images showing a side view of the vehicles.

3 Scene Model

We focus our investigations for the moment on the evaluation of the gas station scene depicted in Figure 1 where many occlusions by static scene objects occur. It appears not to be easy to track vehicle images properly, if significant parts of those images are occluded by stationary scene components. In this case, the matching process, which tries to minimize the difference between synthetic model gradients and the actual gray value gradient of the image at the estimated position of the vehicle, will associate at least parts of the model gradient field of the occluded vehicle to the gradient of the image of the occluding stationary scene object. Figure 3 (a) shows an enlarged section containing a vehicle which moves from right to left on the front lane of the gas station. Due to the occlusion by the advertising post and the bush, the superimposed vehicle model falls back compared to the actual position of the vehicle image. In this case, the synthetic gradient of the vehicle model has been associated with the gray value gradient of the stationary post.

In order to take potential occlusions caused by *static* scene components into account, we first had to extend a previous 2D road model to a 3D scene model of the depicted gas station.

The static objects which can cause occlusion of moving cars are represented by generic 3D surface models (see Figure 1). We thus need only one prototype model for each class of scene component, e.g., petrol pumps or bushes. The actual scene object is then represented by an instantiation of this generic model. Each model can be described by its corners (specified by length parameters relative to a fixed object coordinate system) and its surfaces (specified by the corners). The object coordinates have to be transformed to the common world coordinate system. With exception of the gas station roof which is slanted, the remaining 3D objects need only to be translated. The roof model is rotated and translated.

All mentioned 3D scene objects including parts of the gas station building are projected into the image, see Figure 1.

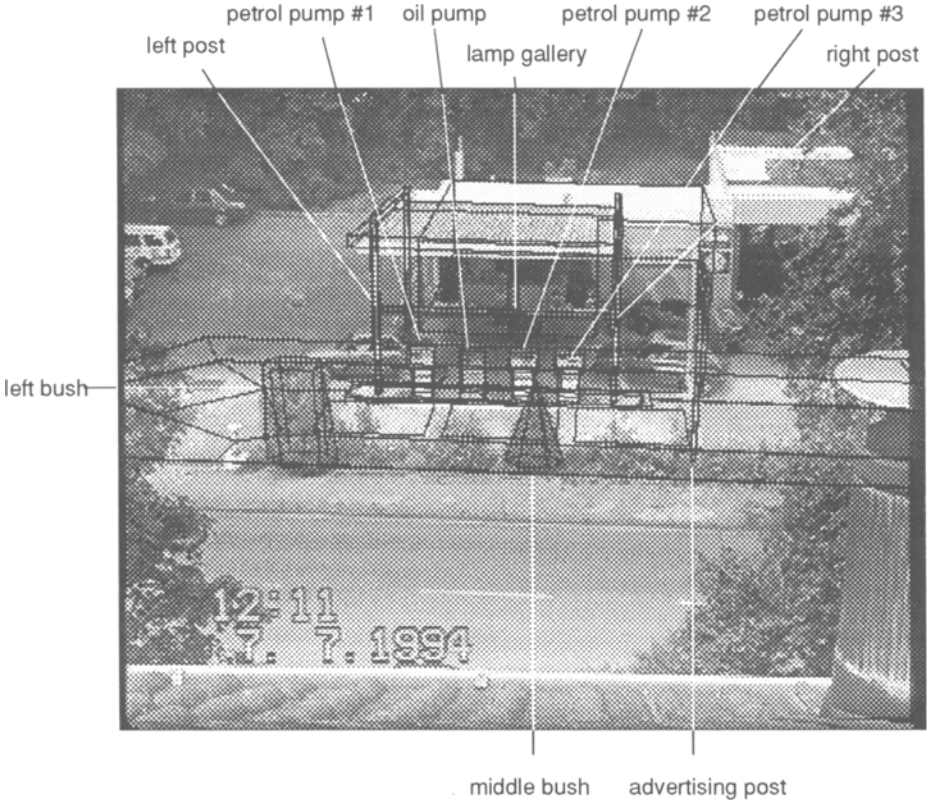

Fig. 1.: Frame #3605 of the gas station image with superimposed 3D scene model as well as road model.

By means of a pixel–oriented raytracer procedure, the distance of each modeled surface point from the projection center is computed, yielding a depth map of the observed scene. With this information, we are able to take the occluded parts of the car into consideration during our tracking process.

4 Occlusion Characterization

If we were able to *predict* the temporal development of a considered occlusion between two objects, we could exploit this knowledge in order to stabilize the tracking process. Therefore, we began to classify possibly occurring occlusions. This yielded occlusion situations which we arranged in a transition diagram in order to obtain all potential transitions between the occlusion situations due to the movements of the concerned objects. In order not to exceed the scope of this contribution, we demonstrate the approach with only a single occlusion predicate, namely *passively_decreasing_occlusion(X, Y, t)* which expresses the fact that at timepoint t the moving object Y *is occluded* by the stationary object X, so that the occlusion decreases with respect to time (see Figure 2):

$$\begin{aligned} \mathit{passively\_decreasing\_occlusion}(X, Y, t) = {} & \neg \mathit{moving}(X, t) \wedge \mathit{moving}(Y, t) \\ & \wedge\, \mathit{occlusion}(X, Y, t) \end{aligned}$$

The predicate *moving(X,t)* implies a non–zero velocity of Object X and *occlusion(X, Y, t)* holds if object X occludes object Y.

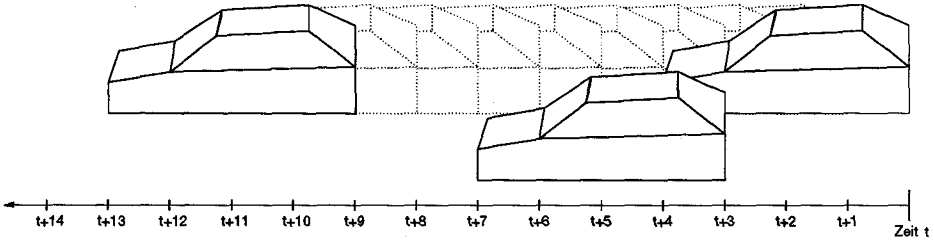

Fig. 2.: Object X is standing while object Y is passing. The occlusion decreases.

Similarly, the predicate actively_decreasing_occlusion(X,Y,t) would describe an occlusion situation in which the *occluded* object Y is standing while the *occluding* object X moves. Again, the occlusion will decrease.

Altogether, we found seven different occlusion predicates (composed out of four primitives) and 15 transition predicates which control the transitions between these occlusion situations. The (primitive and composed) predicates are modelled by means of *fuzzy sets* in order to abstract from quantitative details like *velocity.*

5 Results

5.1 Gas Station Image Sequence

Our first experiments focus on the gas station scene (see Figure 1). Without modeling the occurring occlusions, we had difficulties in tracking such vehicle images properly. For instance, the former tracking process lost the image of the vehicle shown in Figure 3 (a) due to the occlusion by the advertising post. After we introduced an explicit 3D model of the static scene components and excluded the occluded parts of the vehicle image from the matching process, we were able to track this vehicle without any problems (see Figure 3 (b)).

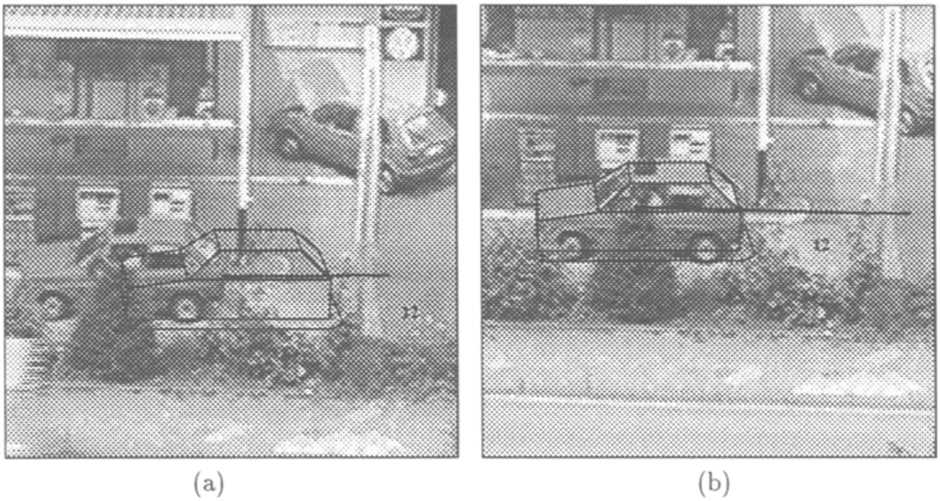

Fig. 3.: (a) Enlarged section showing a vehicle driving on the front lane of the gas station. The vehicle has been occluded by the advertising post on the right hand side and, therefore, could not be tracked properly by the tracking process. (b) Same vehicle as in (a), but now with the superimposed trajectory obtained by our automatic tracking process which took modeled 3D scene components, like the advertising post, into account.

A vehicle which is even more severely occluded is shown in Figure 4 (a). This car drives on the back lane of the gas station from the right to the left. On its way to the petrol pumps, it is occluded by the advertising post, by the right post carrying the roof, by two petrol pumps, by the oil pump, and by the vehicle driving on the front lane. The superimposed trajectory and the vehicle model in Figure 4 (b) show that the tracking process succeeds once we take the occluding scene components into account.

5.2 Downtown Image Sequence

In a second experiment, our approach is illustrated by a test image sequence of a much frequented multi-lane inner-city street intersection (see Figure 5).

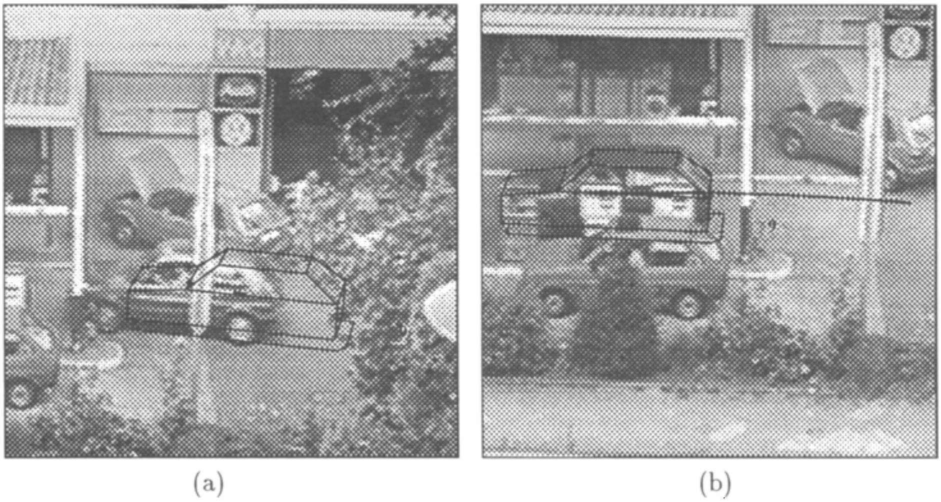

Fig. 4.: (a) Enlarged section of frame #314 of the gas station image sequence. Due to the occlusion, the tracking process associates parts of the vehicle model gradients to the stationary post and looses the moving vehicle image. (b) Enlarged section of frame #463 with superimposed trajectory for the vehicle driving on the back lane. Although the vehicle is heavily occluded, it can be tracked properly after the 3D models of the relevant scene components have been taken into account.

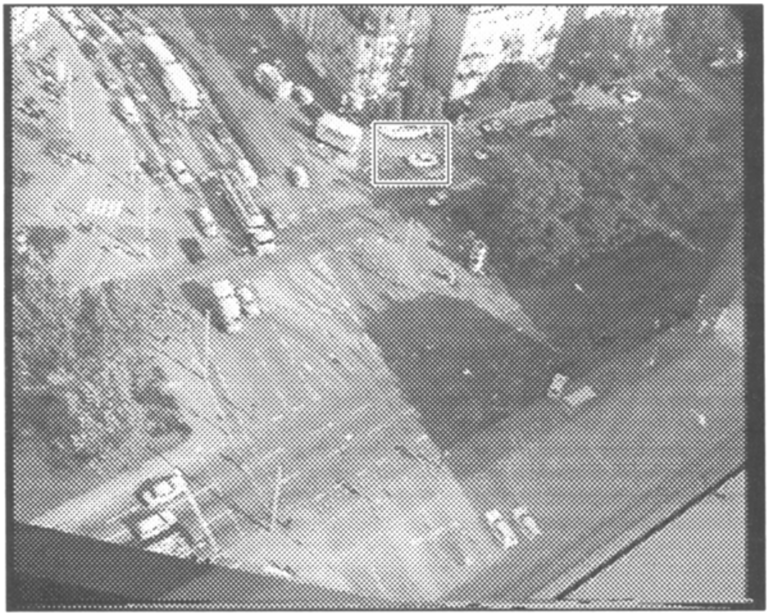

Fig. 5.: First frame of an image sequence recorded at an inner-city intersection.

The vehicle image contained in the marked rectangle in Figure 5 covers only 35×15 pixels. Even a human observer has problems to classify it as a saloon or a fast-back. In our experiments it turned out that the vehicle could not be correctly tracked by optimizing the size and shape of the vehicle model (Figure 6 (a)). After considering the occlusion by the street post in the pose estimation process, we were able to track this vehicle, see Figure 6 (b).

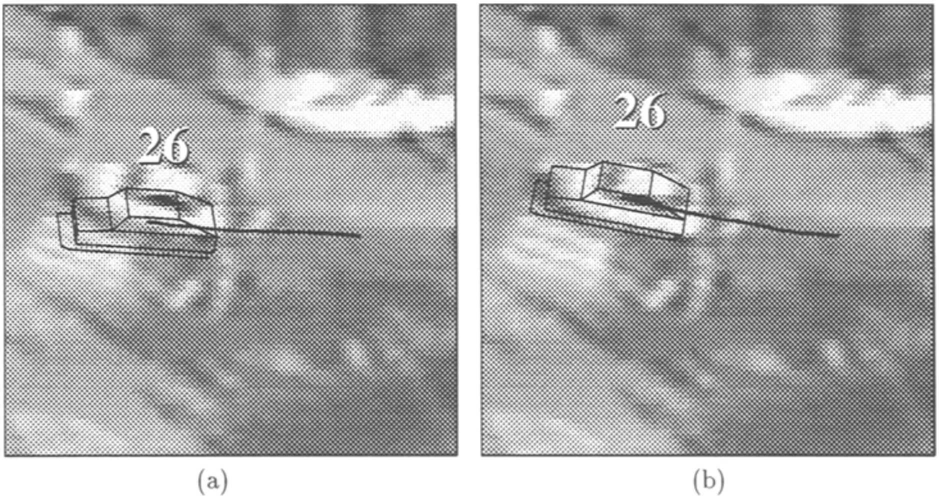

Fig. 6.: Enlarged section of the upper right quadrant of the image sequence depicted by its first frame in Figure 5 showing the tracking results of object #26 at halfframe #195. (a) The traffic light post which occludes the vehicle is *not* considered in the Kalman-Filter update step: the tracking fails. (b) As soon as the occlusion is modeled the vehicle is correctly tracked.

Figure 8 (a) shows the results of the tracking process applied to the bright transporter (object #12) which is driving side by side with the truck. Since the transporter is permanently occluded by the big truck while it is turning left, the tracking fails. In this case, the occlusion is caused by a *dynamic* scene component, namely the moving truck. Therefore, we could not determine an a–priori scene–model as in the case of the gas station scene. In this case, we first computed automatically the trajectory of the big truck where no occlusion problems arise. After this, we knew the estimated position and orientation of the truck for each time instant. We thus were able to determine the occluded parts of the image of the bright transporter by applying our raytracing algorithm to each frame and by considering the current position and orientation of the truck model at the corresponding point of time. During the tracking of the bright transporter, those parts of the vehicle image were excluded from the Kalman filter update step which were currently occluded by the projected truck model. Figure 7 shows in which parts of the vehicle image the model gradient has been suppressed due to the occlusion.

Moreover, one can see that we do not only exploit the surfaces of each car within the matching process, but in addition the information contained in its shadow. The case of an occluded shadow is treated similarly to the occlusion of the vehicle image itself. Figure 8 (b) shows the results of a successful tracking of the occluded bright transporter and the truck.

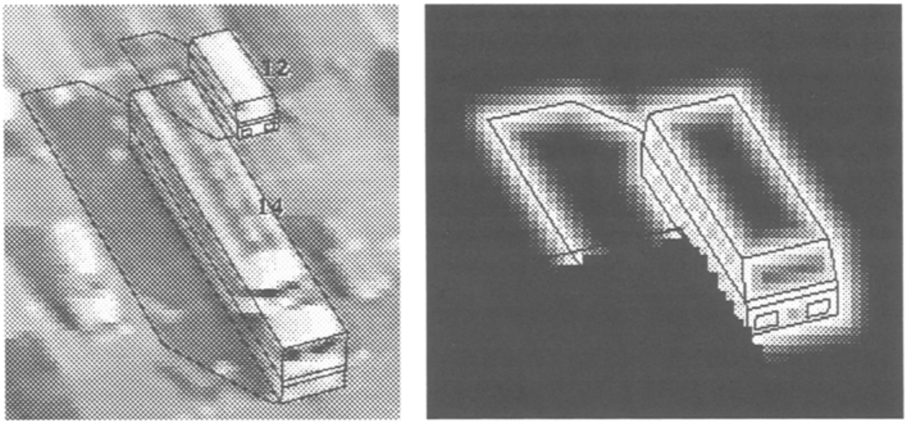

Fig. 7.: The model gradient norm of the image of the minivan (object #12) is constrained to such image points which correspond to the image of object #12 considering the occlusion by object #14. Notice that even parts of the shadow cast by object #12 are suppressed in the tracking process, due to the occlusion by object #14, although the unoccluded parts of the shadow cast by object #12 contribute to the update process.

We currently extend this approach to a *parallel* tracking of all object candidates in the scene in order to treat even bilateral occlusions where first one object occludes the other and subsequently a converse occlusion relation may occur.

5.3 Occlusion Characterization

Due to space limitations, we concentrate on the occlusion characterization between the advertising post and object #12 (see Figure 3). In order to examine the validity of the predicate $passively\_decreasing\_occlusion(advertising\_post, object\ \#12, t)$ defined in section 4, we have to consider the estimated speed of object #12 as a function of the (half-) frame number as shown in Figure 9. The predicate $\neg moving(advertising\_post, t)$ obviously holds for all t. The vehicle is occluded by the advertising post between halfframes #5 and #49. Due to the low degree of validity for $occluded(advertising\_post, object\ \#12, t)$, the degree of validity of all predicates which characterize the occlusion between the post and the vehicle will be zero in subsequent halfframes (#50 through #175). The resulting fuzzy membership function, after evaluating the considered predicate, is shown in Figure 9.

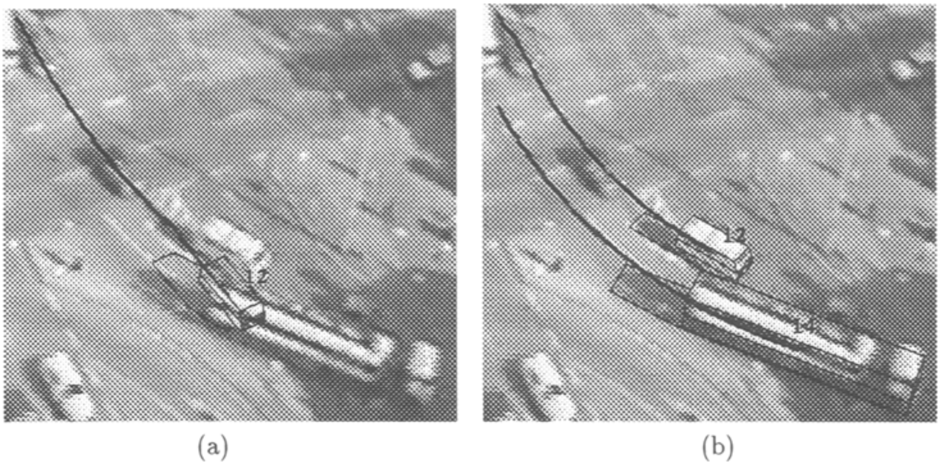

Fig. 8.: (a) Tracking of object #12 of the downtown image sequence fails if the occlusion of the minivan by the truck remains unconsidered. (b) Halfframe #293 with overlaid trajectory obtained by modelling the occlusion.

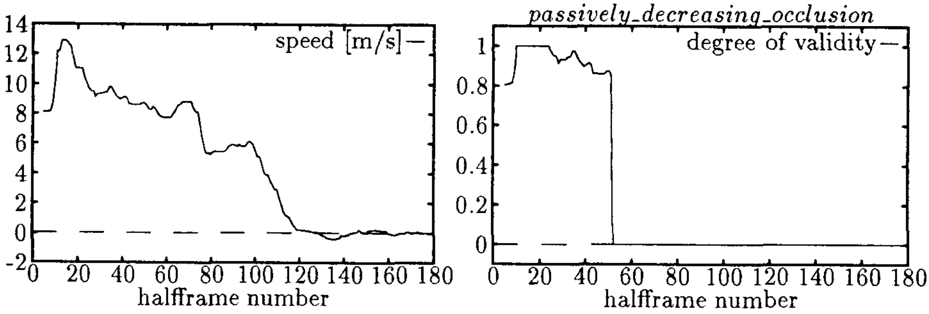

Fig. 9.: Estimated speed of object #12 in the gas station scene during the halfframes #5 through #172 (left) and the resulting characterization of the occlusion between the advertising post and the vehicle (right).

6 Conclusion

In this contribution, we explicitly modeled occlusions occuring between objects in real traffic-scenes. We showed in several experiments that the performance of the vehicle tracking process could be improved significantly by taking occlusions between static and dynamic scene components into account. We could extend this approach even to the case of moving occluding objects and to the shadow cast by a partially occluded object.

However, there still exist cases in which the occlusion modeling is not sufficient for proper vehicle tracking, e.g., when there are significant disturbances in the underlying image data or if there are influences not yet modeled, like image features in the background of the scene, for example strong road markings.

It is still an open question how we can obtain 3D models of the scenes *automa-*

tically, i.e. how we can eliminate the need for modeling the 3D scene components interactively.

Acknowledgment

We thank H. Damm for providing the gas station image sequence and for the 2D road model of the gas station.

References

[Cédras & Shah 95] C. Cédras, M. Shah, Motion-Based Recognition: A Survey, *Image and Vision Computing* **13**:2 (1995) 129–155.

[Du *et al.* 93] L. Du, G.D. Sullivan, K.D. Baker, Quantitative Analysis of the Viewpoint Consistency Constraint in Model-Based Vision, in *Proc. Fourth International Conference on Computer Vision (ICCV '93)*, Berlin, Germany, 11–14 May 1993, pp. 632–639.

[Dubuisson & Jain 95] M.-P. Dubuisson, A.K. Jain, Contour Extraction of Moving Objects in Complex Outdoor Scenes, *International Journal of Computer Vision* **14**:1 (1995) 83–105.

[Koller *et al.* 93] D. Koller, K. Daniilidis, H.-H. Nagel, Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes, *International Journal of Computer Vision* **10**:3 (1993) 257–281.

[Koller *et al.* 94] D. Koller, J. Weber, J. Malik, Robust Multiple Car Tracking with Occlusion Reasoning, in J.-O. Eklundh (Ed.), *Proc. Third European Conference on Computer Vision (ECCV '94)*, Vol. I, Stockholm, Sweden, May 2-6, 1994, Lecture Notes in Computer Science **800**, Springer-Verlag, Berlin, Heidelberg, New York/NY and others, 1994, pp. 189–196.

[Kollnig *et al.* 94] H. Kollnig, H.-H. Nagel, and M. Otte, Association of Motion Verbs with Vehicle Movements Extracted from Dense Optical Flow Fields, in J.-O. Eklundh (ed.), *Proc. Third European Conference on Computer Vision ECCV '94*, Vol. II, Stockholm, Sweden, May 2-6, 1994, Lecture Notes in Computer Science **801**, Springer-Verlag, Berlin, Heidelberg, New York/NY, and others, 1994, pp. 338–347.

[Kollnig & Nagel 95] H. Kollnig, H.-H. Nagel, 3D Pose Estimation by Fitting Image Gradients Directly to Polyhedral Models, in *Proc. Fifth International Conference on Computer Vision (ICCV '95)*, Cambridge/MA, June 20-23, 1995, pp. 569–574

[Meyer & Bouthemy 94] F. Meyer, P. Bouthemy, Region-Based Tracking Using Affine Motion Models in Long Image Sequences, *CVGIP: Image Understanding* **60**:2 (1994) 119–140.

[Sullivan 92] G.D. Sullivan, Visual Interpretation of Known Objects in Constrained Scenes, *Philosophical Transactions Royal Society London* (B) **337** (1992) 361–370.

[Sullivan *et al.* 95] G.D. Sullivan, A.D. Worrall, and J.M. Ferryman, Visual Object Recognition Using Deformable Models of Vehicles, in *Proc. Workshop on Context-Based Vision*, 19 June 1995, Cambridge/MA, pp. 75–86.

[Toal & Buxton 92] A. F. Toal, H. Buxton, Spatio-temporal Reasoning within a Traffic Surveillance System, in G. Sandini (Ed.), *Proc. Second European Conference on Computer Vision (ECCV '92)*, S. Margherita Ligure, Italy, May 18-23, 1992, Lecture Notes in Computer Science **588**, Springer-Verlag, Berlin, Heidelberg, New York/NY and others, 1992, pp. 884–892.