Int J Comput Vis (2008) 80: 211–225
DOI 10.1007/s11263-007-0112-6

Initialization of Model-Based Vehicle Tracking in Video Sequences of Inner-City Intersections

A. Ottlik · H.-H. Nagel

Received: 4 December 2006 / Accepted: 19 November 2007 / Published online: 29 December 2007

Abstract A *fully automatic initialization* approach for 3D-model-based vehicle tracking has been developed, based on Edge-Element and Optical-Flow association. An entire automatic initialization and tracking *system* incorporating this approach achieves results comparable to those obtained by earlier experiments based on *semi-interactive* initialization, provided the assessment criteria are roughly equivalent. Experiences with a large testing sample—about 15 minutes of inner-city traffic videos—are discussed in detail.

Keywords Tracking · Model-based · Road-vehicle · Kalman-filter · Initialization · Traffic videos · Optical-flow-field segmentation

1 Introduction

In 1987, a conference invitation to discuss over a decade of research on image sequence evaluation provided an opportunity to document experiences and expectations (Nagel 1988). Continuation of this research lead us to shift our emphasis from data-driven to model-based vehicle tracking as documented in Nagel (2004). These are but two alternatives from a larger set of options which comprise 'model-free tracking', 2D-model-based tracking in the image plane, 3D-model-based tracking in the scene domain, and 'hybrid' approaches. The latter exploit specific assumptions about the relation between the agent images to be tracked and the agent pose in the scene space. A hybrid approach assumes, e.g., that a 2D-blob in the image domain represents a vehicle on a 'ground plane' whose parameters are known in the camera coordinate system due to a camera calibration—see, e.g., Magee (2004), Kumar et al. (2004), or Pece and Worrall (2006).

In combination with knowledge about the internal and external parameters of the camera, the use of a *3D vehicle model* allows to determine the vehicle image for *any* relative vehicle pose with respect to the camera. Mismatches thus can be diagnosed more easily as being due either to initialization or tracking errors or to inappropriate parameter estimates for vehicle model or/and camera.

A new system, `Motris`, had been designed, implemented, and shown to achieve at least the *tracking rate* obtained earlier with `Xtrack`, as reported in Dahlkamp et al. (2007). Usually video-based vehicle tracking assumes that a vehicle will continue its steady-state motion. The current estimates for the parameters representing the vehicle state—regardless whether this state refers to an image-plane coordinate system or to a 3D-world coordinate system—are combined with a geometric motion-model in order to predict the vehicle state for the next observation time. In most cases this motion-model assumes a straight-line or stationary circular movement at constant speed. Extensive experiments have been performed to evaluate different methods to exploit these assumptions for short-term prediction of 3D-model-based vehicle tracking. The scoring functions exploited for these experiments depend non-linearly on their arguments, implying a dependency on the initialization of the tracking step proper. In order to disentangle the initialization effects from variations in the formulation and parameterization of the tracking step proper, the initial conditions for each vehicle had been determined interactively beforehand and have been re-used for different tracking methods and associated parameter variations. In these experiments, the initialization

A. Ottlik · H.-H. Nagel (✉)
Institut für Algorithmen und Kognitive Systeme, Universität Karlsruhe (TH), Postfach 6980, 76128 Karlsruhe, Germany
e-mail: nagel@iaks.uni-karlsruhe.de

information comprised the vehicle model to be used, the frame-number where tracking should start and terminate as well as whether the vehicle had been recorded in bright sunshine (facilitating the exploitation of shadow-casting to robustify vehicle tracking) or in diffuse illumination without extended shadow-casting. As a result of these experiments, tracking proper does no longer appear to be the most important bottleneck. It thus appeared natural to use `Motris` for a systematic investigation of a *fully automatic initialization* whose treatment constitutes the remainder of this contribution.

Following a terse discussion of related publications in Sect. 2, Sect. 3 describes the component processes required in order to *automatically* initialize 3D-model-based tracking. This automatic initialization approach has been extended in Sect. 4 to provide a cue for the detection of potential tracking irregularities. The capabilities of the enlarged system have been tested in a series of experiments which evaluate the same video sequences already studied in Haag and Nagel (1999), although now *without interactive help* by the experimenter during the initialization phase. Experimental results obtained by such fully automatic initialization and tracking are presented in Sect. 5, with an assessment following in Sect. 6.

2 Discussion of Related Publications

Although road vehicle tracking in traffic videos has considerable appeal for various applications, see Kastrinaki et al. (2003) for this application domain in general or Sun et al. (2006) for on-board vehicle detection and tracking, this manuscript focuses on *basic research* approaches to *3D-model-based* tracking of vehicles recorded by a *single stationary* video camera. This specification excludes coverage of multi-ocular tracking, for example using a *moving stereo-camera pair* as in Leibe et al. (2006), as well as data-driven approaches, in particular ones operating only in the image plane. Regarding the latter the reader is referred to the special issue Collins et al. (2000), to Elgammal et al. (2002), or to Nadimi and Bhanu (2004) who pay special attention to the separation between moving vehicles or people and the co-moving shadows cast by their bodies.

A recent survey of object motion and behaviors by Hu et al. (2004) comprises surprisingly few references to 3D-model-based vehicle tracking since Haag and Nagel (1999). Our own search for relevant archival publications—see Dahlkamp et al. (2007)—confirms this finding and here extends its validity even up to the immediate past. Our subsequent discussion will be restricted to publications treating questions which are directly associated with our subject. We attempt, however, to at least mention some important publications treating system aspects beyond our current focus. In a notable exception from this scarcity of publications concerned with 3D-model-based tracking in monocular video sequences, Pece and Worrall (2002), see too Pece and Worrall (2006), postulate a probability distribution of Edge Element (EE) locations around model segments which have been projected into the image plane according to the current vehicle state estimate. The state estimate is updated by maximizing the postulated likelihood for observing grayvalue transitions. A recently published variant studies a marginalized likelihood model which takes into account possible deviations of vehicle shapes from the modeled one, see Pece (2006).

Although this latter approach has been studied for tracking, it could be considered, too, to offer a methodological alternative for *initialization* in comparison with an earlier, search-based approach reported in Tan et al. (1998). This would require, however, at least a rough idea where to look for a vehicle image. Such a cue usually is obtained by background subtraction—see, e.g., Stauffer and Grimson (2000) or Elgammal et al. (2002)—and combined with the 'ground-plane assumption' in what has been denoted as a hybrid approach above in order to obtain an initial estimate for a vehicle's location in the scene. This latter approach has been used, e.g., in the experiments reported by Pece and Worrall (2006).

Kumar et al. (2004), too, start from foreground regions obtained by background subtraction based on a 'home-grown' approach for background estimation and maintenance. Attributed graphs are used to group the resulting foreground regions for initialization of image-plane tracking. A Kalman-Filter exploits hypotheses about 2D position, movement, and shape continuity. Based on tracking results, 2D-blob trajectories in turn are associated with hypothesized 3D category models moving on the ground plane of the scene, i.e. realizing a *hybrid* approach differing from the one used by Pece and Worrall (2006) regarding the '2D-blob substrate' for the transition from 2D to 3D. Dynamic programming techniques facilitate to maintain the characteristics of a category instance even in cases where the 2D-blob substrate may change substantially due to merging and splitting caused by, e.g., occlusions or artefacts of the blob generation steps.

In contradistinction to the exploitation of background subtraction as a means to generate pose cues for moving vehicles, the estimation, segmentation, and analysis of Optical-Flow (OF) fields can provide the required *rough* initial estimate of vehicle pose in a manner less sensitive to illumination changes because it is based on a *motion*-cue and not on a less specific *change*-cue. A useful recent discussion of related questions can be found in Renno et al. (2006).

3 Component Processes for Automatic Initialization

The entire automatic initialization process consists of a number of component processes which realize the following series of steps, see Ottlik (2005):

1. Optical-Flow-field estimation;
2. Optical-Flow-field segmentation – resulting in so-called 'Object Image Candidates (OICs)';
3. 'Confidence accumulation' by short-time image-plane-tracking of OF-field segments, i.e. OICs;
4. A consistently trackable OIC provides velocity and orientation information which is exploited to reduce the search space for a Hough-Transform-determination of vehicle *type* and *location* in the 3D-scene.

It is the task of these processes to provide *initial* estimates for the five components of a vehicle state-vector **x** which comprises the x- and y-location of the vehicle reference point on the road plane, the vehicle orientation, its speed, and the steering angle for the front-wheels (i.e. the angle between the longitudinal axis of the vehicle and the intersection of the wheel plane with the road plane). The initialization value for the steering angle is always assumed to be zero degrees. As a consequence, initial values have to be estimated only for the *first four* components of the state vector.

3.1 OF-Field Estimation Based on the Gray-level Structure Tensor

Good OF-vectors can be estimated using the Gray-level Structure Tensor (GST), provided the masks for the derivative operators and the integration area are chosen suitably, see Middendorf and Nagel (2001); Middendorf (2004). In analogy to the pseudo-inverse-based OF-estimation approach used in Haag and Nagel (1999), entries into the operator masks used for the estimation of spatiotemporal grayvalue derivatives take into account the line-structure of the interlaced video recording such that the required derivatives can be determined at each half-frame (i.e. 'field') time in *full-frame spatial resolution*. Whenever 'frame' is mentioned in the sequel without further qualification, it corresponds to a time-interval of 20 ms!

In principle, this approach has the additional advantage that an analysis of the GST-eigenvalue structure allows to determine whether the local grayvalue distribution provides enough information to reliably estimate an OF-vector. It turned out, however, that this information does not need to be evaluated *explicitly* because less reliably estimated OF-vectors usually were excluded already during the OF-field segmentation phase from being incorporated into OICs.

3.2 OF-Field Segmentation

The OF-field is estimated and segmented for each frame (i.e. time-point) of the image sequence. In order to avoid multiple initializations for the *same* vehicle, all OF-vectors within an already accepted OIC are discarded from consideration during a continuation of the segmentation process. Furthermore, only OF-vectors exceeding a minimum norm are taken into account. In the experiments reported here, this threshold has been set to a minimum of 0.2 pixel per frame, which is equivalent to a speed of about 4 km/h for a body moving in the depicted scene.

An OF-vector within a segment is compared to its neighboring vectors. In case a neighboring vector is compatible with respect to norm and orientation, it is assigned to this same segment. This compatibility check is carried out using separate thresholds for the maximum difference regarding the norm and the orientation. Figure 1(left panel) illustrates

Fig. 1 (*Left panel*) OF-segmentation at frame 511 in image sequence `stau02`. For illustration purposes, OF-vectors have been suppressed if their norm was smaller than a threshold. This leaves the stationary background—including momentarily stationary vehicles—uncovered by OF-vectors. Note that in some cases OF-segments extend beyond a vehicle image while in other cases they do not cover the vehicle image completely. (*Right panel*) Vehicle models used for vehicle *type selection*: (1) Hatchback, (2) Van, (3) Transporter, and (4) Tramway

the segmentation result for frame 511 in image sequence `stau02`.

3.3 Spatiotemporal Analysis of OF-Field Segments

In order to select OF-segments with a high plausibility to correspond to an OIC, the overlap between OF-segments obtained for adjacent frames is analysed. In case the overlap is large enough (a minimum overlap of 90% was required for the experiments reported in the sequel), it is assumed that these OF-segments belong to the same vehicle. Due to the parameter settings used for OF-computation, only grayvalue structure displacements up to an image-plane shift-velocity of 3.5 pixel per frame could be determined reliably. Thus even in the case of small (i.e. 40 pixel long) and fast moving vehicles, an overlap of $\frac{40-3.5}{40} \approx 91\%$ should be observable.

In order to determine the time when a vehicle has become *fully* visible, it is assumed

- that an OF-segment can be tracked for several consecutive frames (for a minimum of 10 frames in the experiments reported in the sequel) and
- that its size stays *almost* constant throughout this phase. This assumption has been concretized to the requirement that the size of an OF-field segment can have changed by at most 10% during this period.

The OF-field-segmentation and OF-segment-tracking processes have to identify those vehicles which have become *completely visible* within the Field-of-View (FoV). Problems resulting from more difficult situations—such as, e.g., splitting or merging of OF-segments—are not taken into account because it is assumed that these problems are often due to occlusions by stationary objects or other moving vehicles. Moreover, the segment contour does not need to be determined very precisely because edge information will be exploited during a subsequent vehicle position estimation step. The OF-field-segmentation and -tracking algorithms thus could be kept rather simple.

The clusters of OF-vectors obtained according to the preceding steps provide already a good initial estimate for *velocity* and *orientation* of a vehicle, provided vehicles are oriented essentially parallel to their velocity vector in the 3D-scene (see Fig. 2). It thus remains to obtain an initial estimate for the *vehicle location* on the road plane.

3.4 Vehicle Localization

OF-field segmentation itself often does not result in a precise enough vehicle localization. Whenever two vehicle images are close to or even overlap each other and the vehicles drive at almost the same speed, a segmentation based solely on OF-vectors will lead to a single segment. Localization, therefore, exploits edge information and vehicle models in addition to OF-field information.

The initial orientation of the object candidate is estimated by projecting the mean OF-vector of the OF-segment, positioned at its centroid, into the 3D-scene. Using a 3D-vehicle-model and the initial orientation obtained by the preceding step, this image of tentatively visible model edge segments—i.e. of a set of *expected* EEs—is tessellated.

In a next step, orientations of the expected EEs are determined by calculating the derivative of the tessellated model-segment image using the *same* derivative algorithm as for EE extraction from the image sequence. This step has been introduced, because gradient directions close to a corner do not correspond to the visible model segment normals, due to the convolution used for the determination of derivatives.

In a further step, a Hough-Transformation process finds a centroid of this tessellated image maximizing the number of EEs which are compatible with the 'synthetic EE-image'. The Hough space represents the possible pixel positions of the tessellated image's centroid. In case an EE's orientation is close to one of the expected EE's orientation, a fixed score is added to the bin which corresponds to the expected EE's centroid. In order to accommodate discretization errors and to take into account that the 3D-vehicle-models provided for this search phase do not perfectly correspond in general to the observed vehicles, each EE votes for a 3×3-region in the Hough space. Figure 3 illustrates the described algorithm.

3.5 Selection of a Vehicle Model

Several vehicle models (see Fig. 1, right panel) are provided to the system and are checked for compatibility with EEs extracted from the current frame within the OF-segment in a predefined order, namely from the largest (tramway) to the smallest (hatchback). Compatibility is checked using thresholds, one for the minimum score in the Hough space and a second one for the overlap between OF-segment and object candidate. The first (i.e. largest) vehicle model passing these compatibility tests will be chosen for the initialization of a tracking process.

4 Optical-Flow-Vector Coverage Rate (OFCR) as Irregularity Cue

A state-vector $\mathbf{x}$ can be assessed by determining how many compatible features can be found in the image sequence. In case of motion estimation, OF-vectors $\mathbf{u}$ extracted at the position of the object candidate projection are checked for compatibility with the expected displacement rate field. A displacement rate vector $\mathbf{v}(\boldsymbol{\xi}, \mathbf{x})$ at pixel position $\boldsymbol{\xi}$ is computed based on the initialized 3D-vehicle-model and the current state-vector $\mathbf{x}$. It is assumed that the difference $\Delta\mathbf{u} = (\Delta u_x, \Delta u_y)^T = \mathbf{v} - \mathbf{u}$ between the displacement rate vector and the OF-vector $\mathbf{u}(\boldsymbol{\xi})$ at pixel position $\boldsymbol{\xi}$ is normally distributed with expectation $\mathbf{0}$ and covariance Σ. Thus the

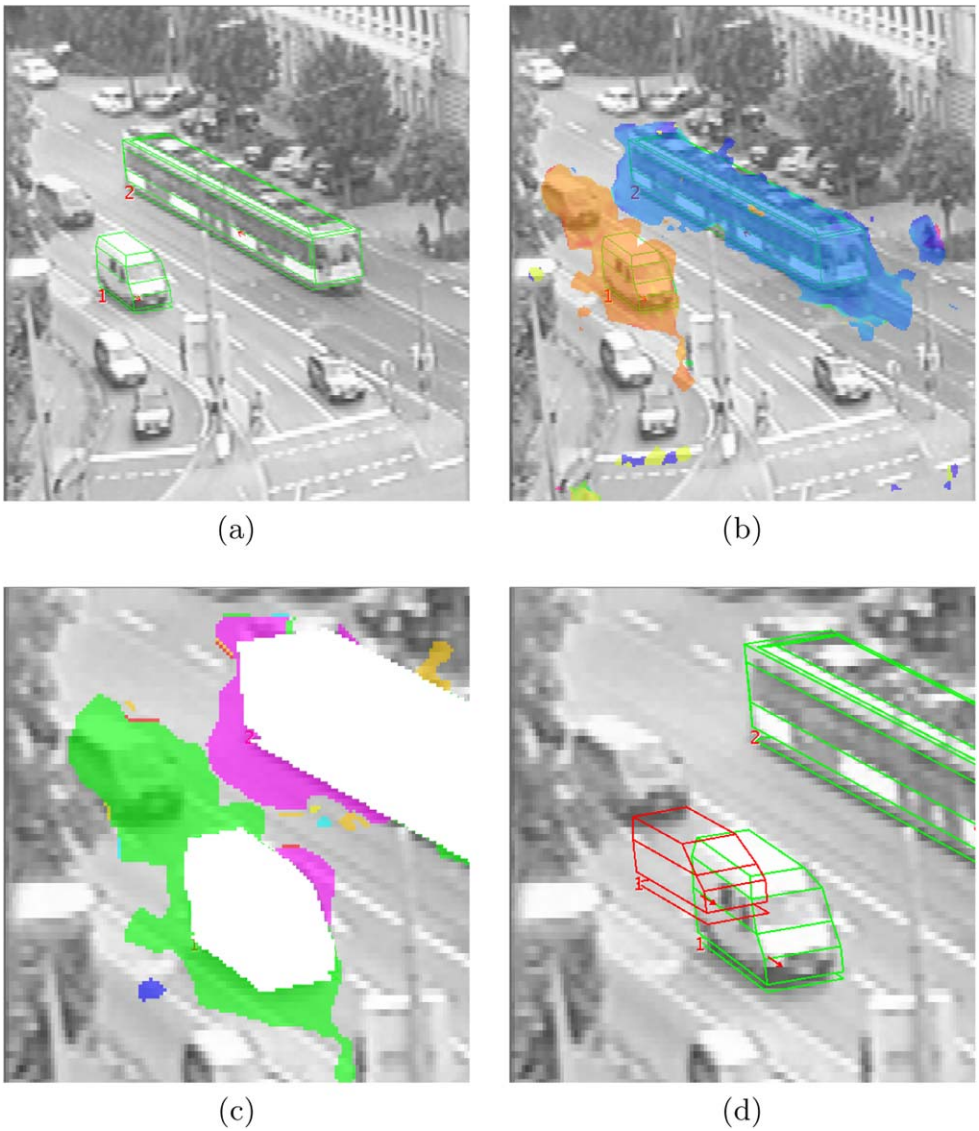

Fig. 2 Detection phase: **a** Section enlargement of frame 99 in image sequence `stau02` with tracking results overlaid. **b** OF-vectors exceeding the minimum displacement norm. The orientation is color encoded. **c** OF-tracking results. White pixels mark all OF-vectors discarded from segmentation and tracking. The green segment exceeds the image of the van since both cars covered by this OF-field segment drive at almost the same speed and thus their OF-vectors coalesce into a single segment. **d** State estimation obtained from green OF-segment in panel (**c**). As can be seen, orientation can be estimated appropriately in contrast to the position which has to be refined in further steps

induced Mahalanobis distance M_{OF} is

$$M_{OF} = \Delta\mathbf{u}^T \cdot \Sigma^{-1} \cdot \Delta\mathbf{u}. \tag{1}$$

The covariance Σ can be computed by taking into account measurement uncertainty Σ_{OF} and the uncertainty P^+ of the updated state estimate $\hat{\mathbf{x}}^+$ (see Ottlik 2005):

$$\begin{aligned}\Sigma &= \frac{\partial\Delta\mathbf{u}}{\partial\mathbf{u}}(\mathbf{u},\hat{\mathbf{x}}^+)\cdot\Sigma_{OF}\cdot\left(\frac{\partial\Delta\mathbf{u}}{\partial\mathbf{u}}(\mathbf{u},\hat{\mathbf{x}}^+)\right)^T \\ &\quad+\frac{\partial\Delta\mathbf{u}}{\partial\mathbf{x}}(\mathbf{u},\hat{\mathbf{x}}^+)\cdot P^+\cdot\left(\frac{\partial\Delta\mathbf{u}}{\partial\mathbf{x}}(\mathbf{u},\hat{\mathbf{x}}^+)\right)^T, \\ &= \Sigma_{OF}+\frac{\partial\Delta\mathbf{u}}{\partial\mathbf{x}}(\mathbf{u},\hat{\mathbf{x}}^+)\cdot P^+\cdot\left(\frac{\partial\Delta\mathbf{u}}{\partial\mathbf{x}}(\mathbf{u},\hat{\mathbf{x}}^+)\right)^T. \end{aligned}\tag{2}$$

In case the Mahalanobis distance exceeds a threshold with respect to the $(1-\alpha)$ quantile of a $\chi^2(2)$ distribution (two degrees of freedom due to the two independent components of $\Delta\mathbf{u}$), the displacement rate vector and the OF-vector are assumed to be incompatible. The ratio r_t between the number of compatible OF-vectors and the number of expected vectors—the so-called 'Optical-Flow-vector Coverage Rate (OFCR)', see Fig. 5—can be used to assess the state estimate.

Of particular interest is a closer analysis of the OFCR temporal development shown in the left panel of Fig. 5: when the vehicle enters into the occlusion situation, the OF-vectors determined for the pixel positions on the vehicle image increasingly correspond to the ones estimated from the—essentially stationary—tree foliage with the result that these estimates are close to zero and thus are incompatible with the expected displacement rate on the vehicle surface.

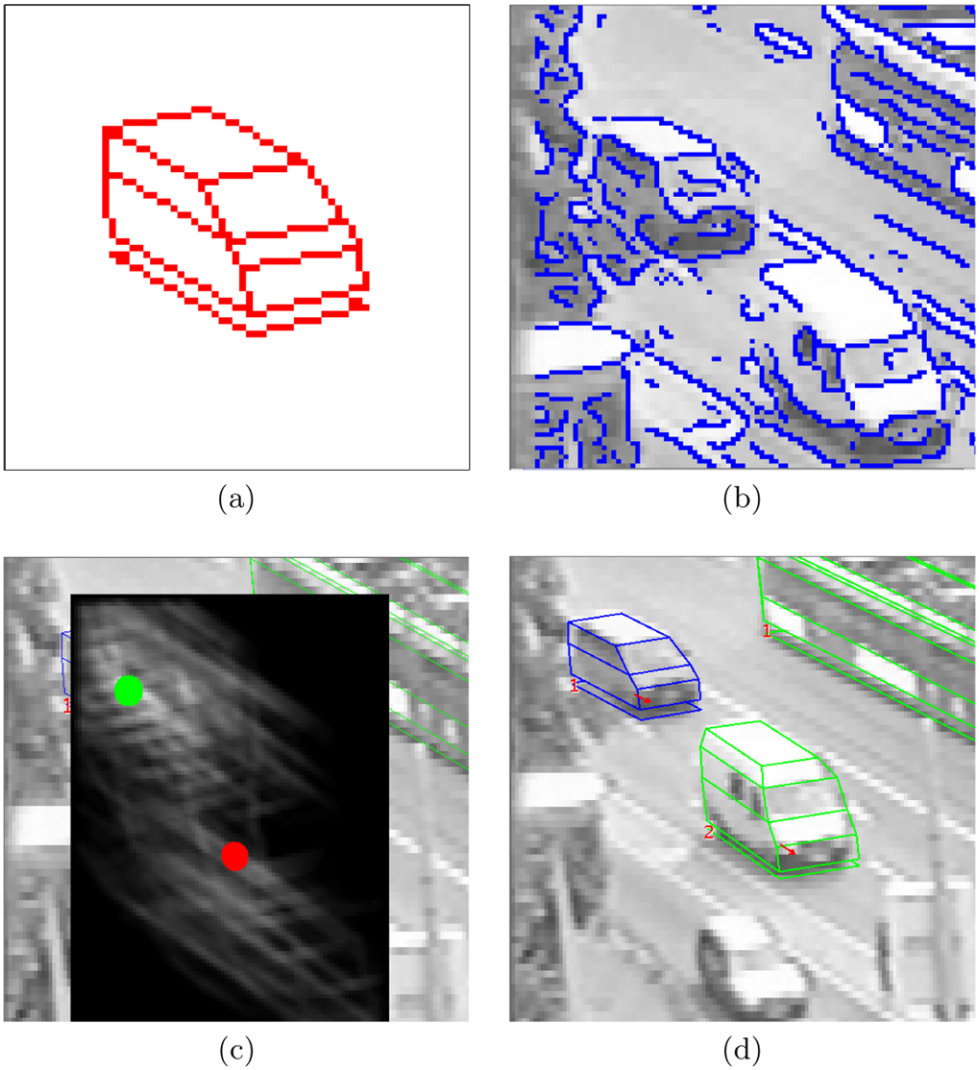

Fig. 3 Localization using a Hough-Transformation: **a** Vehicle model image generated using the state estimation derived from the corresponding OF-segment. **b** Edge elements (EEs) extracted from the current video frame. **c** Accumulator-Array as a result of the Hough-Transformation, with a pixel having been painted the brighter the higher the score in this bin; *red dot*: OF-segment's centroid, see Fig. 2; *green dot*: position of maximum. **d** New state estimation using the localization obtained from the Hough-Transformation

The number of compatible OF-estimates thus decreases and, thereby, the stabilizing effect which the OF-estimates usually exert on model-based tracking. Tracking begins to fail with the effect that the estimated model speed drops to zero, too. This in turn, however, results in a (fake) compatibility between the incorrectly estimated model state and the (near-)zero OF-estimates which can be seen, e.g., in a sudden increase of the OFCR around frame number 1300 in Fig. 4.

This assessment can be used to identify situations where tracking has *failed* or where a vehicle has been *occluded* almost completely. In such situations, tracking of the corresponding vehicle is discontinued. In order not to stop tracking when a vehicle is occluded for only a short subsequence of frames—by, e.g., a mast—the *assessment figure-of-merit* a_t is accumulated over time using a decay factor γ:

$$a_t = (1-\gamma)\cdot a_{t-1} + \gamma \cdot r_t. \tag{3}$$

Figure 5 illustrates the assessment for several vehicles which are temporarily occluded by a tree.

5 Experiments With the `stauXX` Video Sequences

The incorporation of an automatic initialization process into our 3D-model-based vehicle tracking system `Motris` should make its use easier by obviating the necessity to provide (partial) initialization information for each vehicle.

An initial experimental evaluation of the *fully automatic initialization* described in Sect. 3 has been presented in Ottlik (2005)—albeit based on a set of only three video sequences. It thus appeared desirable to investigate the hypothesis that this new automatic initialization yields tracking results which are comparable in quality to those reported earlier in Haag and Nagel (1999). These earlier results had been obtained with the `Xtrack`-system and a *semi*-automatic initialization procedure on a relatively large sample size of 394 vehicles.

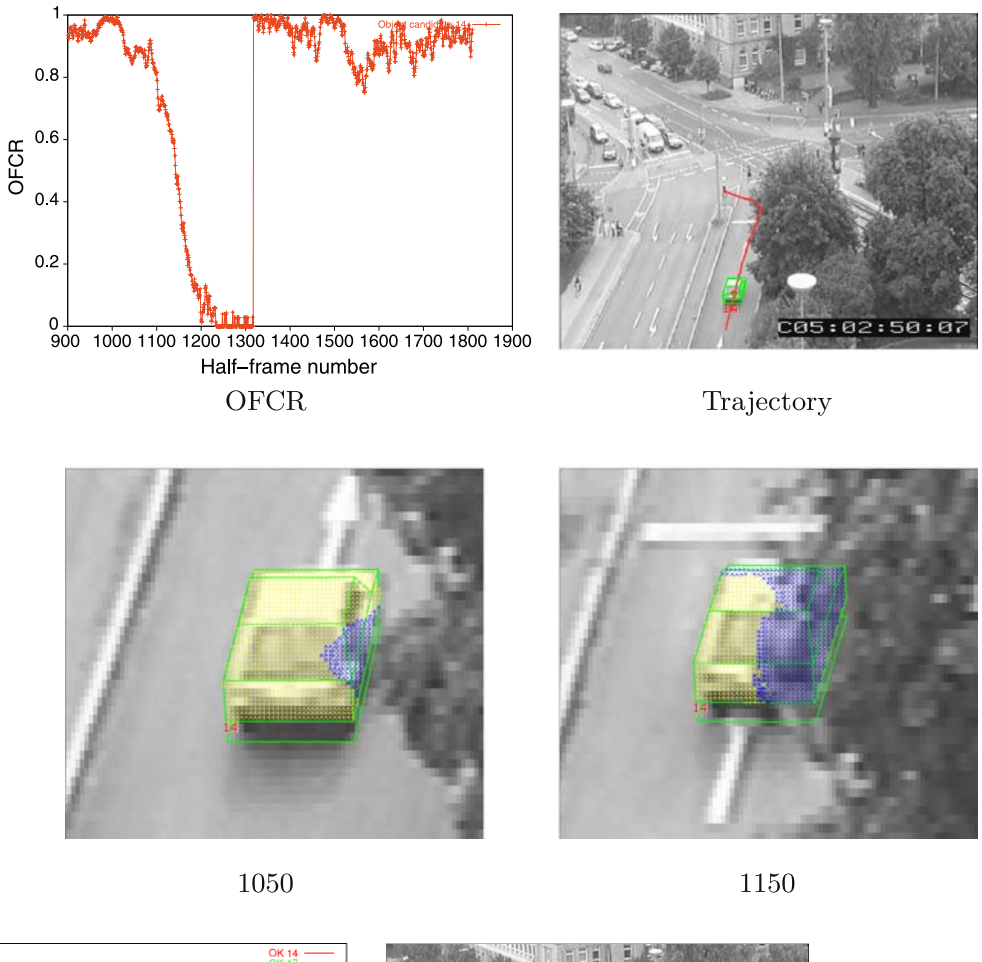

Fig. 4 Optical-Flow-vector Coverage Rate (OFCR) (*top left panel*) illustrated for vehicle # 14 in image sequence `stau02`. This vehicle becomes partially occluded by a tree as shown in the *top right panel*. The *bottom panels* illustrate OF-vectors which are compatible (colored in *yellow*) or incompatible (colored in *blue*) with the displacement vectors determined for the corresponding pixels on the basis of the current state vector estimate. In the *bottom left panel*, only a smaller fraction of this vehicle is occluded by a tree whereas the *bottom right panel* illustrates OFCR for a somewhat later point in time when a larger fraction of this vehicle has been occluded by a different, more extending branch of the tree. The estimated vehicle trajectory shown in the *top right panel* indicates that this occlusion resulted in a tracking failure

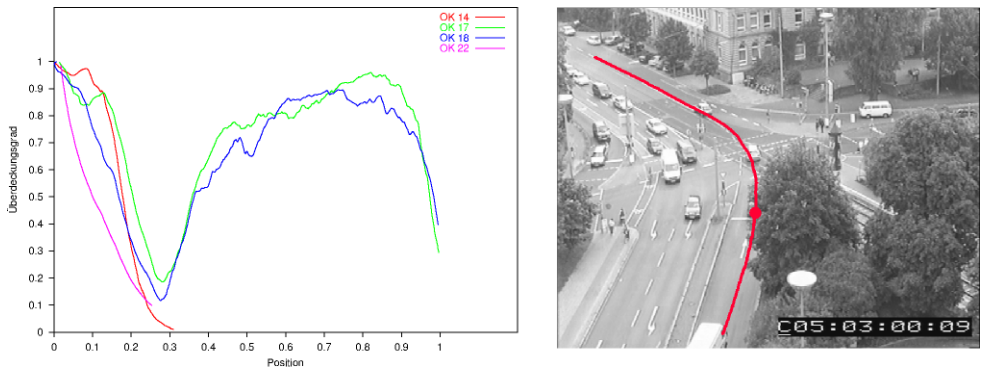

Fig. 5 The *left panel* shows the 'Optical Flow Coverage Rate (OFCR)', i.e. the ratio between the number of OF-vectors (which are considered to be compatible with their corresponding displacement vector for vehicles driving along the lane indicated in the *right panel*) and the expected total number of displacement vectors within the vehicle image derived from the current vehicle state vector estimate. The abscissa denotes the position of a vehicle on the lane (see *red line* in *right panel*). The ratio has been plotted for four different object candidates, each one plotted in a different color. The *red blob* in the *right panel* indicates the image area where vehicles are occluded by a tree. As soon as vehicles are occluded, the plotted ratio decreases significantly. Two vehicles can be successfully tracked after re-emerging from this occlusion situation: this fact is reflected by the increase of the plotted ratio after the corresponding vehicles have passed the occluding tree

It is not possible, however, to ascribe any difference between the results obtained by the `Xtrack`-system and here *solely* as being due to the effects of an automatic initialization, because `Motris` differs in several respects from the `Xtrack`-system.

5.1 Boundary Conditions for a Comparison

During the design and implementation of `Motris`, a multitude of local modifications and improvements have been realized in order to create a system which is easier to un-

frame 757 frame 850

Fig. 6 Example for tracking assessed as 'very good' (++): Object Candidate (OC) 17 in image sequence `stau12`

derstand, to use, and to maintain than the latest version of `Xtrack`. In addition, the redesign aimed at creating a system which also operates more robustly than `Xtrack` to the extent possible without introducing changes of a more fundamental nature. The combined *effects* of these local algorithmic changes have been studied extensively with the conclusion that the `Motris`-system *without* automatic initialization yields results at least as good as those obtained earlier using the `Xtrack`-system, provided the same initialization steps are used in both systems, see Dahlkamp et al. (2007).

For the record, it is mentioned that we used `Motris` here with the 'steering angle' option whereas the results reported by Haag and Nagel (1999) for the `Xtrack`-system had been obtained with the older 'angular velocity' option. It is possible to replace the state vector component 'angular velocity' by the 'steering angle'. The vehicle orientation then remains essentially constant even for very small or zero speeds whereas a non-zero angular velocity value caused by noise effects may result in orientation changes for a (near) zero speed value. The principal advantage of the 'steering angle' option consists in the fact that we no longer need a separate threshold in order to suppress erroneous orientation updates for very small speed values. Previous extensive experiments have shown that the results obtained with the steering angle option are at least as good as those obtained using the angular velocity. We thus posit that this methodological difference with respect to `Xtrack` does not create any significant bias for the conclusions to be drawn from the experiments reported in the sequel.

5.2 Evaluation of the Automatic Initialization Process

All experiments have been carried out using the same set of parameters.

5.2.1 *Assessment Procedure*

As in Haag and Nagel (1999), tracking results are assessed as 'very good' (denoted as ++), 'good' (+), 'tolerable' (o), 'bad' (–), or as a tracking 'failure' (- -). Twelve image sequences (`stau01` – `stau12`) have been evaluated comprising about 44 000 frames (i.e. about 15 minutes of video). Altogether 394 vehicles are visible in these sequences. Unfortunately, 33 vehicles do not correspond to any vehicle type for which models have been provided currently to the system. These 'non-standard types' comprise, e.g., trucks with trailers or busses. Tracking such vehicles with the 'closest' model among those currently admitted usually fails because these vehicles are larger than the provided models which often leads to multiple initializations. Such 'non-standard' vehicles have not been taken into account for the assessment. We are thus left with a Reference Set (RS) comprising $394 - 33 = 361$ vehicles admitted for assessment.

Regarding a comparison with results reported in Haag and Nagel (1999), an additional aspect has to be taken into account for an overall assessment, namely the delay between the time when a vehicle has become fully visible in the FoV of the recording camera and the time point of actual initialization. In case a vehicle

1. has been initialized within 50 frames (i.e. 1 second) after it has become fully visible and
2. the most appropriate vehicle type has been chosen among those admitted and
3. the object candidate covers the visible vehicle image very well,

tracking is assessed as *very good* (see Fig. 6).

Smaller discrepancies from the (interactively determined) real state for the entire period of visibility or even intermittent larger discrepancies, which subsequently have been corrected automatically, are denoted as *good* tracking (see Fig. 7).

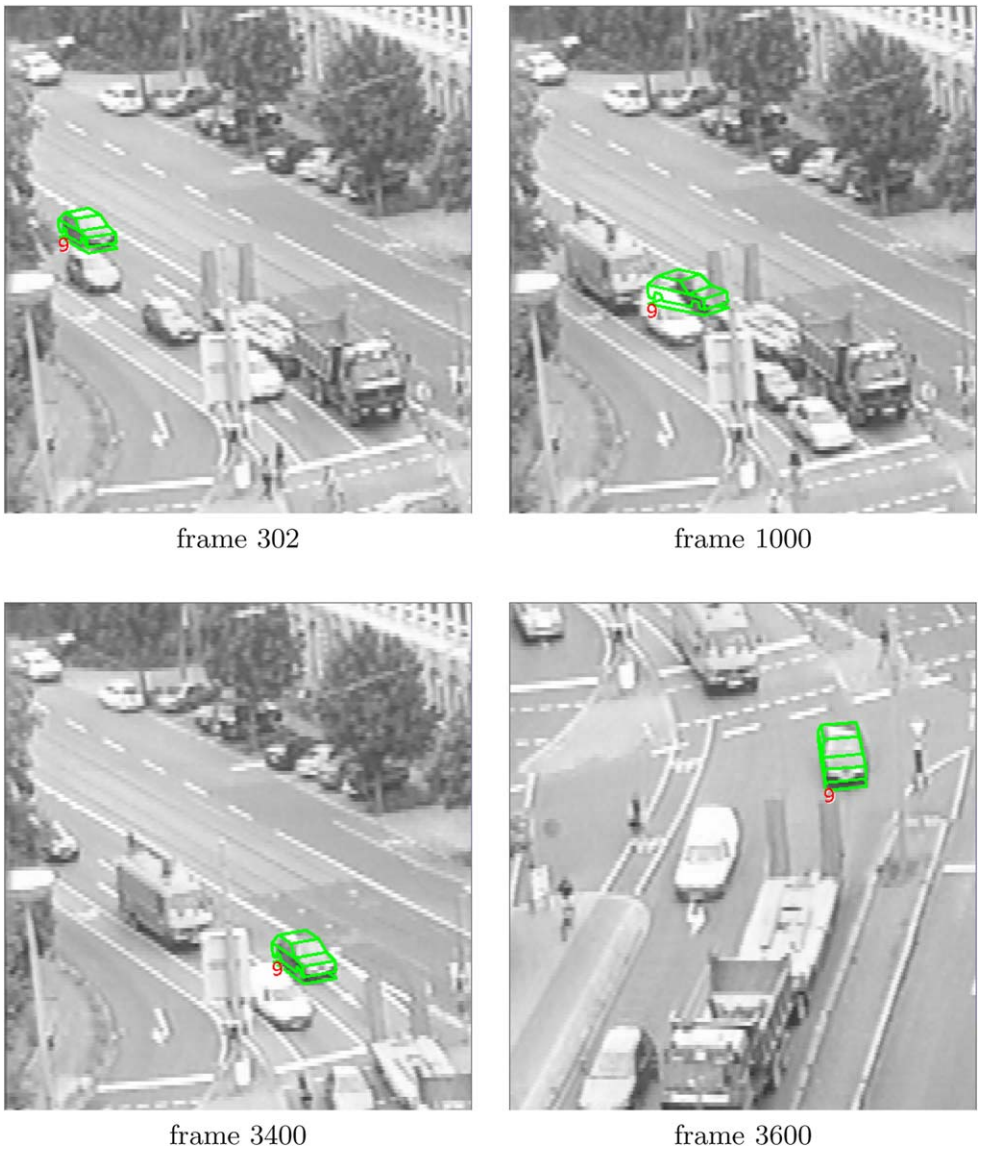

Fig. 7 Example for a tracking result which has been assessed as 'good' (+): Object Candidate (OC) 9 in image sequence `stau03`. Initialization and tracking is good until the car halts at the traffic light. During the halt the model is distracted by other vehicles. After the vehicle restarts, tracking can be improved again

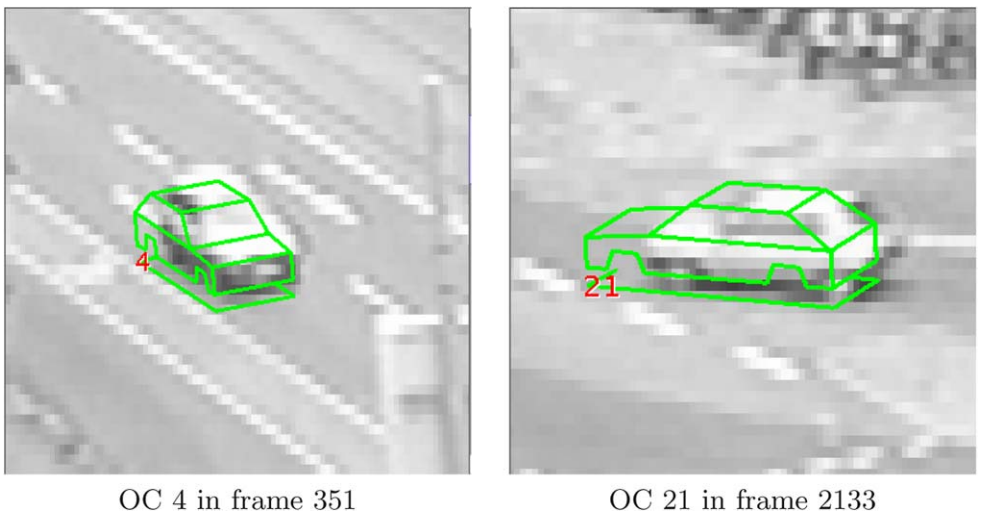

Fig. 8 Examples for tracking assessed as 'tolerable' (o): Object Candidates (OCs) 4 and 21 in image sequence `stau05`

In those cases where the object candidate does not fit precisely but still covers at least half of the vehicle image, tracking results are assessed as *tolerable* (see Fig. 8).

When (i) initialization takes place later than 50 frames after the vehicle has become fully visible, if (ii) the wrong vehicle type has been chosen, or if (iii) the object candidate does not cover most of the vehicle image, the result is considered as *bad* (see Fig. 9).

In case (a) no object candidate at all has been initialized for a visible vehicle, if (b) several candidates have been initialized for the same vehicle, or if (c) tracking fails and this failure has not been detected automatically, tracking is assessed as a *failure* (see Fig. 10).

If a vehicle becomes temporally occluded—by, e.g., a tree—tracking is discontinued when the vehicle is no longer visible and a new object candidate is initialized after the vehicle has 'reappeared'. Thus two object candidates have to be initialized for one vehicle. In case tracking is assessed differently for these two object candidates, the overall assessment will be chosen as the worse assessment among these two (see Fig. 11).

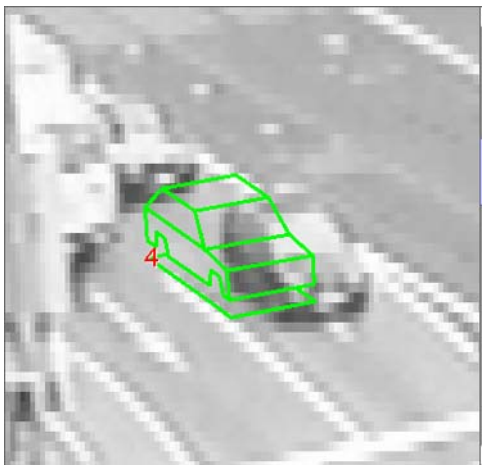

Fig. 9 Example for tracking assessed as 'bad' (-): Object Candidate (OC) 4 in image sequence `stau12`

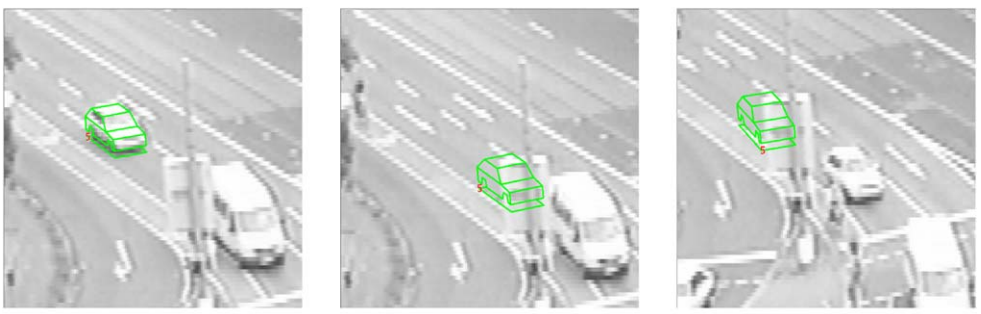

Fig. 10 Example for a tracking 'failure' (--): Object Candidate (OC) 5 in image sequence `stau07` at frame 369 (*left panel*), 500 (*center*), and 3410 (*right*)

5.2.2 Results

About 15% of all vehicles in the Reference Set (RS) can be tracked as 'very good' and 23% as still 'good'. Assignment to the categories 'tolerable', 'bad' and 'failure' is more subtle and will be discussed below in more detail.

Initialization Only 3 vehicles have not been initialized at all, i.e. less than 1% out of 361. In 16 cases (i.e. 4%), multiple initializations occurred—mainly due to a wrong choice of the vehicle type or due to bad tracking results. A bad *initial localization* occurred for 73 vehicles (i.e. about 20%) due to background structure. 48 vehicles have been initialized lately. In 22 cases the only reason for a 'bad' assessment is that these vehicles were partially occluded by masts shortly after they had become visible (see Fig. 12).

In 7 cases, a *re*-initialization has taken place *lately*, otherwise tracking of those vehicles would have had to be assessed better, see Fig. 13.

Model Selection In about 16% the wrong vehicle type has been selected. Type selection is a difficult task because the vehicle models supplied to the system rarely match precisely the vehicles seen in the videos. It thus can happen, e.g., that in the case of a station wagon observed in the image sequence the hatchback model used does not match very well but the van model might match better. In 38 cases the wrong type selection was the only failure for a vehicle. These vehicles have been assessed as satisfactory in the variant 'fair assessment'.

Automatic Detection of Tracking Failures There have been 19 cases overall where tracking failures have not been detected automatically. In 9 of these cases, an occlusion situation by, e.g., a tree or a traffic sign has not been detected. Unfortunately, the tracking of vehicles which were only *partially* occluded by a *thin* mast has been discontinued, too, in 8 cases. For two additional vehicles, tracking has been discontinued automatically although they were fully visible (false alarm). Each of these two cases has been treated as a failure.

Fig. 11 Error of type selection in second initialization leads to an overall error: Object Candidate (OC) 1 (re-initialized as OC 25, *top right panel*) in `stau02` and OC 18 (re-initialized as OC 42, *bottom right panel*) in `stau12`

6 Assessment of Results Obtained with the `stauXX` Sequences

The investigations reported here attempt to answer the question to which extent the automatic initialization described in Sect. 3 allows to replace the *semi*-interactive initialization process realized for the `Xtrack`-system used by Haag and Nagel (1999). The two experiments compared for this purpose differ in assessment details, in particular regarding the time-point—i.e. the frame number—when a vehicle is considered to have been detected acceptably. `Xtrack`-results can not be judged on such a criterion because there the initialization time-point had been determined interactively for each vehicle. A similar argument applies to the automatic choice of vehicle model which again had been determined interactively for the `Xtrack`-experiments.

In addition to the newly relevant initialization time, another aspect has to be taken into account. The Hough-transform for the localization of a vehicle model searches for the best match between the expected edge map (obtained from a tentative projection of the vehicle model into the image plane with suppression of occluded model segments) and the EE-map obtained from the image frame (see Sect. 3.4). This search procedure delivers a suitable location estimate for each vehicle even in cases where two vehicles enter the FoV (almost) simultaneously with roughly comparable velocities on two neighboring lanes. In contradistinction to this version with automatic initialization, `Xtrack` assumed that the center of a suitably selected OF-blob, back-projected into the road-plane, would provide an initial guess for the location of one vehicle. In this latter case, it is practically impossible to split the resulting OF-vector cluster unless a-priori knowledge about the lane structure is used, based on the *assumption* that an OF-cluster extending across two neighboring lanes is due to two neighboring vehicles. This assumption had been exploited by `Xtrack` in order to split such an OF-blob along the intermediate lane boundary.

Such an assumption is no longer necessary if knowledge about the vehicle model enters already into the localiza-

tion phase—which had become justifiable due to the improved OF-estimates and the improved OF-clusters obtained therefrom. As a consequence, the automatic initialization *no longer requires a lane-model*.

Fig. 12 Example of late initialization due to masts. *Top panel*: vehicle initialized correctly. *Bottom panel*: vehicle initialized lately due to partial occlusion by masts

In view of these complications for a juxtaposition of the experiments reported in Haag and Nagel (1999) and here, three different overall comparisons for the `stauXX` set of image sequences are presented:

'*Strict*' assessment: Out of the Reference Set of 361 admitted visible vehicles, 65% on average have been initialized automatically and tracked satisfactorily, see Table 1 (entries in column $1+i, i=1..12$, correspond to results obtained for image sequence `stau{0}i`).

'*Tolerant*' assessment: Because only four vehicle types have been supplied to the system, these usually do not fit the observed vehicles very well. Vehicle type selection is a difficult task under such conditions and, therefore, tracking is assessed as satisfactory even though a vehicle type has been selected which is not fully appropriate. Under these conditions, 76% of the Reference Set could be tracked at least satisfactorily, see Table 2.

'*Fair*' assessment: In addition to the cases accepted in the assessment characterized as 'tolerant' in the preceding item, late initializations—for example at the pedestrian crossing in the upper right quadrant of Fig. 12 (with 'thin

Fig. 13 Example of a late re-initialization (*right panel*) after an occlusion

Table 1 *Strict* assessment (see text for detailed explanation)

Assessment	1	2	3	4	5	6	7	8	9	10	11	12	∑
++	3	3	6	5	5	5	4	6	4	4	4	4	53
+	4	7	11	2	6	9	3	12	3	6	10	12	85
o	10	5	4	3	10	16	3	15	5	7	17	3	98
–	11	9	6	7	3	6	2	6	3	10	14	12	89
- -	4	1	2	4	1	1	2	3	3	8	6	1	36
++/+/o	17	15	21	10	21	30	10	33	12	17	31	19	236 (65%)
∑	32	25	29	21	25	37	14	42	18	35	51	32	361

Table 2 *Tolerant* assessment (see text for detailed explanation)

Assessment	1	2	3	4	5	6	7	8	9	10	11	12	∑
++	3	3	6	5	5	5	4	6	4	4	4	4	53
+	4	7	11	2	6	9	3	12	3	6	10	12	85
o	15	8	6	3	12	18	5	20	7	13	21	8	136
–	6	6	4	7	1	4	0	1	1	4	10	7	51
- -	4	1	2	4	1	1	2	3	3	8	6	1	36
++/+/o	22	18	23	10	23	32	12	38	14	23	35	24	274 (76%)
∑	32	25	29	21	25	37	14	42	18	35	51	32	361
Initialization failure (multiple countings)													
no initialization at all	0	0	1	1	0	0	0	0	0	1	0	0	3
multiple	2	1	0	1	0	1	1	3	1	2	4	0	16
localization	8	5	4	2	7	11	1	9	2	7	13	4	73
late	8	5	5	7	1	4	0	1	1	3	7	6	48
Other failure reasons (multiple countings)													
wrong model	8	3	4	0	4	3	2	6	4	9	9	6	58
undetected	3	0	2	3	0	0	1	1	1	2	5	1	19
tracking failures													
tracking failure	0	1	0	0	2	5	2	6	3	3	2	2	26

Fig. 14 Results on `stau02` image sequence at frame 1600

occlusions' of vehicles by masts, see left panel of Fig. 1)—have been counted, too, as satisfactory because many of these vehicles had also been initialized semi-interactively *at the same position* in the case of `Xtrack`. Furthermore, late *re*-initializations related to temporary occlusion of a vehicle have also been assessed as satisfactory (see Fig. 13). Based on such a 'fair' assessment, 83% of all vehicles from the Reference Set have been initialized and tracked satisfactorily by the `Motris` system as described above, see Table 3.

7 Conclusions

If we do not punish the automatic initialization approach for difficulties which an interactive initialization with an adapted vehicle model can not yet handle either—for example a series of occlusions by traffic-signal masts and pedestrians as illustrated in the upper right corner of the FoV in Fig. 1(left panel)—then the *automatic* initialization results in a tracking rate of about 80% or better *despite the fact* that it does not require an explicit lane model and, moreover, uses a more general polyhedral model for *all passenger* vehicles. This approach thus constitutes a significant advance compared to the earlier `Xtrack`-system.

In a sense, the system approach described here marks the beginning of a transition phase following two decades of research on 3D-model-based vehicle tracking in road traffic videos: This system comprises all components required for *geometry-controlled* vehicle tracking which is mature

Table 3 Detailed results obtained for a *fair* assessment (see text)

Assessment	1	2	3	4	5	6	7	8	9	10	11	12	$\sum$
++	3	3	6	5	5	5	4	6	4	4	4	4	53
+	4	7	11	2	6	9	3	12	3	6	10	12	85
o	20	13	8	8	13	21	5	21	7	14	23	12	165
–	1	1	2	2	0	1	0	0	1	3	8	3	22
- -	4	1	2	4	1	1	2	3	3	8	6	1	36
++/+/o	27	23	25	15	24	35	12	39	14	24	37	28	303 (83%)
$\sum$	32	25	29	21	25	37	14	42	18	35	51	32	361

enough to justify its evaluation on such a large testing sample. At the same time, the discussion in the preceding section illustrates numerous bottlenecks which need careful attention by future research.

The transition aspect is seen in the fact that the necessary improvements and modifications can now be evaluated within an *entire systems framework*, as to be distinguished from limited tests and assessments of isolated components.

Acknowledgements The investigations have been partially supported by the European Union FP5-Project CogViSys—IST-2000-29404 and by the European Union Project HERMES (6thFP-IST-027110).

References

Collins, R. T., Lipton, A. J., & Kanade, T. (2000). Introduction to the special section on video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 745–746.

Dahlkamp, H., Nagel, H.-H., Ottlik, A., & Reuter, P. (2007). A framework for model-based tracking experiments in image sequences. *International Journal of Computer Vision*, *73*(2), 139–157.

Elgammal, A., Duraiswami, R., Harwood, D., & Davis, L. S. (2002). Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, *90*(7), 1151–1163.

Haag, M., & Nagel, H.-H. (1999). Combination of edge element and optical flow estimates for 3D-model-based vehicle tracking in traffic image sequences. *International Journal of Computer Vision*, *35*(3), 295–319.

Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *34*(3), 334–352.

Kastrinaki, V., Zervakis, M., & Kalaitzakis, K. (2003). A survey of video processing techniques for traffic applications. *Image and Vision Computing*, *21*(4), 359–381.

Kumar, P., Ranganath, S., Sengupta, K., & Weimin, H. (2004). Cooperative multi-target tracking and classification. In T. Pajdla & J. Matas (Eds.), *Lecture notes in computer science: Vol. 3021–3024. Proceedings of the 8th European conference on computer vision* (pp. I:376–389), Prague, Czech Republic, 11–14 May 2004. Berlin: Springer.

Leibe, B., Cornelis, N., Cornelis, K., & Van Gool, L. (2006). Integrating recognition and reconstruction for cognitive traffic scene analysis from a moving vehicle. In K. Franke, K.-R. Müller, B. Nickolay, & R. Schäfer (Eds.), *Lecture notes in computer science: Vol. 4174. Pattern recognition, proceedings of the 28th DAGM-symposium (DAGM 2006)* (pp. 192–201), Berlin, Germany, 12–14 September 2006. Berlin: Springer.

Magee, D. R. (2004). Tracking multiple vehicles using foreground, background and motion models. *Image and Vision Computing*, *22*(2), 143–155.

Middendorf, M. (2004). *Zur Auswertung lokaler Grauwertstrukturen*. Juli 2004 (in German, ISBN 3-8334-1175-9).

Middendorf, M., & Nagel, H.-H. (2001). Estimation and interpretation of discontinuities in optical flow fields. In *Proceedings of the eighth international conference on computer vision* (Vol. I, pp. 178–183), Vancouver, BC, Canada, 9–12 July 2001. Los Alamitos: IEEE Computer Society.

Nadimi, S., & Bhanu, B. (2004). Physical models for moving shadow and object detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(8), 1079–1087.

Nagel, H.-H. (1988). From image sequences towards conceptual descriptions. *Image and Vision Computing*, *6*(2), 59–74.

Nagel, H.-H. (2004). Steps toward a cognitive vision system. *AI-Magazine*, *25*(2), 31–50.

Ottlik, A. (2005). Zur modellgestützten Initialisierung von Fahrzeugverfolgungen in Videoaufzeichnungen. In *Dissertationen zur Künstlichen Intelligenz (DISKI)* (Vol. 291). Berlin: Akademische Verlagsgesellschaft Aka GmbH. Dissertation, Februar 2005, Fakultät für Informatik der Universität Karlsruhe (TH) (in German).

Pece, A. E. C. (2006). Contour tracking based on marginalized likelihood ratios. *Image and Vision Computing*, *24*(3), 301–317.

Pece, A. E. C., & Worrall, A. D. (2002). Tracking with the EM contour algorithm. In A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Eds.), *Lecture notes in computer science: Vol. 2350–2353. Proceedings of the 7th European conference on computer vision (ECCV 2002, Parts I–IV)* (pp. I:3–17), Copenhagen, Denmark, 27 May–2 June 2002. Berlin: Springer.

Pece, A. E. C., & Worrall, A. D. (2006). A comparison between feature-based and EM-based contour tracking. *Image and Vision Computing*, *24*(11), 1218–1232.

Renno, J., Lazarevic-McManus, N., Makris, D., & Jones, G. A. (2006). Evaluating motion detection algorithms: issues and results. In G. A. Jones, (Ed.), *Proceedings of the sixth IEEE international workshop on visual surveillance (VS 2006)* (pp. 97–104),

Graz, Austria, 13 May 2006. Faculty of Computing, Information Systems and Mathematics, Kingston University, Penrhyn Road, Kingston upon Thames, Surrey, UK KT1 2EE.

Stauffer, C., & Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 747–757.

Sun, Z., Bebis, G., & Miller, R. (2006). On road vehicle detection: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 694–711.

Tan, T. N., Sullivan, G. D., & Baker, K. D. (1998). Model-based localization and recognition of road vehicles. *International Journal of Computer Vision*, *27*(1), 5–25.