International Journal of Computer Vision, 10:3, 257–281 (1993)

Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes

D. KOLLER, K. DANIILIDIS AND H.-H. NAGEL†
Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlsruhe (TH), Postfach 6980, D-7500 Karlsruhe 1, Federal Republic of Germany
†Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB), Karlsruhe

Abstract
Moving vehicles are detected and tracked automatically in monocular image sequences from road traffic scenes recorded by a stationary camera. In order to exploit the a priori knowledge about shape and motion of vehicles in traffic scenes, a parameterized vehicle model is used for an intraframe matching process and a recursive estimator based on a motion model is used for motion estimation. An interpretation cycle supports the intraframe matching process with a state MAP-update step. Initial model hypotheses are generated using an image segmentation component which clusters coherently moving image features into candidate representations of images of a moving vehicle. The inclusion of an illumination model allows taking shadow edges of the vehicle into account during the matching process. Only such an elaborate combination of various techniques has enabled us to track vehicles under complex illumination conditions and over long (over 400 frames) monocular image sequences. Results on various real-world road traffic scenes are presented and open problems as well as future work are outlined.

1 Introduction

Image sequence analysis provides intermediate results for a conceptual description of events in a scene. A system that establishes such higher-level descriptions based on the tracking of moving objects in the image domain has been described in (Koller, Heinze, & Nagel 1991). Here we introduce three-dimensional models about the structure and the motion of the moving objects as well as about the illumination of the scene in order to verify the hypotheses for object candidates and to robustly extract smooth trajectories of such objects.

In order to record and analyze nontrivial events in road traffic scenes we have to cope with the following dilemma: Either we must fixate the camera on an interesting agent by applying gaze control so that the agent remains in the field of view, or we must use a stationary camera with a field of view that is large enough to capture significant actions of moving agents. The immediate shortcomings of the passive approach that is pursued in this work are the small size and the low resolution of the area covered by the projection of the moving agent. Image-domain cues like gray-value edges and corners are short and can be barely detected. Additionally, in a road traffic scene, we have to cope with a highly cluttered environment full of background features as well as with occlusions and disocclusions. This renders the task of figure-background discrimination extremely difficult. The use of models representing our a priori knowledge appears necessary in order to accomplish the difficult task of detecting and tracking under real-world conditions.

Our approach consists of the following main steps: A cluster of coherently moving image features provides a rough estimate for moving regions in the image. The assumption that such a cluster is due to a hypothetical object moving on a planar road in the scene yields a rough estimate for the position of the hypothetical object in the scene: the center of the group of the moving image features is projected back into the scene, based on a calibration of the camera. The assumption of a forward motion yields the orientation of the principal

axis of the model, which is assumed to be parallel to the motion.

Straight-line edge segments extracted from the image are matched to the 2-D model edge segments—a view sketch—obtained by projecting a 3-D polyhedral model of the vehicle into the image plane, using a hidden-line algorithm to determine their visibility. The matching of image edge segments is based on the Mahalanobis distance of line segment attributes as described in (Deriche & Faugeras 1990). The midpoint representation of line segments is suitable for using different uncertainties parallel and perpendicular to the line segments. In order to avoid incorrect matches between model segments and image edge segments which arise from shadows of the vehicles, we enrich the applied a priori knowledge by including an illumination model. This provides us with a geometrical description of the shadows of the vehicles projected onto the street plane. In this way we have been able to track even vehicles that are either mapped onto very small areas in the image or exhibit salient shadow edges.

The 3-D generic-vehicle model is parameterized by 12 length parameters. This enables the instantiation of different vehicles, for example sedan, hatchback, bus, or van from the same generic-vehicle model. The estimation of model shape parameters is possible by including them into the state estimation process.

We establish a motion model which describes the dynamic vehicle motion in the absence of knowledge about the intention of the driver. In the stationary case, in which the steering angle remains constant, the result is a simple circular motion with constant magnitude of velocity and constant angular velocity around the normal of a plane on which the motion is assumed to take place. The unknown intention of the driver in maneuvering the car is captured by the introduction of process noise. The motion parameters for this motion model are estimated using a recursive maximum a posteriori (MAP) estimator.

In the subsequent section we discuss related approaches with a comparable goal. In the third section we begin with the modeling of the scene and describe the generic 3-D vehicle model and the motion model for the vehicle. The fourth section is dedicated to the matching process. We describe how we exploit information from a motion-segmentation step to formulate an initial model hypothesis and explain the iterative matching algorithm. In the fifth section we illustrate the reason that caused us to introduce an illumination model. The sixth section contains the description of the time-recursive motion estimation. The results of our experiments are illustrated in section 7. A previous version of our tracking algorithm and some preliminary results of our experiments can be found in (Koller et al. 1992).

2 Related Investigations

In this section we discuss related investigations about tracking and recognizing object models from image sequences. The reader is referred to the excellent book by Grimson (1990b) for a complete description of research on object recognition from a single image.

Gennery (1982) has proposed the first approach for tracking 3-D objects of known structure. A constant velocity six degrees of freedom (DOF) model is used for prediction and an update step similar to the Kalman filter—without addressing the nonlinearity—is applied. Edge elements closest to the predicted-model line segments are associated as corresponding measurements. During the last ten years, Gennery's approach evolved and one can find in (Gennery 1992) the most elaborate version of this approach in estimating the motion of a known object, with particular emphasis on a time-efficient implementation of the recursive estimation and on the propagation of uncertainty. The used force- and torque-free motion model is the same as in (Gennery 1982) and can be applied to robot activities in space.

Thompson and Mundy (1987) emphasize the object-recognition aspect of tracking by applying a pose-clustering technique. Candidate matches between image and model vertex pairs define points in the space of all transformations. Dense clusters of such points indicate a correct match. Object motion can be represented by a trajectory in the transformation space. Temporal coherence then means that this trajectory should be smooth. Predicted clusters from the last time instant establish hypotheses for the new time instants, which are verified as matches if they lie close to the newly obtained clusters. The images we have been working on did not contain the necessary vertex pairs in order to test this novel algorithm. Furthermore, we have not been able to show that the approach of Thompson and Mundy (1987) can be extended to the handling of parameterized objects.

Verghese, Gale, and Dyer (1990) have implemented two approaches for tracking 3-D-known objects in real time. Their first method is similar to the approach of Thompson and Mundy (1987) (see the preceding para-

graph). Their second method is based on the optical flow of line segments. Using line-segment correspondences, of which initial (correct) correspondences are provided interactively at the beginning, a prediction of the model is validated and spurious matches are rejected.

Lowe (1991) built the system that has been the main inspiration for our matching strategy. He does not enforce temporal coherence however, since he does not imply a motion model. Pose updating is carried out by minimization of a sum of weighted least squares including a priori constraints for stabilization. Line segments are used for matching but distances of selected edge points from infinitely extending model lines are used in the minimization. Lowe (1990) uses a probabilistic criterion to guide the search for correct correspondences and a match-iteration cycle similar to ours.

A gradient-ascent algorithm is used by Worrall et al. (1991) in order to estimate the pose of a known object in a car sequence. Initial values for this iteration are provided interactively at the beginning. Since no motion model is used, the previous estimate is used at every time instant to initialize the iteration. Marslin, Sullivan, and Baker (1991) have enhanced the approach by incorporating a motion model of constant translational acceleration and angular velocity. Their filter optimality, however, is affected by use of the speed estimates as measurements instead of the image locations of features.

Schick and Dickmanns (1991) use a generic parameterized model for the object types. They solve the more general problem of estimating both the motion and the shape parameters. The motion model of a car moving on a clothoid trajectory is applied including translational as well an angular acceleration. The estimation machinery of the simple extended Kalman filter (EKF) is used. So far, however, their approach has only been tested on synthetic line images.

The following approaches do not consider the correspondence search problem but concentrate only on the motion estimation. A constant-velocity model with six DOF is assumed by Wu et al. (1988), Harris and Stennet (1990), and Evans (1990), whereas Young and Chellappa (1990) use a precessional-motion model. A problem similar to that of Young and Chellappa (1990) with stereo images is solved by Zhang and Faugeras (1992), where closed-form solutions for the prediction step are established using a constant angular velocity and constant translational acceleration-motion model. By means of the motion estimates of the tracked 3-D line segments, Zhang and Faugeras (1992) obtain groupings of the 3-D line segments into single objects.

A quite different paradigm is followed by Murray, Castelow, and Buxton (1989). They first try to solve the structure from the motion problem from two monocular views. In order to accomplish this, they establish temporal correspondence of image-edge elements and use these correspondences to solve for the infinitesimal motion between the two time instants and the depths of the image points. On the basis of this reconstruction, Murray et al. (1989) carry out a 3-D–3-D correspondence search. Their approach has been tested with camera motion in a laboratory set-up.

We have restricted our brief survey to approaches reasoning in the 3-D-scene domain. Further approaches exist for tracking moving objects in the image domain by using hypotheses about the change of the projections of the underlying 3-D objects.

3 Models for the Vehicles and their Motion

3.1 The Parameterized-Vehicle Model

A first step in modeling the scene is to define spatial models for the mobile-scene components, these are the vehicles in our domain region. We use a parameterized 3-D generic model to represent various types of vehicles moving in traffic scenes that have been recorded by us. The parameters are given in figure 1. An *internal* vehicle-model instantiation is established by a set of

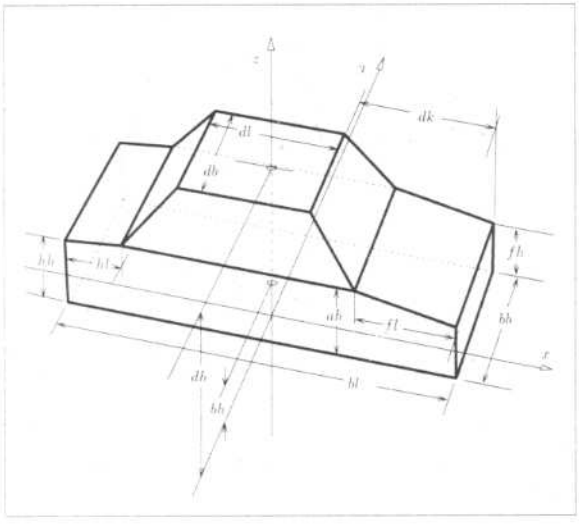

Fig. 1. We use a parameterized 3-D generic model to represent the various types of vehicles moving in traffic scenes. The model comprises 12 length parameters.

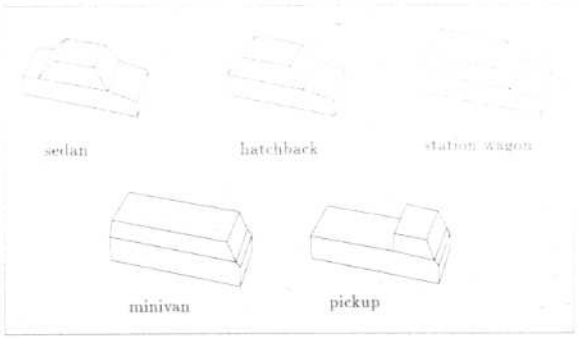

Fig. 2. Example of five different vehicle models derived from the same generic model.

values for the 12 length parameters. In the current implementation we use a fixed, interactively selected, set of values for the internal vehicle-model parameters for each vehicle in the scene. We will denote the position and orientation of the vehicle-model coordinate frame with respect to the scene coordinates as the *external* vehicle-model instantiation.

Different types of vehicles are generated from this representation by varying the 12 length parameters of our model. Figure 2 shows an example of five different specific vehicle models derived from the same generic model.

3.2 The Motion Model

We use a motion model that describes the dynamic behavior of a road vehicle without knowledge about the intention of the driver. Since we further assume that the motion is constrained onto the street plane, we get in the stationary case a simple circular motion with a constant magnitude of the translational velocity v and a constant angular velocity ω. The remaining three degrees of freedom of the external vehicle model instantiation are described by the position $\mathbf{p}$ of the vehicle center on the street plane and the orientation angle ϕ around the normal of the plane (the z-axis) through the vehicle center, that is, the orientation of the principal axis of the vehicle model with respect to the scene coordinate system. In this way we have only one angular velocity $\omega = \dot{\phi}$.

The position of the object center $\mathbf{p}(t) = (p_x(t), p_y(t), 0)^T$ in the street plane is described by (see figure 3):

$$\mathbf{p}(t) = \mathbf{C} + \rho \begin{pmatrix} \sin \phi(t) \\ -\cos \phi(t) \\ 0 \end{pmatrix} \tag{1}$$

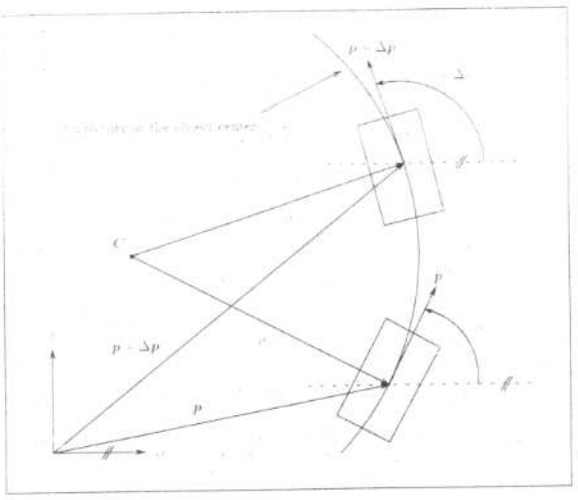

Fig. 3. Stationary circular motion as a motion model. During the time interval $\tau = t_{k+1} - t_k$ the center of the object has shifted about $\Delta\mathbf{p}$ and rotated about $\Delta\phi$ around the normal (the z-axis) through the object center of the plane on which the motion takes place.

Differentiation of (1) with respect to the time t and elimination of the radius ρ by using $v = |\mathbf{v}| = |\dot{\mathbf{p}}| = \rho\dot{\phi} = \rho\omega$ results in the motion model described by the following differential equation:

$$\begin{aligned} \dot{p}_x &= v \cos \phi \\ \dot{p}_y &= v \sin \phi \\ \dot{v} &= 0 \\ \dot{\phi} &= \omega \\ \dot{\omega} &= 0 \end{aligned} \tag{2}$$

The deviation of this idealized motion from the real motion is captured by process noise associated with v and ω. In order to recognize pure translational motion in noisy data, we evaluate the angle difference $\omega\tau$ ($\tau = t_{k+1} - t_k$ is the time interval between the frame times t_{k+1} and t_k). In case $\omega\tau$ is smaller than a threshold, we assume pure translational motion with the estimated (constant) angle ϕ and $\dot{\phi} = \omega = 0$.

In equation (2) it is assumed that the principal axis of the model through the model center is tangential to the circle that is described by the moving object center. In the general case this is not true, but the deviation—the so-called slip angle β—could easily be compensated by shifting the center of rotation along the principal axis of the model to a position at which the principal axis of the model is tangential to the circle along which the vehicle drives. In case of slow stationary circular motion this rotation center lies in the intersection of the

principal axis with the rear wheel axis of the vehicle (Mitschke 1990). We assume a slow stationary circular motion and use this shifted rotation center in order to compensate for the slip angle β.

4 The Matching Process

4.1 Matching Primitives

The matching between model data and image data is performed on edge segments. The model edge segments are the edges of the 3-D polyhedral model which are projected from the 3-D scene back into the 2-D image. We will call them *model segments* (see the lower left image of figure 4). Invisible model edge segments are removed by a hidden-line algorithm. The image edge segments are extracted and approximated using the method of Korn (1988). We will call them *data segments* (see, e.g., the lower right image of figure 4).

4.2 Finding Correspondences Between Matching Primitives

Correspondences between model and data segments are established using the Mahalanobis distance between attributes of the line segments as described by Deriche and Faugeras (1990). We use the representation $\mathbf{X} = (c_x, c_y, \theta, l)$ of a line segment, defined as

$$c_x = \frac{x_1 + x_2}{2}$$

$$c_y = \frac{y_1 + y_2}{2}$$

$$\theta = \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right)$$

$$l = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (3)$$

where $(x_1, y_1)^T$ and $(x_2, y_2)^T$ are the endpoints of a line segment. The advantage of this representation is to

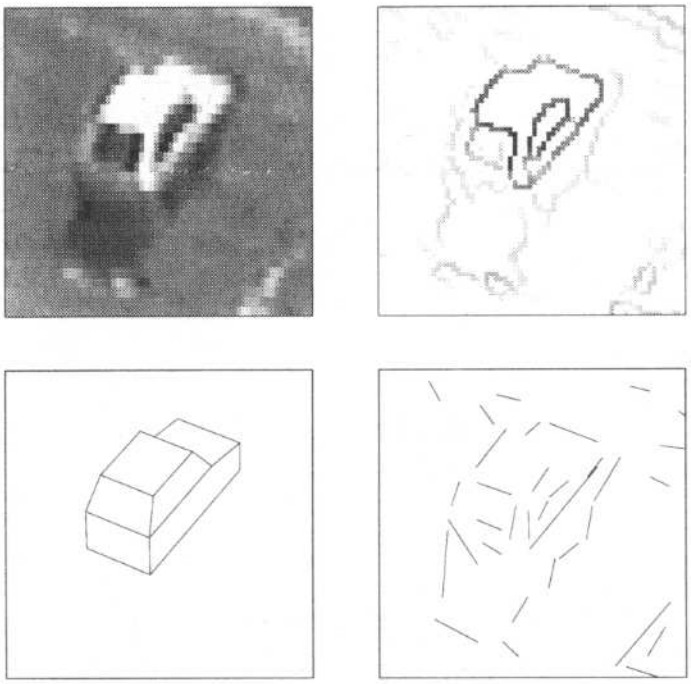

Fig. 4. The upper left image shows a small enlarged image section with a moving vehicle. The lower left image shows the projected model, which yields the so-called *model segments*. The upper right figure shows the grey-coded maxima gradient magnitude in the direction of the gradient of the image function; and the lower right image shows the straight-line segments extracted from these data, which we call the *data segments*. This example also illustrates the complexity of the task of detecting and tracking small moving objects. The shown object spans no more than 20×40 pixels in the image.

distinguish between the, in general quite different, uncertainties of the endpoints parallel and perpendicular to the line-segment direction.

Denoting by $\sigma_{||}$ the uncertainty in the position of the endpoints along an edge chain and by $\sigma_{\perp}$ the positional uncertainty perpendicular to the linear edge chain approximation, a covariance matrix Λ is computed, depending on $\sigma_{||}$, $\sigma_{\perp}$, θ, and l. Given the attribute vector $\mathbf{X}_{m_i}$ of a model segment and the attribute vector $\mathbf{X}_{d_j}$ of a data segment, the Mahalanobis distance between $\mathbf{X}_{m_i}$ and $\mathbf{X}_{d_j}$ is defined as

$$d_{ij}^2 = (\mathbf{X}_{m_i} - \mathbf{X}_{d_j})^T (\Lambda_{m_i} + \Lambda_{d_j})^{-1} (\mathbf{X}_{m_i} - \mathbf{X}_{d_j}) \quad (4)$$

Given the uncertainties in the position of the endpoints $\sigma_{||}$ and $\sigma_{\perp}$ of the data segment j the covariance matrix Λ_{d_j} is defined by equation 12 in (Deriche & Faugeras 1990):

$$\Lambda_{d_j} = \begin{pmatrix} \frac{\sigma_{||}^2 \cdot \cos^2\theta + \sigma_{\perp}^2 \cdot \sin^2\theta}{2} & \frac{(\sigma_{||}^2 - \sigma_{\perp}^2) \cdot \sin\theta\cos\theta}{2} & 0 & 0 \\ \frac{(\sigma_{||}^2 - \sigma_{\perp}^2) \cdot \sin\theta\cos\theta}{2} & \frac{\sigma_{\perp}^2 \cdot \cos^2\theta + \sigma_{||}^2 \cdot \sin^2\theta}{2} & 0 & 0 \\ 0 & 0 & \frac{2\sigma_{\perp}^2}{l^2} & 0 \\ 0 & 0 & 0 & 2\sigma_{||}^2 \end{pmatrix} \quad (5)$$

The covariance matrix Λ_{m_i} of a model segment i is expressed by the actual covariance matrix P of the state estimation $\mathbf{x}$ by the following equation (cf. equations (11) and (13) in section 6):

$$\Lambda_{m_i} = \left(\frac{\partial \mathbf{X}_{m_i}}{\partial \mathbf{x}}\right) P \left(\frac{\partial \mathbf{X}_{m_i}}{\partial \mathbf{x}}\right)^T \quad (6)$$

$\partial \mathbf{X}_{m_i}/\partial \mathbf{x}$ is the $4 \times n_{\text{state}}$ submatrix of the Jacobian matrix $H = \partial \mathbf{X}_m / \partial \mathbf{x}$ of the measure function $\mathbf{h}$ according to the ith model segment (see next section), n_{state} is the state dimension, that is, the number of state parameters.

Let $\mathfrak{M}$ denote the set of n model segments

$$\mathfrak{M} = \{M_i\}_{i=1\ldots n} \quad (7)$$

and $\mathfrak{D}$ denote the set of p data segments

$$\mathfrak{D} = \{D_i\}_{i=1\ldots p} \quad (8)$$

then we define a model interpretation $\mathfrak{I}_j$ as the set of correspondences of model segments M_i and data segments D_{i_j}:

$$\mathfrak{I}_j = \{(M_i, D_{i_j})\}_{i=1\ldots n} \quad (9)$$

A correspondence in the interpretation $\mathfrak{I}_j$ from a model segment M_i is established to the data segment D_{i_j} in such a way that D_{i_j} is the data segment that minimizes the Mahalanobis distance to the model segment, provided that the Mahalanobis distance d is less than a threshold d_τ:

$$D_{i_j} = D_{k_{\min}} \in \mathfrak{D} \cup \text{NIL} \quad \text{with}$$
$$k_{\min} = \arg\min \{d_{ik} \mid d_{ik} < d_\tau\}_{k=1\ldots p} \quad (10)$$

The NIL element is included in order to cope with nonmatches.

Due to the structure of vehicles this is not always the best match. The known vehicles and their models consist of two essential sets of parallel line segments. One set along the orientation of the modeled vehicle and one set perpendicular to this direction. This yields not in all cases a unique best-model interpretation. But evidence from our experiments so far supports our hypothesis that in most cases appropriate initial assumptions about the position and orientation can be derived. These are good enough to obviate the necessity for a combinatorial search as for example in (Grimson 1990a).

The search window for corresponding line segments in the image is a rectangle around the projected model segments. The dimensions of this rectangle are intentionally set by us to a higher value than the values obtained from the estimated uncertainties in order to overcome the optimism of the IEKF as explained in section 6.

4.3 Computing Initial Values

We will now describe how we derive initial values for the position and orientation for an initial model instantiation which enables us to formulate an appropriate object hypothesis.

We obtain these initial values (the position $\mathbf{p}_0$ and orientation ϕ_0 of the vehicle on the street plane) by exploiting the information from a motion-segmentation step. In this motion-segmentation step the displacement of extracted image features is used to establish displacement vectors between consecutive frames. The assumption of a dominant translational motion of the vehicles enables the use of a clustering approach in which neighboring moving image features are grouped under the

assumption, that they represent projected features of a single moving vehicle, using certain consistency criteria and relations as described in (Sung 1988; Koller, Heinze, & Nagel 1991). The enclosing rectangle of such a clustered group of moving image features is then assumed to represent the image of a moving vehicle as shown in the left image of figure 5. In order to obtain corresponding descriptions of moving vehicles in scene coordinates, we exploit the fact that the depth variation of the object is very small compared to the distance to the camera. The centers of the image features can then be projected back to the street plane, assuming the same average height of the corresponding scene features (the average height of features on a car is taken to be 0.8 meter) as seen in the upper right of figure 5. The camera parameters are provided by an external camera calibration program. The formulation of an object hypothesis is now performed using the center of the back-projected image features (i.e., the scene features) as position $\mathbf{p}_0$ and the average displacement direction of the scene features as the orientation angle ϕ_0 of the object as shown in the lower left of figure 5. This assumes a positive velocity of the object, that is, the object is moving forward.

4.4 The Matching Algorithm

Similar to the approach of Lowe (1987), we use an iterative approach to find the set with the best correspondence between 3-D model edge segments and 2-D image edge segments. However, we do not update the correspondence after each step of the applied minimization. Using the same correspondences we perform a complete state update based on a MAP estimation by means of the Levenberg-Marquardt minimization method. At the end of each *minimization iteration*—the inner loop—a new correspondence set (i.e., a new model interpretation) is determined according to the estimated state of position and orientation and then the state is updated again by the minimization iteration. This outer loop, which we call *interpretation loop*, is necessary to take into account the visibility of edge segments depending on the viewing direction and the

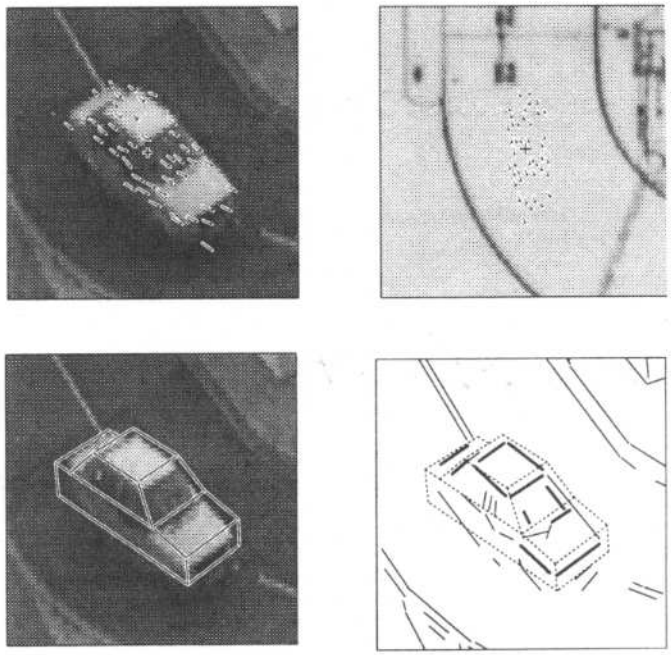

Fig. 5. The computation of initial values for initializing a model instantiation in the iterative matching process is based on image regions provided by a motion-segmentation step of optical-flow vectors and are assumed to represent the image of a moving object (upper left image). Back projection of the image features yields corresponding descriptions in scene coordinates (overlayed to a digitized image of an official map in the upper right image). The center of these scene features determines the position and the averaged displacement of the scene features determines the orientation of the initial model instantiation (lower left image). The lower right image exhibits the established model interpretation as the set of correspondences between the projected 3-D model segments according to the initial model instantiation (dotted lines) and the extracted image edge segments (fat lines).

$i \leftarrow 0$
$\mathcal{I}_i \leftarrow$ get_correspondences($\boldsymbol{x}^-$)
DO
  $\boldsymbol{x}_i^+ \leftarrow$ update_state($\mathcal{I}_i$)
  $r_i \leftarrow$ residual($\mathcal{I}_i$)
  $\mathcal{I}_{i+1} \leftarrow$ get_correspondences($\boldsymbol{x}_i^+$)
  $i \leftarrow i + 1$
WHILE$((\mathcal{I}_i \neq \mathcal{I}_j;\ j = 0, 1 \ldots, i) \wedge i \leq \mathrm{IMAX})$
$i_{min} \leftarrow \arg\min(r_j);\ j = 0, 1 \ldots, \mathrm{IMAX}$
$\boldsymbol{x}^+ \leftarrow \boldsymbol{x}_{i_{min}}^+$

Fig. 6. First, initial values for the position and orientation of a vehicle model instantiation $\mathbf{x}^-$ are computed either by back projection of coherently moving image features or by prediction using the transition function according to our motion model. Second, an initial model interpretation is established by building correspondences of model and data segments. The model interpretation that leads to the smallest residual is then determined by an iteration with respect to combinations of model and data segments. At the end of each iteration step, a new model interpretation is established according to the updated estimated state of position and orientation of the object model in the scene. The iteration stops if the new model interpretation has been treated already or a threshold number of iteration steps has been reached. The state update step is based on a MAP estimation using a Levenberg-Marquardt minimization method.

current estimated state of position and orientation, respectively. The interpretation loop is terminated if a certain number of loop steps has been reached or the established new model interpretation has been treated already.

Our of the set of model interpretations investigated in the interpretation loop, the model interpretation that results in the smallest residual is then used as a state update. The algorithm is sketched in figure 6. We use the average residual per matched edge segment, multiplied by a weight which takes the lengths of edge segments into account, as a criterion for the selection of the smallest residual.

5 The Illumination Model

In this section we illustrate the reason that caused us to introduce an illumination model.

During initial experiments with video sequences from real-world traffic scenes in which vehicles exhibit salient shadow edges, for instance in the upper frames of figure 7, sometimes incorrect model interpretations were established. Such incorrect model interpretations are characterized by correspondences between model line segments and image edge segments which arise from the shadows of the vehicles on the street plane. The left column of figure 7 shows an incorrect model interpretation due to incorrect matches between shadow edges and model line segments. This results in incorrect estimates for the position and orientation of the vehicle model. This problem can be overcome by including the shadow into the modeling of the scene based on an illumination model. The right column of figure 7 shows the result of the best match, using the same initial values but with the inclusion of shadow edges into the modeling of the scene. As can be seen, much better estimates for the position and orientation of the vehicle model are obtained.

We use a simple illumination model which assumes parallel incoming light, and compute the visible contour of the 3-D vehicle model projected onto the street plane (see figure 8). This contour is then decomposed into straight line segments. The two parameters for the illumination direction are set interactively off-line and are assumed to be constant during the entire image sequence as well as to be the same for all vehicles in the scene.

6 Recursive Motion Estimation

In this section we elaborate the recursive estimation of the vehicle motion parameters. As we have already described in section 3.2, the assumed model is a uniform motion of a known vehicle model along a circular arc.

The state vector $\mathbf{x}_k$ at time point t_k is a five-dimensional vector consisting of the position $(p_{x,k}, p_{y,k})$ and orientation ϕ_k of the model as well as the magnitudes v_k and ω_k of the translational angular velocities, respectively:

$$\mathbf{x}_k = (p_{x,k} \quad p_{y,k} \quad \phi_k \quad v_k \quad \omega_k)^T \tag{11}$$

By integrating the differential equations (2) we obtain the following discrete plant model describing the state transition from time point t_k to time point t_{k+1}:

$$\begin{aligned}
p_{x,k+1} &= p_{x,k} + v_k \tau \cdot \frac{\sin(\phi_k + \omega_k \tau) - \sin \phi_k}{\omega_k \tau} \\
p_{y,k+1} &= p_{y,k} - v_k \tau \cdot \frac{\cos(\phi_k + \omega_k \tau) - \cos \phi_k}{\omega_k \tau} \\
\phi_{k+1} &= \phi_k + \omega_k \tau \\
v_{k+1} &= v_k \\
\omega_{k+1} &= \omega_k
\end{aligned} \tag{12}$$

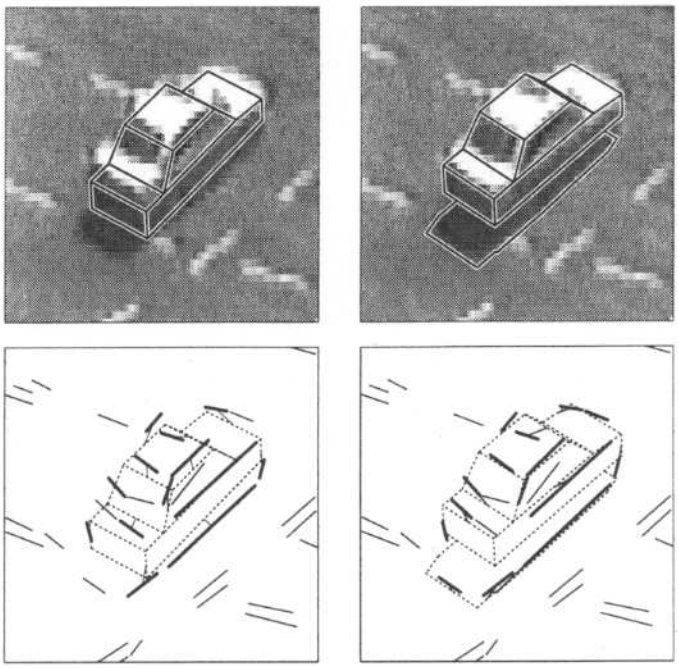

Fig. 7. The two left-hand subfigures illustrate problems created by incorrectly estimated positions and orientations of the vehicle model due to a model interpretation corrupted by matching data segments which arise from the shadow of the vehicle on the street plane. By including the shadow into the scene model, the shadow-image edge segments can be treated correctly in a model interpretation. This results in much better estimates for the position and orientation of the vehicle model as seen in the two images on the right-hand side.

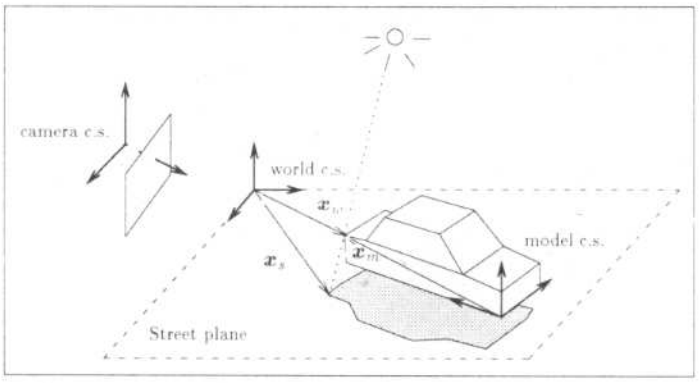

Fig. 8. Description of coordinate systems (c.s.) and illustration of the illumination model applied to our investigations. We assume parallel incoming light and compute the visible contour of the 3-D vehicle model projected onto the street plane.

We introduce the usual dynamical systems notation (see, e.g., (Gelb 1974)). The symbols $(\hat{\mathbf{x}}_k^-, P_k^-)$ and $(\hat{\mathbf{x}}_k^+, P_k^+)$ are used, respectively, for the estimated states and their covariances before and after updating based on the measurements at time t_k.

By denoting the transition function of (12) by $\mathbf{f}(\cdot)$ and assuming white Gaussian process noise $\mathbf{w}_k \sim \mathfrak{N}(\mathbf{0}, Q_k)$, the prediction equations read as follows:

$$\begin{aligned} \hat{\mathbf{x}}_{k+1}^- &= \mathbf{f}(\hat{\mathbf{x}}_k^+) \\ P_{k+1}^- &= F_k P_k^+ F_k^T + Q_k \end{aligned} \tag{13}$$

where F_k is the Jacobian $\partial\mathbf{f}/\partial\mathbf{x}$ at $\mathbf{x} = \hat{\mathbf{x}}_k^+$.

The four dimensional parameter vectors $\{\mathbf{X}_i\}_{i=1\ldots m}$ from m matched line segments in the image plane (i.e., the set of data segments $\{D_k\}_{i=1\ldots m}$ for which there is a corresponding model segment in a certain model

interpretation) are combined into a $4m$-dimensional measurement vector $\mathbf{z}_k$ assumed to be equal to the measurement function $\mathbf{h}_k(\mathbf{x}_k)$ plus white Gaussian measurement noise $\mathbf{v}_k \sim \mathfrak{N}(\mathbf{0}, R_k)$.

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k \tag{14}$$

$$\mathbf{h}_k(\mathbf{x}_k) = \begin{pmatrix} \mathbf{X}_1(\mathbf{x}_k) \\ \vdots \\ \mathbf{X}_m(\mathbf{x}_k) \end{pmatrix}$$

$$\mathbf{X}_i(\mathbf{x}_k) = \begin{pmatrix} c_{x_i}(\mathbf{x}_k) \\ c_{y_i}(\mathbf{x}_k) \\ \theta_i(\mathbf{x}_k) \\ l_i(\mathbf{x}_k) \end{pmatrix} \quad i = 1 \ldots m \tag{15}$$

The measurement noise covariance matrix R_k is block-diagonal. Its blocks are 4×4 covariance matrixes as in equation (5). As already formulated in section 4, the line-segment parameters are functions of the endpoints of a line segment. We will briefly explain how these endpoints are related to the state (11). A point (x_i, y_i) in the image plane at time instant t_k is the projection of a point $\mathbf{x}_{w_{i,k}}$ described in the world coordinate system (see figure 8).

The parameters of this transformation have been obtained off-line, based on the calibration procedure of Tsai (1987), using dimensional data extracted from a construction map of the depicted roads. In this way we constrain the motion problem even more because we not only know that the vehicle is moving on the road plane, but we know the normal of this plane as well. The point $\mathbf{x}_{w_{i,k}}$ is obtained by the following rigid transformation from the model coordinate system:

$$\mathbf{x}_{w_{i,k}} = \begin{pmatrix} \cos\phi_k & -\sin\phi_k & 0 \\ \sin\phi_k & \cos\phi_k & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{x}_{m,i} + \begin{pmatrix} p_{x,k} \\ p_{y,k} \\ 0 \end{pmatrix} \tag{16}$$

where $(p_{x,k}, p_{y,k}, \phi_k)$ are the state parameters and $\mathbf{x}_{m,i}$ are the known positions of the vehicle vertexes in the model coordinate system.

As already mentioned, we have included the projection of the shadow contour among the measurements in order to obtain more predicted edge segments for matching and to avoid false matches to data edge segments arising from shadows that lie in the neighborhood of predicted model edges. The measurement function of projected shadow edge segments differs from the measurement function of the projections of model vertexes in one step. Instead of only one point in the world coordinate system we get two. One point $\mathbf{x}_s$ as vertex of the shadow on the street and a second point $\mathbf{x}_w = (x_w, y_w, z_w)$ as vertex on the object which is projected onto the shadow point $\mathbf{x}_s$. We assume a parallel projection in shadow generation. Let the light-source direction be $(\cos\alpha \sin\beta, \sin\alpha \sin\beta, \cos\beta)^T$ where α and β—set interactively off-line—are the azimuth and polar angle, respectively, described in the world coordinate system. The following expression for the shadow point in the xy-plane (the road plane) of the world coordinate system can be easily derived:

$$\mathbf{x}_s = \begin{pmatrix} x_w - z_w \cos\alpha \tan\beta \\ y_w - z_w \sin\alpha \tan\beta \\ 0 \end{pmatrix} \tag{17}$$

The point $\mathbf{x}_w$ can then be expressed as a function of the state using (16). A problem arises with endpoints of line segments in the image that are not projections of model vertexes but intersections of occluding line segments. Due to the small length of the possibly occluded edges (for example, the side edges of the hood and of the trunk of the vehicle) we cover this case by the already included uncertainty $\sigma_{||}$ of the endpoints in the edge direction. A formal solution would require a closed-form expression for the endpoint position in the image as a function of the coordinates of the model vertexes belonging to the occluded and occluding edge segments. Such a closed-form solution has not yet been implemented in our system.

The measurement function $\mathbf{h}_k$ is nonlinear in the state $\mathbf{x}_k$. Therefore, we have tested three possibilities for the updating step of our recursive estimation. In all three approaches we assume that the state after the measurement $\mathbf{z}_k$ is normally distributed around the estimate $\hat{\mathbf{x}}_{k-1}^+$ with covariance P_{k-1}^+, which is only an approximation to the actual a posteriori probability density function (pdf) after an update step based on a nonlinear measurement. An additional approximation is the assumption that the pdf after the nonlinear prediction step remains Gaussian. Thus we state the problem as the search for the maximum of the following a posteriori pdf after measurement $\mathbf{z}_k$:

$$p(\mathbf{x}_k \mid \mathbf{z}_k) =$$

$$\frac{1}{c} \exp\left\{-\frac{1}{2}\,[\mathbf{z}_k - \mathbf{h}_k(\mathbf{x}_k)]^T\, R_k^{-1}\, [\mathbf{z}_k - \mathbf{h}_k(\mathbf{x}_k)]\right\}$$

$$\times \exp\left\{-\frac{1}{2}\,(\mathbf{x}_k - \hat{\mathbf{x}}_k^-)^T\, P_k^{-1}\, (\mathbf{x}_k - \hat{\mathbf{x}}_k^-)\right\} \quad (18)$$

where c is a normalizing constant.

This is a MAP estimation and can be stated as the minimization of the objective function

$$[\mathbf{z}_k - \mathbf{h}_k(\mathbf{x}_k)]^T\, R_k^{-1}\, [\mathbf{z}_k - \mathbf{h}_k(\mathbf{x}_k)]$$

$$+ (\mathbf{x}_k - \hat{\mathbf{x}}_k^-)^T\, P_k^{-1}\, (\mathbf{x}_k - \hat{\mathbf{x}}_k^-) \rightarrow \min_{\mathbf{x}_k} \quad (19)$$

resulting in the updated estimate $\hat{\mathbf{x}}_k^+$. In this context the well-known *iterated extended Kalman filter* (IEKF) (Jazwinski 1970; Bar-Shalom & Fortmann 1988) is actually the Gauss-Newton iterative method (Scales 1985) applied to the above objective function whereas the *extended Kalman filter* (EKF) corresponds only to the first iteration step in the framework of this method. We have found such a clarification (Jazwinski 1970) of the meaning of EKF and IEKF to be important for understanding the performance of each method.

In order to control a potential divergence observable during application of the Gauss-Newton minimization—to which the IEKF is equivalent—we have considered as a third possibility the Levenberg-Marquardt iteration method which we call *modified IEKF*. The Levenberg-Marquardt technique applied to (19) yields the following iteration step:

$$\hat{\mathbf{x}}_k^{i+1} - \hat{\mathbf{x}}_k^i = \left\{(H_k^i)^T\, R_k^{-1}\, H_k^i + P_k^{-1} + \mu I\right\}^{-1}$$

$$\left\{-(H_k^i)^T\, R_k^{-1}\, [\mathbf{z}_k - \mathbf{h}_k(\hat{\mathbf{x}}_k^i)]\right.$$

$$\left. + P_k^{-1}\, (\hat{\mathbf{x}}_k^i - \hat{\mathbf{x}}_k^-)\right\} \quad (20)$$

where i is the index for the *minimization iteration* at time point t_k. The parameter μ is increased until a step in a descent direction is taken. For large μ, such a step is very small and leads to slow but guaranteed convergence. If steps are taken in a descent path then μ is decreased. At the limit $\mu \rightarrow 0$ the steps are in the direction determined by a pure Gauss-Newton approach and convergence is accelerated. If the initial values are in the close vicinity of the minimum, the IEKF and modified IEKF yield the same result.

The state convariance is updated as follows

$$\left[(H_k^{i+1})^T\, R_k^{-1}\, H_k^{i+1} + P_k^{-1}\right]^{-1} \quad (21)$$

where H_k^{i+1} is the Jacobian of the measurement function evaluated at the updated state $\hat{\mathbf{x}}_k^{i+1}$.

Due to the mentioned approximations, all three methods are suboptimal and the computed covariances are optimistic (Jazwinski 1970). This fact practically affects the matching process by narrowing the search region and making the matcher believe that the current estimate is much more reliable than it actually is. Practical compensation methods include the addition of artificial process noise or a multiplication with an amplification matrix. We did not apply such methods in our experiments, in order to avoid a severe violation of the smoothness of the trajectories. We have just added process noise to the velocity magnitude v and ω (about 10% of the actual value) in order to compensate the inadequacy of the motion model with respect to the real motion of a vehicle.

We have tested all three methods (Thórhallson 1991) and it turned out that the IEKF and the modified IEKF are superior to the EKF regarding convergence as well as retainment of a high number of matches. As Maybank (1990) suggested, these suboptimal filters are the closer to the optimal filter, in a minimum mean square error sense, the nearer the initial value lies to the optimal estimate. This criterion is actually satisfied by the initial position and orientation values in our approach obtained by back projecting image features clustered into objects onto a plane parallel to the street. In addition to the starting values for position and orientation, we computed initial values for the velocity magnitudes v and ω during a bootstrap process. During the first n_{boot} ($= 2$, usually) time frames, position and orientation are statically computed. Subsequently, initial values for the velocities are determined by the discrete time derivatives of these positions and orientations. Alternatively, a reasonable initialization could be obtained by applying a least-squares batch algorithm on the first n_{boot} frames and using the associated *Cramer-Rao lower bounds* as initial values for the covariances, as proposed by Broida, Chandrashekhar, and Chellappa (1990).

Concluding the estimation section we should mention that the above process requires only a slight modification for the inclusion of the shape parameters of the model as additional unknowns in the state vector. Since shape parameters remain constant, the prediction step is the same, and the measurement function must be modified by substituting the model points $\mathbf{x}_{m,i}$ as functions of the shape parameters instead of considering them to have constant coordinates in the model coordinate system.

7 Experiments and Results

Leaving Car

As a first experiment we used an image sequence of about 80 frames in which one car is moving from the left to the right leaving a parking area (see the three upper images of figure 9). The image of the moving car covers about 60×100 pixels of a (512×576) frame. In this example it was not necessary, and due to the

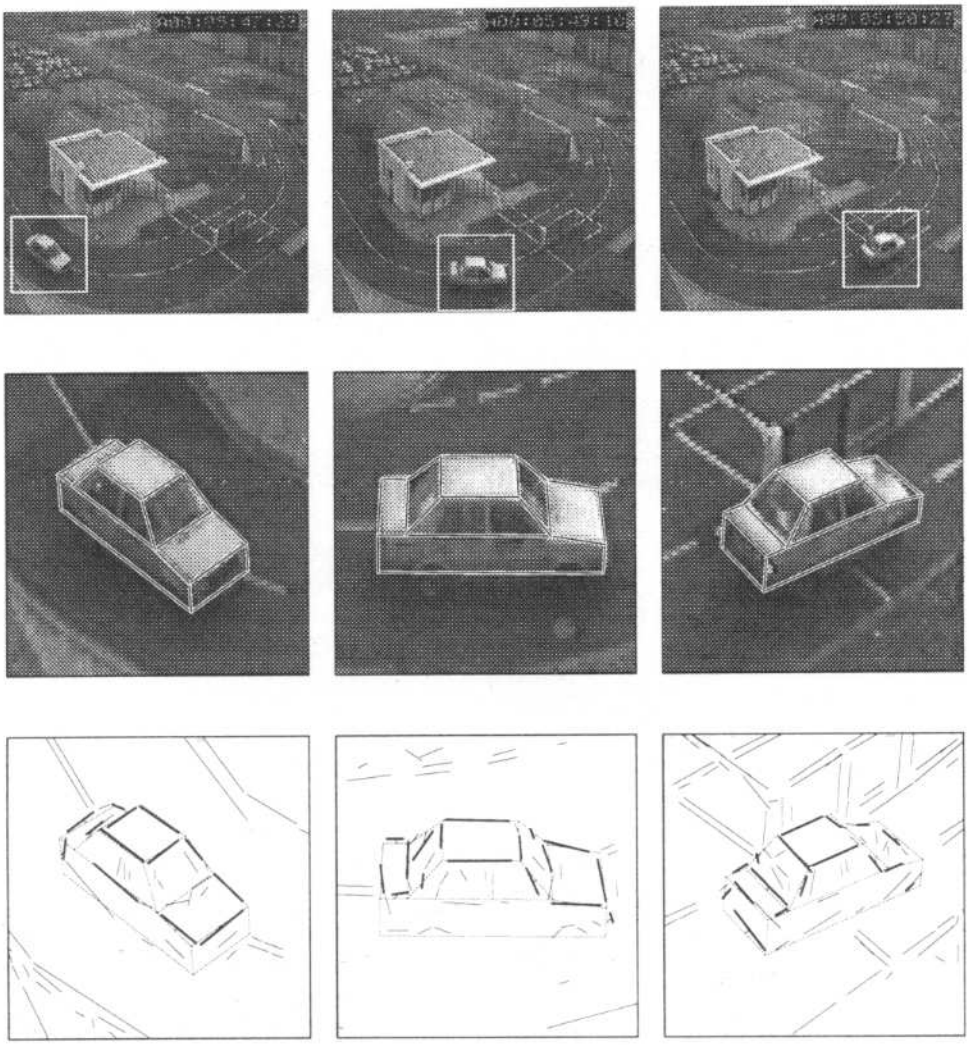

Fig. 9. The first row shows the 4th, 41st, and 79th frame of an image sequence. The three images in the middle row give an enlarged section of the model matched to the car moving in the image sequence. The lower three figures exhibit the image line segments and model segments (dashed lines) in the same enlarged section as in the middle row; image line segments that have been matched to the model are given by fat lines.

illumination conditions not even possible, to use shadow edges in the matching process. The matched models for the three upper frames are illustrated in the middle row of figure 9, with more details given in the lower three figures where we see the extracted straight lines, the back projected model segments (dashed lines), and the matched data segments, emphasized by thick lines.

The resultant object trajectories will be used as inputs to the process of associating motion verbs to trajectory segments (Koller, Heinze, & Nagel 1991; Kollnig 1992). Since such subsequent analysis steps are very sensitive to noise, we attempt to obtain smoother object trajectories using small process noise for the magnitude of the velocity v and the angular velocity ω. In this and in most of the subsequent experiments, we therefore use a process noise of $\sigma_v = 10^{-3}$ (m/frame) and $\sigma_\omega = 10^{-4}$ (rad/frame). Given this σ_v and σ_ω, the majority of the translational and angular accelerations are assumed to be $\dot{v} < \sigma_v/\tau^2 = 0.625$ m/s^2 and $\dot{\omega} < \sigma_\omega/\tau^2 = 6.25 \cdot 10^{-2}$ rad/s^2, respectively, with $\tau = t_{k+1} - t_k = 40$ ms.

The bootstrap phase is performed using the first three frames in order to estimate initial magnitudes of the velocities v and ω. Since the initially detected moving region does not always correctly span the image of the moving object, we used values $\sigma_{p_{x_0}} = \sigma_{p_{y_0}} = 0.1$ m. An initial value for the covariance in the orientation ϕ is roughly estimated by considering the differences in the orientation between the clustered displacement vectors, that is, $\sigma_{\phi_0} = 0.03$ rad. We used $\sigma_{||} = 2.4$ and $\sigma_{\perp} = 0.8$ as the uncertainties of a data-edge segment parallel and perpendicular to the edge direction. As a threshold for the computed Mahalanobis distance used for establishing correspondences of model and data segments—equation (4)—we have chosen $d_\tau = 6$.

The car has been tracked during the entire sequence of 80 frames with an average number of about 16 line segment correspondences per frame. The computed trajectory for this moving car is given in figure 10 and the estimated translational velocity v as well as the angular velocity ω is given in figure 11.

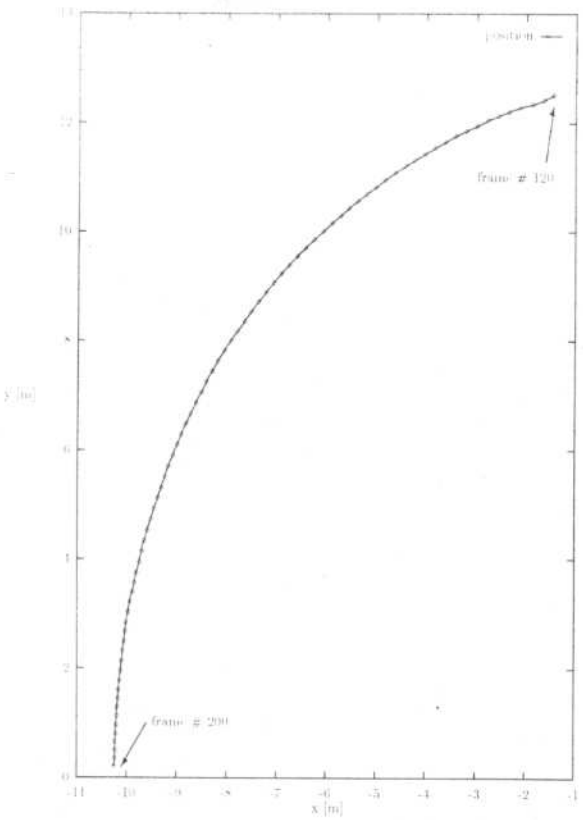

Fig. 10. The estimated position of the moving car of figure 9.

Incoming Car and Parking Car

In order to further test our motion model we used an image subsequence of about 400 frames of the same parking area. In this subsequence, one car is waiting in front of the barrier until the barrier raises, while in the background a second car appears behind the gatekeeper's lodge, turns to the right and manuevers backward into a parking slot. Here one has to cope with two different motions which cannot be explicitly captured by our motion model: the high angular acceleration during the transition from circular to straight translational motion carried out by the incoming car (actually a clothoidal path motion) and the transition from positive to negative translational velocity during the manuevering of the parking car. But exactly such cases illustrate the real meaning of the introduction of process noise into recursive estimation. By increasing the process noise we have been able to overcome the weakness of our motion model: the standard deviations used are $\sigma_v = 3.3 \cdot 10^{-3}$ m/frame and $\sigma_\omega = 10^{-2}$ rad/frame for the incoming car and $\sigma_v = 2 \cdot 10^{-3}$ m/frame and $\sigma_\omega = 2 \cdot 10^{-4}$ rad/frame for the parking car.

In order to overcome the tracking difficulty due to the tiny viewing angle of the projection of the parking car, we have slightly increased the threshold for the Mahalanobis distance to $d_\tau = 7$ as well as the start covariances to $\sigma_{p_{x_0}} = \sigma_{p_{y_0}} = 0.2$ m and $\sigma_{\phi_0} = 0.03$ rad. The uncertainty in the start position of the incoming

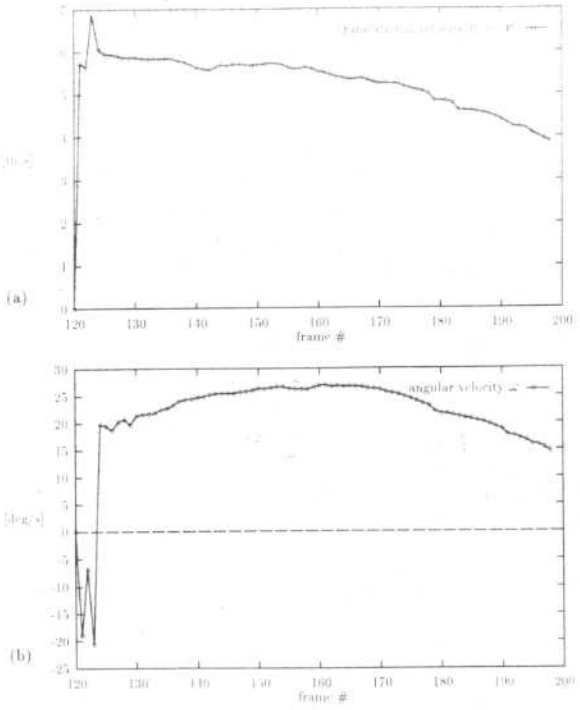

Fig. 11. The estimated translational (a) and angular (b) velocity of the moving car of figure 9.

car caused by the occluding barrier is captured by higher values used for the start covariances: $\sigma_{p_{x_0}} = \sigma_{p_{y_0}} = 0.5$ m and $\sigma_{\phi_0} = 0.1$ rad. The other parameters have remained the same as in the previous experiment. In this experiment we used the shadow edges as additional line segments in the matching process as described in section 5.

Some intermediate results of tracking the parking car in the background are given in figure 12 and intermediate results of the incoming car are shown in figure 13. Since the parking car in the background as well as the incoming car initially waiting in front of the barrier are partially occluded during the bootstrap phase, two single static matches have not been sufficient to initialize the motion model reasonably well. The motion model of the parking car in the background, therefore, is initialized during the first four frames and the motion model of the incoming car is initialized during the first five frames. The resultant trajectories of the two cars are shown in figure 14 while the estimated translational and angular velocities are given in figure 15.

The trajectory of the incoming car exhibits another problem that arises in cases in which a vehicle disappears from the field of view. This problem is similar to that of a partial occlusion except that no data segments whatsoever could be found outside the image. In this way the tracking algorithm tries to keep the trajectory of an object inside the region of the scene covered by the field of view because data segments at the boundary of the image are matched to some model segments.

Multilane Street Intersection

The next group of experiments involved an image subsequence of about 50 frames of a much frequented multilane street intersection. In this sequence there are several moving vehicles with different shapes and dimensions, all vehicles turning to the left (figure 16).

The size of the images of the moving vehicles varies in this sequence from 30×60 to 20 ×40 pixels in a frame. Figure 4 shows some intermediate steps in extracting the line segments. We explicitly present this figure in order to give an idea of the complexity of the task of detecting and tracking a moving vehicle spanning such a small area in the image. We used the same values for the process noise and the threshold for the Mahalanobis distance as in our first experiment: $\sigma_v = 10^{-3}$ m/frame, $\sigma_\omega = 10^{-4}$ rad/frame and $d_r = 6$. Due to a larger uncertainty in the estimation of initial values caused by the tiny viewing angle, we used the same initial covariances as for the incoming car in the previous experiment: $\sigma_{p_{x_0}} = \sigma_{p_{y_0}} = 0.5$ *m and* $\sigma_{\phi_0} = 0.1$ rad. It turned out that in this way static matches in the first two frames are sufficient for an initial estimation of v and ω. In this experiment we also used the shadow edges as additional line segments in the matching process as described in section 5.

In the upper part of figure 17 we see three frames out of the same image sequence. In the middle part of figure 17, the matched model of a taxi is given as an enlarged section from the three upper images. In the lower three figures the correspondences of image line segments and the model line segments are given.

Figure 18 shows the resultant object trajectories of the cars numbered #5, #6, and #8 tracked in figures 16, 17, and 21, respectively. The estimated translational velocity v and angular velocity ω is given in figure 19. It can be clearly observed in Figure 19 that the estimates for v and ω stabilize with increasing frame number, that is, with (observation) time.

Five of the vehicles—actually the five brightest ones—appearing in the first frame of this image sequence have been tracked throughout the entire sequence. The

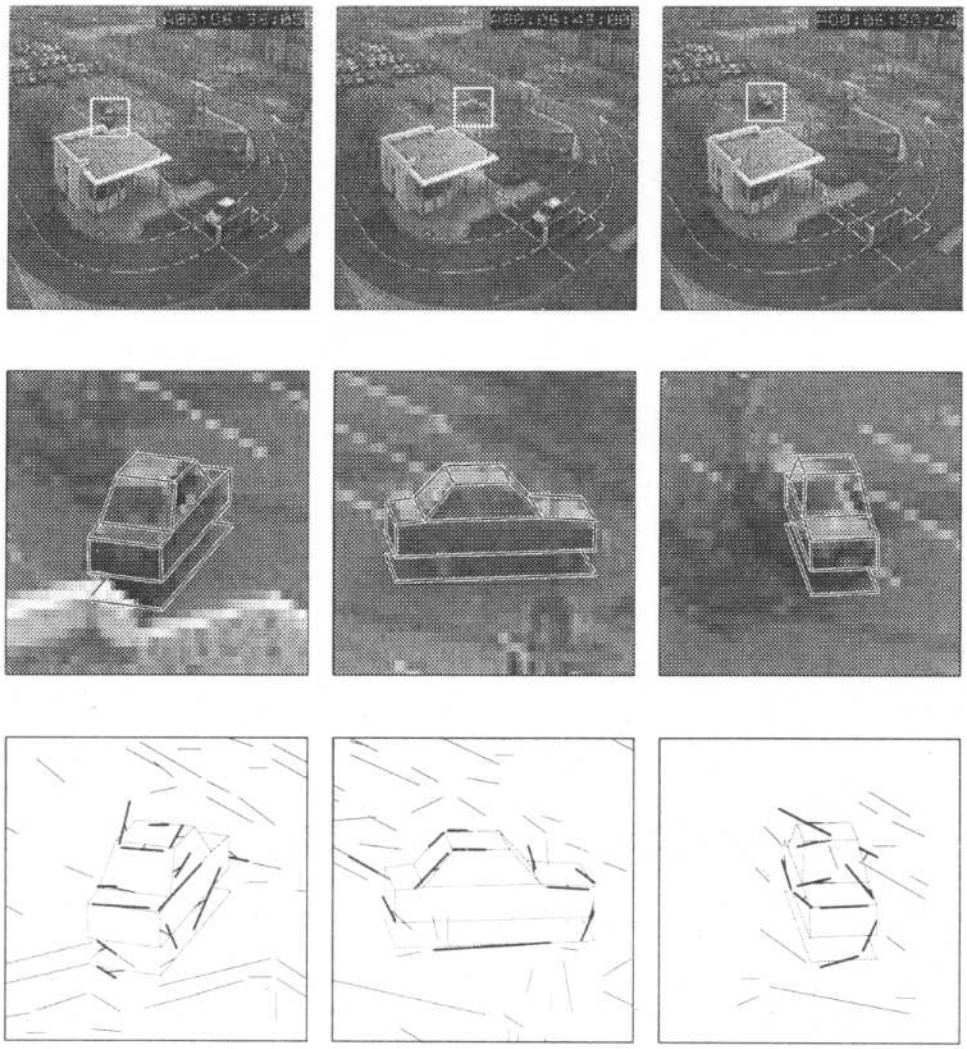

Fig. 12. The first row shows the 30th, 200th, and 399th frame of an image sequence of about 400 frames (frames #1300–1699). The three images in the middle row give an enlarged section of the model matched to the moving car performing a parking maneuver in the image sequence. The lower three figures exhibit image line segments and model segments (dashed lines) in the same enlarged section as in the middle row; image line segments that have been matched to the model are give by fat lines.

reason for the failure in tracking the other vehicles has been the inability of the initialization step to provide the system with appropriate initial values in order to formulate object hypotheses.

In order to test our tracking algorithm even in cases where appropriate initial values have not been provided automatically, for example for the dark car at the lower left corner in the images of figure 21, we provided the estimated initial values interactively. Figure 20 shows the initial value for the dark car #8 represented as an image region assumed to cover the image of the moving object. Since this image region is computed by a segmentation step which clusters coherently moving image features, which—due to the sparse internal structure of the dark car—gives only very few hints to image features, the image region representing the moving car does not correctly span the image of the car (upper left image of figure 20). Since we further exploit only the center and the averaged moving direction of the back projected image features of this image region to compute initial values, it is sufficient to shift the center of the model, as we have done for the initial values in the first three frames for the dark car #8 (see the upper right image of figure 20). The computed optimal position

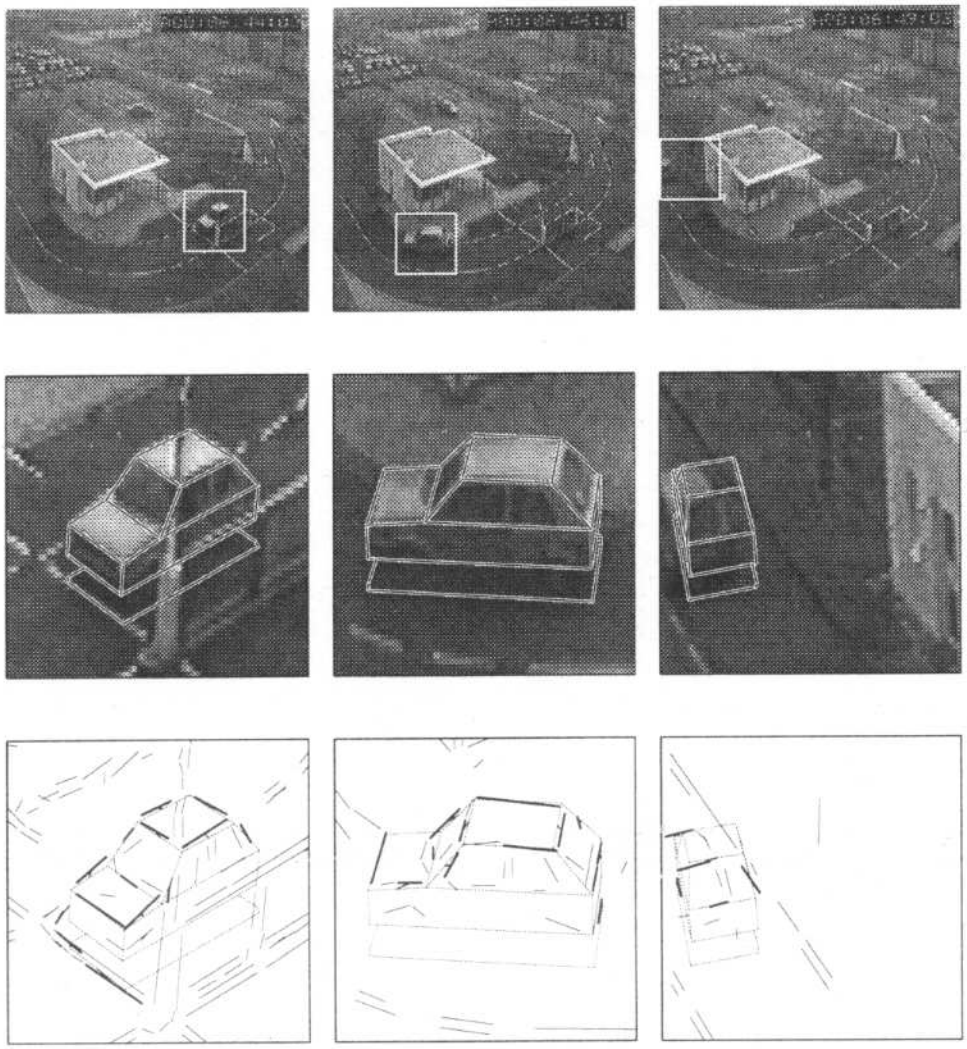

Fig. 13. The first row shows the 1327th, 1396th, and 1653rd frame of the same image sequence as in figure 12. The car in front of the barrier is accelerating from frame #1522 onwards and disappears at about frame #1660 at the left side. The three images in the middle row give an enlarged section of the model matched to the accelerating car moving in the image sequence. The lower three figures exhibit image line segments and model segments (dashed lines) in the same enlarged section as in the middle row; image line segments that have been matched to the model are given by fat lines.

and orientation for the unshifted (left column) and shifted model center (right column) is given in the lower two images of figure 20.

In this way even the dark car with a very poor internal structure—resulting in very few edge segments—could be tracked during this image sequence despite a partial occlusion by a traffic light. This example also shows the necessity of using shadow edges in cases in which the images of vehicles exhibit only a very poor internal structure and the only chance to track such a vehicle is to match image edges of the outer contour.

3 Conclusion and Future Work

Our task has been to build a system that will be able to compute smooth trajectories of vehicles in traffic scenes and will be extendible to incorporate a solution to the problem of classifying the vehicles according to computed shape parameters. We have considered this task to be difficult due to the complex illumination conditions and the cluttered environment of real-world traffic scenes as well as the small effective field of view that is spanned by the projection of each vehicle given

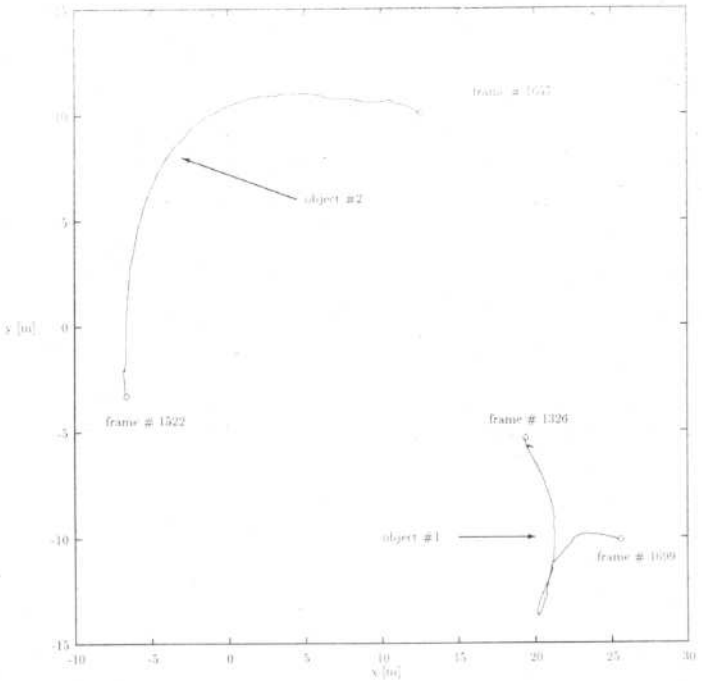

Fig. 14. The estimated position of the parking car (object #1) of figure 12 and the incoming car (object #2 in figure 13. The little arc at the end of the trajectory of the incoming car (object #2) is caused by the fact that it begins to leave the field of view and thus no corresponding data segments are found at the left side of the car to correctly estimate the translational motion.

a stationary camera. In all experiments quoted from the literature (see section 2) the projected area of objects covers a quite high portion of the field of view. Furthermore, only two of them (Worrall et al. 1991; Evans 1990) are tested under outdoor illumination conditions (road traffic and landing of an aircraft, respectively).

In order to accomplish the above-mentioned tasks we have applied the following constraints. We restricted the degrees of freedom of the transformation between model and camera from six to three by assuming that a vehicle is moving on a plane known a priori by calibration. We considered only a simple time-coherent motion model that can be justified by the high sampling rate (25 frames per second) and the knowledge that vehicles do not maneuver abruptly.

The second critical point we have been concerned about is the establishment of good initial matches and pose estimates. Most tracking approaches do not emphasize the severity of this problem of establishing a number of correct correspondences in the starting phase and feeding the recursive estimator with quite reasonable initial values. Again we have used the a priori knowledge of the street-plane position and the results of clustering image-domain cues into object hypotheses of a previous step. Thus we have been able to start the tracking process automatically with a simple matching scheme and feed the recursive estimator with values of low error covariance.

The third essential point we have addressd is the additional consideration of shadows. Data-line segments arising from shadows are not treated any more as disturbing data like markings on the road, but they contribute to the stabilization of the matching process.

Our work will be continued by the following steps. First, the matching process should be enhanced by introducing a search tree. In spite of the good initial pose estimates, we are still confronted occasionally with totally false matching combinations due to the highly ambiguous structure of our current vehicle model. Second, the motion model can be extended to capture translational as well as angular acceleration. Third, the generic-vehicle model enables a simple adaptation to the image data by varying the shape parameters. These shape parameters should be added as unknowns and estimated along time. Preliminary experiments (Koller 1992) have shown that this task can be carried out successfully.

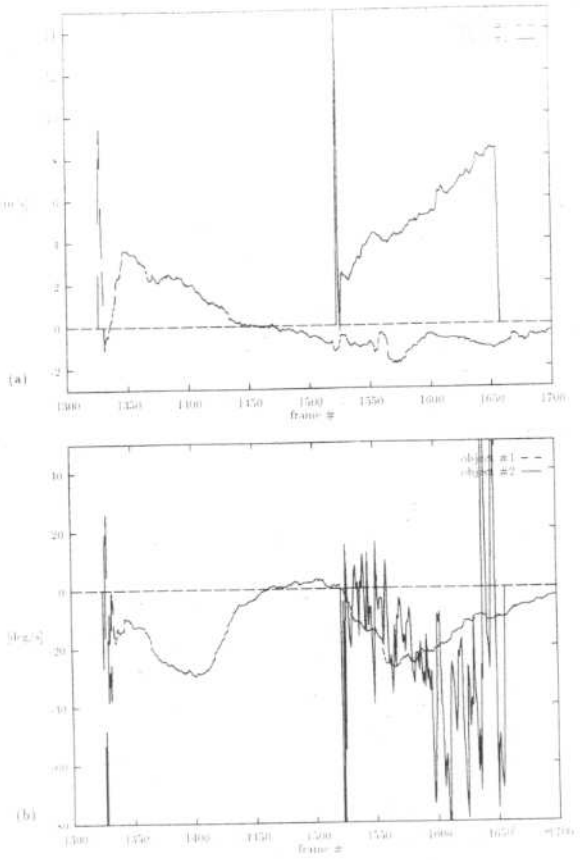

Fig. 15. The estimated translational (a) and angular (b) velocities of the parking car of figure 12 and the incoming car in figure 13. Due to the previously mentioned problems in using too small a value for the process noise, a higher value for process noise is necessary to track the incoming car (object #2). This higher process noise causes a larger uncertainty in the estimation of the angular velocity.

Acknowledgments

We thank G. Winkler for his comments and careful reading of an earlier version of this article. Our thanks go to T. Thórhallson for his contribution to a critical phase of these investigations. The financial support of the first author of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and of the second author by the Deutscher Akademischer Austauschdienst (DAAD, German Academic Exchange Service) are gratefully acknowledged.

Appendix: Computation of Jacobians

The use of a recursive estimator requires the Jacobians of the measurement as well as of the transition function.

The Jacobian of the measurement function (15) has the following block structure:

$$H(\mathbf{x}) := \frac{\partial \mathbf{h}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial \mathbf{X}_1(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial \mathbf{X}_n(\mathbf{x})}{\partial \mathbf{x}} \end{pmatrix} \tag{22}$$

where X_i is the 4×1 parameter vector of the ith edge segment given by

$$\left.\begin{aligned} \mathbf{c}_i &= \frac{1}{2}(\mathbf{x}_{i,1} + \mathbf{x}_{i,2}) \\ \theta_i &= arctan\left(\frac{y_{i,2} - y_{i,1}}{x_{i,2} - x_{i,1}}\right) \\ l_i &= \sqrt{(x_{i,2} - x_{i,1})^2 + (y_{i,2} - y_{i,1})^2} \end{aligned}\right\} \tag{23}$$

Here, $\mathbf{x}_{i,1} = (x_{i,1}, y_{i,1})^T$ *and* $\mathbf{x}_{i,2}, y_{i,2})^T$ are the endpoints of the ith edge segment. Thus the Jacobian is a $4n\times5$ matrix if n is the number of the edge segments and 5 corresponds to the number of the unknown state components.

The following chain of operations gives the dependence of a model point in the image on the state.

$$\mathbf{x}_{w_i} = R_{wm}(\phi)\mathbf{x}_{m_i} + \mathbf{p}_m \tag{24}$$

$$\mathbf{x}_{c_i} = R_{cw}\mathbf{x}_{w_i} + \mathbf{t}_c \tag{25}$$

$$\mathbf{x}_i = \begin{pmatrix} \frac{f_x x_{c_i}}{z_{c_i}} \\ \frac{f_y y_{ci}}{z_{c_i}} \end{pmatrix} + \mathbf{x}_o \tag{26}$$

The indexes m, w, and c are used for the model, world, and camera coordinate system, respectively. The transformation $(R_{cw}, \mathbf{t}_c)$ from world to camera as well as the focal lengths (f_x, f_y) and the optical center $\mathbf{x}_o$ are constant and a priori known by calibration. It turns out that the measurements depend only on position $\mathbf{p}_m$ and orientation ϕ of the object but not on the velocities. Hence, the fourth and fifth column of the Jacobian are zero.

For endpoints of shadow edge-segments one step is added to the chain of the transformations:

$$\mathbf{x}_{s_i} = \mathbf{x}_{w_i} - \frac{z_{w_i}}{\cos\theta_s}\mathbf{n} \tag{27}$$

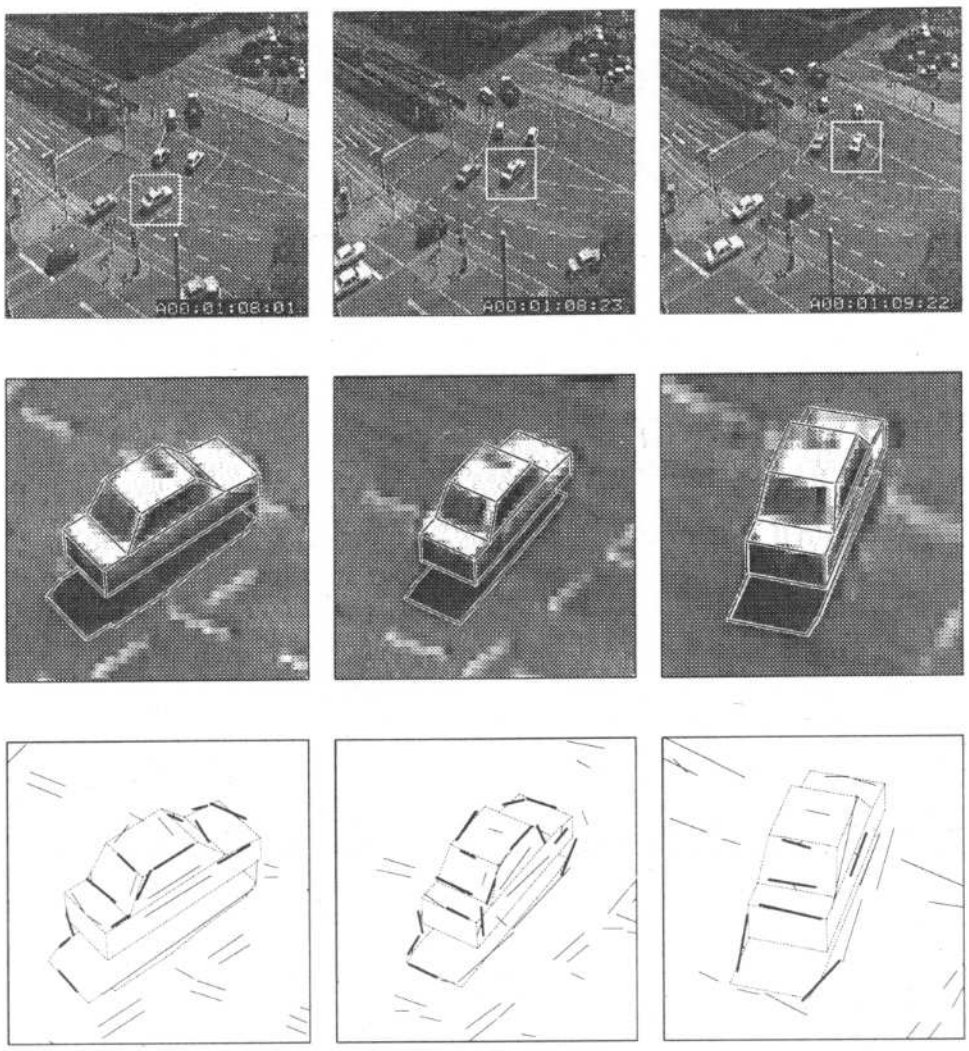

Fig. 17. Analogous to figure 16. The middle row shows an enlarged section of the model matched to the taxi (object #6) moving in the center of the frame.

Analogously, the expression for a shadow edge segment reads

$$\frac{\partial \mathbf{X}_i(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{X}_i(\mathbf{x})}{\partial \mathbf{x}_{c_\alpha}} \frac{\partial \mathbf{x}_{c_\alpha}}{\partial \mathbf{x}_{s_\alpha}} \frac{\partial \mathbf{x}_{s_\alpha}}{\partial \mathbf{x}_{w_\alpha}} \frac{\partial \mathbf{x}_{w_\alpha}}{\partial \mathbf{x}} + \frac{\partial \mathbf{X}_i(\mathbf{x})}{\partial \mathbf{x}_{c_\beta}} \frac{\partial \mathbf{x}_{c_\beta}}{\partial \mathbf{x}_{s_\beta}} \frac{\partial \mathbf{x}_{s_\beta}}{\partial \mathbf{x}_{w_\beta}} \frac{\partial \mathbf{x}_{w_\beta}}{\partial \mathbf{x}} \quad (32)$$

The derivatives of a vertex $\gamma = \{\alpha, \beta\}$ with respect to $(\mathbf{p}_m, \phi)$ are given by

$$\frac{\partial \mathbf{x}_{c_\gamma}}{\partial \mathbf{x}_{w_\gamma}} = R_{cw}$$

$$\frac{\partial \mathbf{x}_{w_\gamma}}{\partial \mathbf{p}_m} = \mathbf{I}_3 \quad (33)$$

$$\frac{\partial \mathbf{x}_{w_\gamma}}{\partial \phi} = \frac{\partial R_{wm}(\phi)}{\partial \phi} \mathbf{x}_{m_\gamma}$$

with

$$\frac{\partial R_{wm}(\phi)}{\partial \phi} = \begin{pmatrix} -\sin\phi & -\cos\phi & 0 \\ \cos\phi & -\sin\phi & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (34)$$

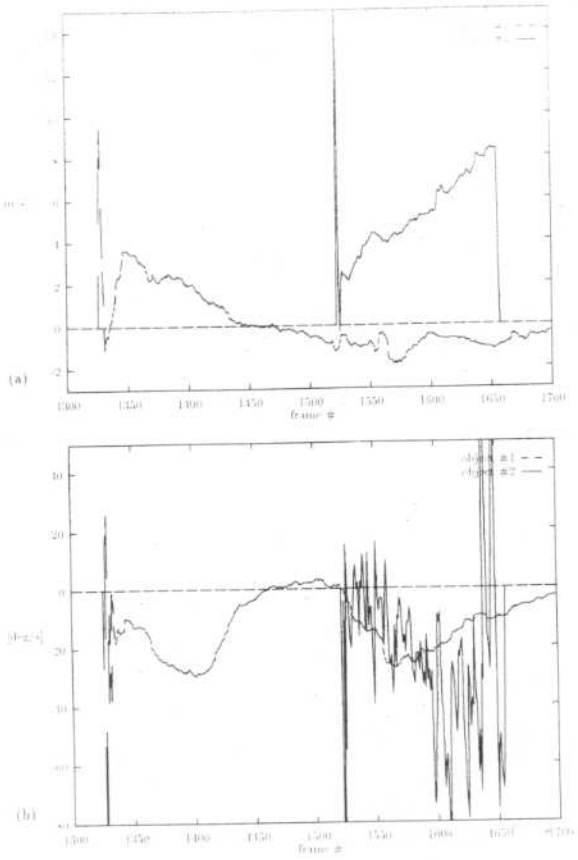

Fig. 15. The estimated translational (a) and angular (b) velocities of the parking car of figure 12 and the incoming car in figure 13. Due to the previously mentioned problems in using too small a value for the process noise, a higher value for process noise is necessary to track the incoming car (object #2). This higher process noise causes a larger uncertainty in the estimation of the angular velocity.

Acknowledgments

We thank G. Winkler for his comments and careful reading of an earlier version of this article. Our thanks go to T. Thórhallson for his contribution to a critical phase of these investigations. The financial support of the first author of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and of the second author by the Deutscher Akademischer Austauschdienst (DAAD, German Academic Exchange Service) are gratefully acknowledged.

Appendix: Computation of Jacobians

The use of a recursive estimator requires the Jacobians of the measurement as well as of the transition function.

The Jacobian of the measurement function (15) has the following block structure:

$$H(\mathbf{x}) := \frac{\partial \mathbf{h}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial \mathbf{X}_1(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial \mathbf{X}_n(\mathbf{x})}{\partial \mathbf{x}} \end{pmatrix} \tag{22}$$

where X_i is the 4×1 parameter vector of the ith edge segment given by

$$\left.\begin{aligned} \mathbf{c}_i &= \frac{1}{2}(\mathbf{x}_{i,1} + \mathbf{x}_{i,2}) \\ \theta_i &= arctan\left(\frac{y_{i,2} - y_{i,1}}{x_{i,2} - x_{i,1}}\right) \\ l_i &= \sqrt{(x_{i,2} - x_{i,1})^2 + (y_{i,2} - y_{i,1})^2} \end{aligned}\right\} \tag{23}$$

Here, $\mathbf{x}_{i,1} = (x_{i,1}, y_{i,1})^T$ *and* $\mathbf{x}_{i,2}, y_{i,2})^T$ are the endpoints of the ith edge segment. Thus the Jacobian is a $4n\times5$ matrix if n is the number of the edge segments and 5 corresponds to the number of the unknown state components.

The following chain of operations gives the dependence of a model point in the image on the state.

$$\mathbf{x}_{w_i} = R_{wm}(\phi)\mathbf{x}_{m_i} + \mathbf{p}_m \tag{24}$$

$$\mathbf{x}_{c_i} = R_{cw}\mathbf{x}_{w_i} + \mathbf{t}_c \tag{25}$$

$$\mathbf{x}_i = \begin{pmatrix} \frac{f_x x_{c_i}}{z_{c_i}} \\ \frac{f_y y_{ci}}{z_{c_i}} \end{pmatrix} + \mathbf{x}_o \tag{26}$$

The indexes m, w, and c are used for the model, world, and camera coordinate system, respectively. The transformation $(R_{cw}, \mathbf{t}_c)$ from world to camera as well as the focal lengths (f_x, f_y) and the optical center $\mathbf{x}_o$ are constant and a priori known by calibration. It turns out that the measurements depend only on position $\mathbf{p}_m$ and orientation ϕ of the object but not on the velocities. Hence, the fourth and fifth column of the Jacobian are zero.

For endpoints of shadow edge-segments one step is added to the chain of the transformations:

$$\mathbf{x}_{s_i} = \mathbf{x}_{w_i} - \frac{z_{w_i}}{\cos\theta_s}\mathbf{n} \tag{27}$$

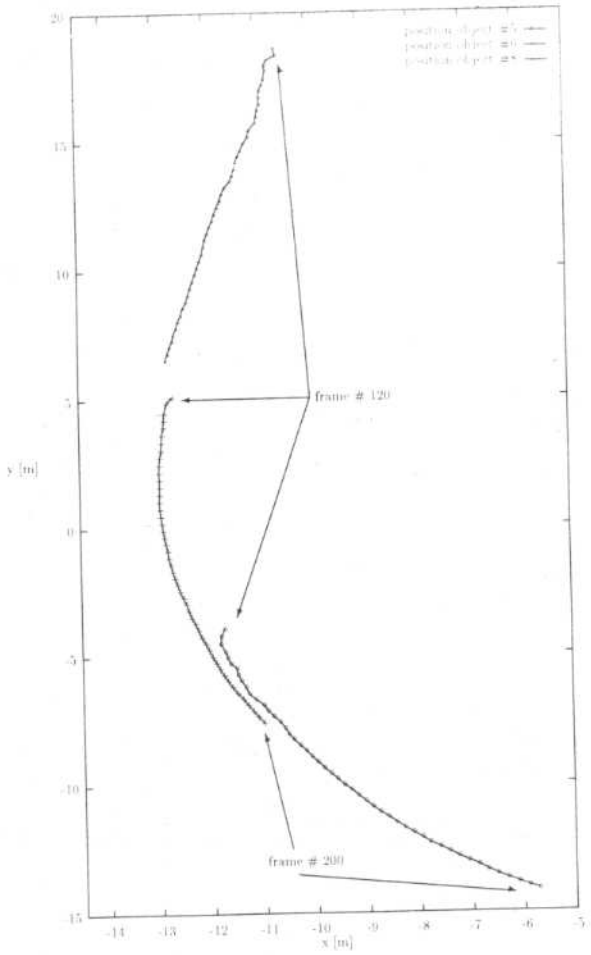

Fig. 18. The estimated positions of the moving cars in figures 16 (object #5), 17 (object #6), and 21 (object #8).

What remains to be computed are the derivatives of the parameter vector $\mathbf{X}_i$ with respect to the camera coordinates $\mathbf{x}_{c_\gamma}$:

$$\frac{\partial \mathbf{X}_i(\mathbf{x})}{\partial \mathbf{x}_{c_\gamma}} = \begin{pmatrix} \frac{\partial \mathbf{c}_i}{\partial \mathbf{x}_{c_\gamma}} \\ \frac{\partial \theta_i}{\partial \mathbf{x}_{c_\gamma}} \\ \frac{\partial l_i}{\partial \mathbf{x}_{c_\gamma}} \end{pmatrix} \tag{35}$$

We differentiate first the midpoint vector and obtain

$$\frac{\partial \mathbf{c}_i}{\partial \mathbf{x}_{c_\gamma}} = \frac{1}{2}\,\Pi_\gamma \tag{36}$$

where Π_γ is the following Jacobian of a perspective projection

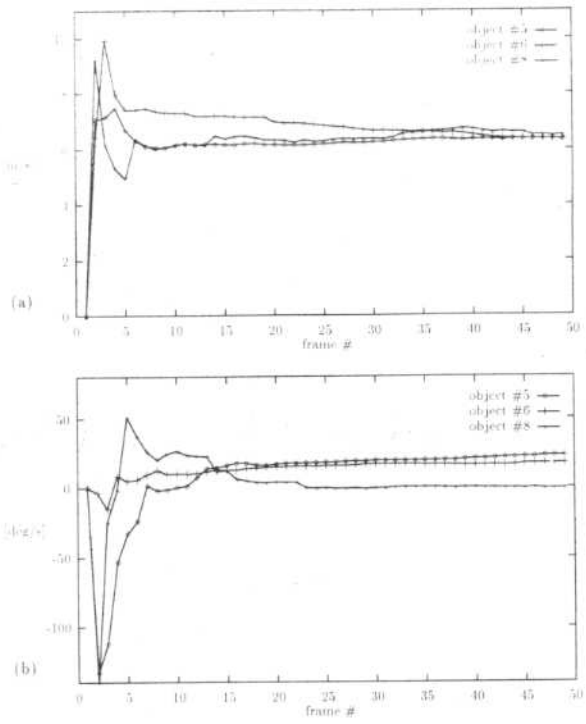

Fig. 19. The estimated translational (a) and angular (b) velocities of the moving cars in figure 16 (object #5), 17 (object #6), and 21 (object #8).

$$\Pi_\gamma := \frac{\partial \mathbf{x}_\gamma}{\partial \mathbf{x}_{c_\gamma}} = \frac{1}{z_{c_\gamma}} \begin{pmatrix} f_x & 0 & -f_x \dfrac{x_{c_\gamma}}{z_{c_\gamma}} \\ 0 & f_y & -f_y \dfrac{y_{c_\gamma}}{z_{c_\gamma}} \end{pmatrix} \tag{37}$$

We, then, differentiate the angle θ_i with respect to the first vertex α

$$\frac{\partial \theta_i}{\partial \mathbf{x}_{c_\alpha}} = \frac{1}{1 + \left(\dfrac{\Delta y_i}{\Delta x_i}\right)^2} \left[\frac{\Delta y_i}{(\Delta x_i)^2}\,,\; -\frac{1}{\Delta x_i} \right] \frac{\partial \mathbf{x}_\alpha}{\partial \mathbf{x}_{c_\alpha}}$$

$$= \frac{1}{l_i^2} \begin{pmatrix} \Delta y_i \\ -\Delta x_i \end{pmatrix}^T \Pi_\alpha \tag{38}$$

with $(\Delta x_i,\, \Delta y_i)^T = \mathbf{x}_\beta - \mathbf{x}_\alpha$ and, accordingly with respect to the second vertex β;

$$\frac{\partial \theta_i}{\partial \mathbf{x}_{c_\beta}} = \frac{1}{l_i^2} \begin{pmatrix} -\Delta y_i \\ \Delta x_i \end{pmatrix}^T \Pi_\beta \tag{39}$$

Last, we compute the derivatives of the length l_i of the line segment

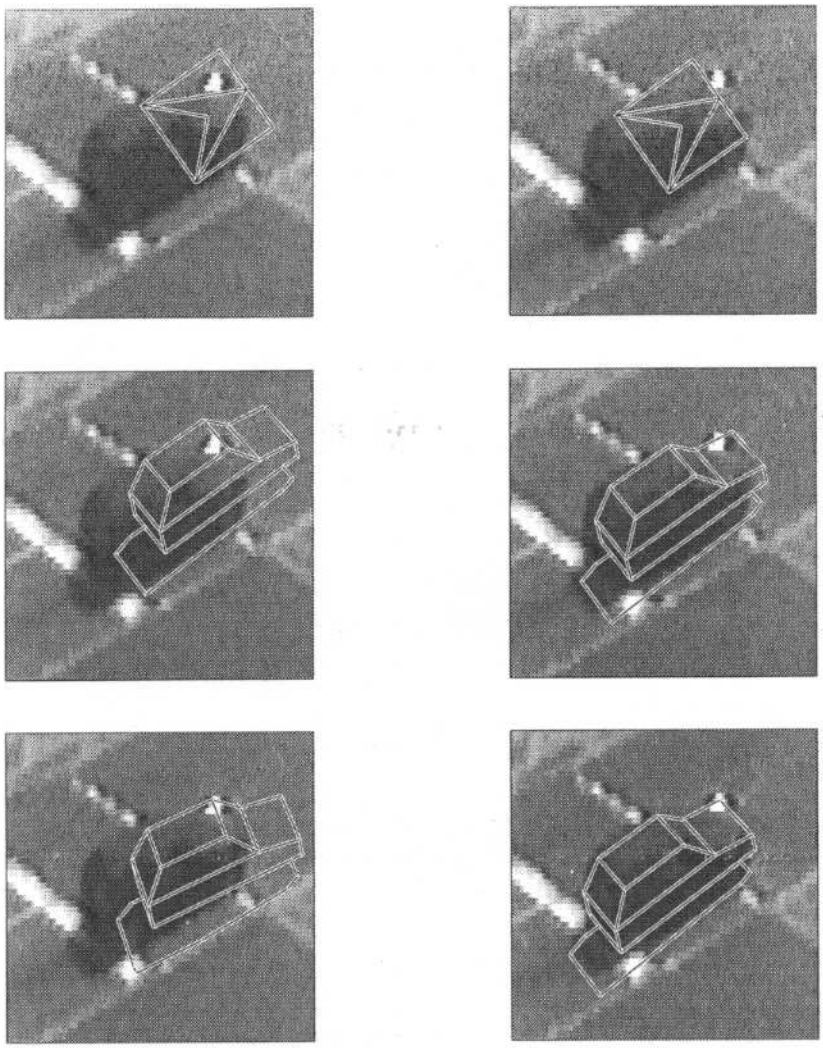

Fig. 20. In order to even test our tracking algorithm in cases in which no appropriate initial values are computed, we are forced to slightly manipulate the computed initial values. The left column shows the initialization without manipulation: in the upper image the initial value—represented as an image segment—is computed using a clustering approach of optical flow vectors. The left image of the middle row shows the corresponding initial model instance while the lower image shows the computed optimal position and orientation for this model initialization. The right column shows the same intermediate steps using a slight shift of the center of the initialized region, resulting in a significantly better fit.

$$\frac{\partial l_i}{\partial \mathbf{x}_{c_\alpha}} = \frac{1}{l_i} \begin{pmatrix} -\Delta x_i \\ -\Delta y_i \end{pmatrix}^T \Pi_\alpha \tag{40a}$$

as well as

$$\frac{\partial l_i}{\partial \mathbf{x}_{c_\beta}} = \frac{1}{l_i} \begin{pmatrix} \Delta x_i \\ \Delta y_i \end{pmatrix}^T \Pi_\beta \tag{40b}$$

and we summarize

$$\frac{\partial \mathbf{X}_i(\mathbf{x})}{\partial \mathbf{x}_{c_\alpha}} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \\ \Delta y_i/l_i^2 & -\Delta x_i/l_i^2 \\ -\Delta x_i/l_i & -\Delta y_i/l_i \end{pmatrix} \Pi_\alpha \tag{41}$$

$$\frac{\partial \mathbf{X}_i(\mathbf{x})}{\partial \mathbf{x}_{c_\beta}} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \\ -\Delta y_i/l_i^2 & \Delta x_i/l_i^2 \\ \Delta x_i/l_i & \Delta y_i/l_i \end{pmatrix} \Pi_\beta \tag{42}$$

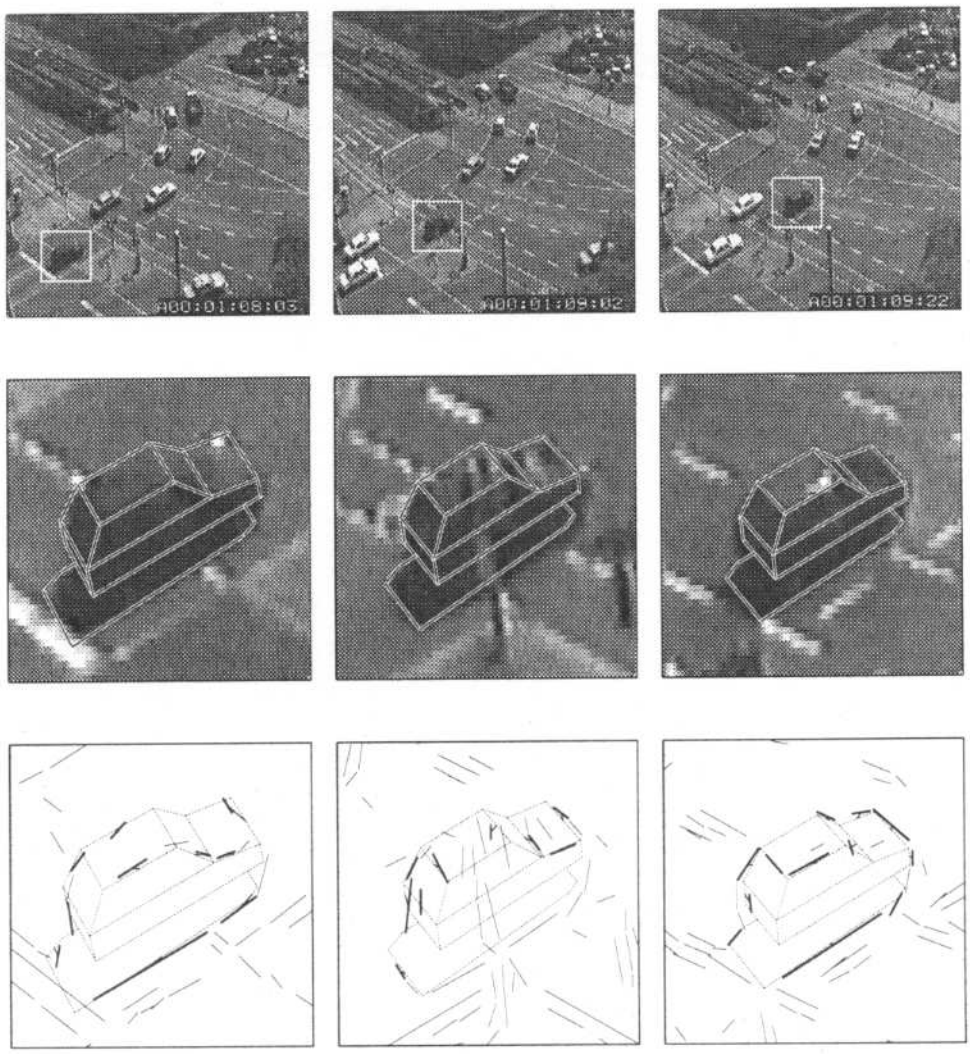

Fig. 21. The middle row (analogous to figure 16) shows an enlarged section of the model matched to the black car (object #8, see upper row) moving at the lower left corner of the frame. This example demonstrates the necessity of using shadow edges in the matching process, since there is no chance to extract internal model edges in a reasonable number and quality to match and track the model.

In order to apply the chain rule (32), an additional derivative of the coordinates of the shadow point with respect to world coordinates is required and obtained from (27):

$$\frac{\partial \mathbf{x}_{s_\gamma}}{\partial \mathbf{x}_{w_\gamma}} = \begin{pmatrix} 1 & 0 & -\cos\phi_s \tan\theta_s \\ 0 & 1 & -\sin\phi_s \tan\theta_s \\ 0 & 0 & 0 \end{pmatrix} \tag{43}$$

The computation of the Jacobian of the transition function is not so cumbersome. We differentiate (12) and obtain

$$F(\mathbf{x}) = \begin{pmatrix} 1 & 0 & v_k\tau C & \tau S & v_k\tau \dfrac{\partial S}{\partial \omega_k} & \mathbf{0}^T \\ 0 & 1 & v_k\tau S & -\tau C & -v_k\tau \dfrac{\partial C}{\partial \omega_k} & \mathbf{0}^T \\ 0 & 0 & 1 & 0 & \tau & \mathbf{0}^T \\ 0 & 0 & 0 & 1 & 0 & \mathbf{0}^T \\ 0 & 0 & 0 & 0 & 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{44}$$

with

$$S = \frac{\sin(\phi_k + \omega_k\tau) - \sin\phi_k}{\omega_k\tau},$$

$$C = \frac{\cos(\phi_k + \omega_k\tau) - \cos\phi_k}{\omega_k\tau} \tag{45}$$

and

$$\frac{\partial S}{\partial \omega_k} = \frac{1}{\omega_k}\left[\cos(\phi_k + \omega_k\tau) - \frac{\sin(\phi_k + \omega_k\tau) - \sin\phi_k}{\omega_k\tau}\right],$$

$$\frac{\partial C}{\partial \omega_k} = \frac{1}{\omega_k}\left[-\sin(\phi_k + \omega_k\tau) - \frac{\cos(\phi_k + \omega_k\tau) - \cos\phi_k}{\omega_k\tau}\right] \tag{46}$$

References

Bar-Shalom, Y., and Fortmann, T.E., 1988. *Tracking and Data Association*. Academic Press: New York.

Broida, T.J., Chandrashekhar, S., and Chellappa, R., 1990. Recursive 3-d motion estimation from a monocular image sequence, *IEEE Trans. Aerospace Electron. Syst.* 26: 639–656.

Deriche, R., and Faugeras, O.D., 1990. Tracking line segments, *Image Vis. Comput.* 8: 261–270.

Evans, R., 1990. Kalman filtering of pose estimates in applications of the rapid video rate tracker, *Proc. Brit. Mach. Vis. Conf.*, Oxford, pp. 79–84, September 24–27.

Gelb, A., ed., 1974. *Applied Optimal Estimation*. MIT Press: Cambridge, MA and London.

Gennery, D.B., 1982. Tracking known three-dimensional objects, *Proc. Conf. Amer. Assoc. Artif. Intell.*, Pittsburgh, pp. 13–17, August 18–20.

Gennery, D.B., 1992. Visual tracking of known three-dimensional objects, *Intern. J. Comput. Vis.* 7: 243–270.

Grimson, W.E.L., 1990a. The combinatorics of object recognition in cluttered environments using constrained search, *Artificial Intelligence* 44: 121–165.

Grimson, W.E.L., 1990b. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press: Cambridge, MA.

Harris, C., and Stennet, C., 1990. RAPID—a video rate object tracker, *Proc. Brit. Mach. Vis. Conf.*, Oxford, pp. 73–77, September 24–27.

Jazwinski, A.H., 1970. *Stochastic Processes and Filtering Theory*. Academic Press: New York and London.

Koller, D., 1992. Detektion, Verfolgung und Klassifikation bewegter Objekte in monokularen Bildfolgen am Beispiel von Straßenverkehrsszenen. Dissertation, Fakultät für Informatik der Universität Karlsruhe (TH), available as vol. DISKI 13, *Dissertationen zur Künstlichen Intelligenz*, infix-Verlag, Sankt Augustin, Germany.

Koller, D., Heinze, N., and Nagel, H.-H., 1991. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs, *Conf. Comput. Vis. Patt. Recog.*, Lahaina, Maui, Hawaii, pp. 90–95, June 3–6.

Koller, D., Daniilidis, K., Thórhallson, T., and Nagel, H.-H., 1992. Model-based object tracking in traffic scenes, *Proc. 2nd Europ. Conf. Comput. Vis.*, S. Margherita, Ligure, Italy, May 18–23. G. Sandini (ed.), *Lecture Notes in Computer Science* **588**, Springer-Verlag: Berlin, Heidelberg, New York.

Kollnig, H., 1992. Berechnung von Bewegungsverben und Ermittlung einfacher Abläufe. Diplomarbeit, Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlsruhe (TH), Karlsruhe.

Korn, A.F., 1988. Towards a symbolic representation of intensity changes in images, *IEEE Trans. Patt. Anal. Mach. Intell.*, 10: 610–625.

Lowe, D.G., 1987. Three-dimensional object recognition from single two-dimensional images, *Artificial Intelligence* 31: 355–395.

Lowe, D.G., 1990. Integrated treatment of matching and measurement errors for robust model-based motion tracking, *Proc. 3rd Intern. Conf. Comput. Vis.*, Osaka, pp. 436–440, December 4–7.

Lowe, D.G., 1991. Fitting parameterized three-dimensional models to images. *IEEE Trans. Patt. Anal. Mach. Intell.* 13: 441–450.

Marslin, R.F., Sullivan, G.D., and Baker, K.D., 1991. Kalman filters in constrained model-based tracking, *Proc. Brit. Mach. Vis. Conf.*, Glasgow, UK, pp. 371–374, September 24–26, Springer-Verlag, Berlin, Heidelberg, New York.

Maybank, S., 1990. Filter-based estimates of depth, *Proc. Brit. Mach. Vis. Conf.*, Oxford, pp. 349–354, September 24–27.

Mitschke, M., 1990. *Dynamik der Kraftfahrzeuge: Band C—Fahrverhalten*. Springer-Verlag: Berlin, Heidelberg, New York.

Murray, D.W., Castelow, D.A., and Buxton, B.F., 1989. From image sequences to recognized moving polyhedral objects, *Intern. J. Comput. Vis.* 3: 181–209.

Scales, L.E., 1985. *Introduction to Non-Linear Optimization*. Macmillan: London.

Schick, J., and Dickmanns, E.D., 1991. Simultaneous estimation of 3D shape and motion of objects by computer vision, *Proc. IEEE Workshop on Visual Motion*, Princeton, NJ, pp. 256–261, October 7–9.

Sung, C.-K., 1988. Extraktion von typischen und komplexen Vorgängen aus einer Bildfolge einer Verkehrsszene. In H. Bunke, O. Kübler, and P. Stucki, (eds.), *DAGM-Symposium Mustererkennung 1988*, pp. 90–96, Zürich, Informatik-Fachberichte **180**, Springer-Verlag: Berlin, Heidelberg, New York.

Thompson, D.W., and Mundy, J.L., 1987. Model-based motion analysis—motion from motion. In *Robotics Research*, R. Bolles and B. Roth (eds.), MIT Press: Cambridge, MA, pp. 299–309.

Thórhallson, T., 1991. Untersuchung zur dynamischen Modellanpassung in monokularen Bildfolgen. Diplomarbeit, Fakultät für Elektrotechnik der Universität Karlsruhe (TH), durchgeführt am Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlsruhe (TH), Karlsruhe.

Tsai, R., 1987. A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, *IEEE Trans. Robot. Autom.* 3: 323–344.

Verghese, G., Gale, K.L., and Dyer, C.R., 1990. Real-time, parallel motion tracking of three dimensional objects from spatiotemporal

images. In V. Kumar, P.S. Gopalakrishnan, and L.N. Kanal (eds.), *Parallel Algorithms for Machine Intelligence and Vision*, pp. 340–359, Springer-Verlag: Berlin, Heidelberg, New York.

Worrall, A.D., Marslin, R.F., Sullivan, G.D., and Baker, K.D., 1991. Model-based tracking, *Proc. Brit. Mach. Vis. Conf.*, pp. 310–318, Glasgow, September 24–26, Springer-Verlag: Berlin, Heidelberg, New York.

Wu, J.J., Rink, R.E., Caelli, T.M., and Gourishankar, V.G., 1988. Recovery of the 3-D location and motion of a rigid object through camera image (an extended Kalman filter approach), *Intern. J. Comput. Vis.* 3: 373–394.

Young, G., and Chellappa, R., 1990. 3-D motion estimation using a sequence of noisy stereo images: models, estimation and uniqueness results, *IEEE Trans. Patt. Anal. Mach. Intell.* 12: 735–759.

Zhang, Z., and Faugeras, O.D., 1992. Three-dimensional motion computation and object segmentation in a long sequence of stereo frames, *Intern. J. Comput. Vis.* 7: 211–241.