Model-Based Initialisation of Vehicle Tracking: Dependency on Illumination

H. Leuck and H.-H. Nagel
Institut für Algorithmen und Kognitive Systeme
Universität Karlsruhe (TH)
Postfach 6980, 76128 Karlsruhe, Germany

Abstract

Although a model-based vehicle tracking approach offers the promise to be more reliable than a purely data-driven one, based on the additional knowledge brought to bear during the tracking phase, a suitable initialisation of the tracking phase still presents considerable problems. Part of these difficulties are related to the appropriate choice of assumptions concerning the prevailing illumination of the recorded scene. We present an approach to automatically detect elongated shadow-casting structures in the scene and exploit these structures to automatically distinguish between directed and diffuse illumination. Extended experiments with real-world traffic scenes illustrate the principal practicality of this approach, but simultaneously reveal unexpected difficulties.

1. Introduction

It is well known that model-based machine vision can offer substantial advantages – provided an appropriate model is used. The resulting challenge thus consists in the provision of a suitably large range of models, together with a reliable and fast selection process. Alternatively, a suitably flexible generic description can be used together with a sufficiently reliable and fast adaptation process.

It will be helpful for the following to clearly distinguish between system-internal representations referring to phenomena in the image plane (2D-models) and representations of phenomena in the depicted 3D world. Among the latter, surfaces and bodies in the 3D-space are usually required although representations of 3D illumination conditions, 3D movements, actions or situated sequences of actions (i. e. behaviors) in 3D-space may be needed, too. Our study uses the application domain of road traffic in order to investigate the interaction between a 3D moving rigid body-model and time-varying illumination conditions, in particular during the initialisation phase of a tracking system.

The fundamentals for 3D model-based tracking of rigid bodies in image sequences [5] are understood well enough by now to encourage the study of important aspects in more detail, in particular the effects of various initialisation approaches on the highly non-linear 3D-model-based tracking processes. Recent publications [8, 1, 6] have shown that different approaches succeed in the initialisation and tracking of isolated larger vehicles which move sufficiently fast in front of a sufficiently uniform background. Problems begin to show up, however, if vehicle images of reduced size have to be detected and localized, if the contrast between vehicles and foreground or background diminishes, or if vehicles begin to cluster densely – for example close to red traffic lights or congested road sections. Three different, but mutually interacting questions arise in this context:

1. For short pre-digitized test image sequences, detection, initialisation, and tracking of vehicles start either in the first frame or when a vehicle in question enters the field of view at an usually a-priori known frame number. In
the long run, this a-priori knowledge will not be available. Each image frame has to be scrutinized whether or not new vehicles can be detected and should be initialized for tracking.

2. It appears advantageous to exploit the knowledge about vehicles expected in the recorded traffic scene – as encoded in the form of vehicle models – for all phases of the initialization process, not only for tracking. A model-dependent shadow cast by a vehicle thus should already influence the initialization.

3. As a consequence of the preceding consideration, the actually prevailing illumination conditions have to be determined and properly evaluated, both for initialisation and for tracking. Publications which report about shadows taken into account for model-based vehicle tracking – like, e.g., [2, 1] – rely on a-priori knowledge about the type of illumination (directed vs. diffuse) and about the direction of the incoming light.

This contribution describes an approach which attempts to cope with all three subproblems. The advantage of such an encompassing approach consists in the possibility to study the interaction between the consequences of each one. Time-consuming systematic experiments reported in Section 4 allow a first assessment of this approach.

2. Detection of the Actual Illumination

Experience has shown that the illumination of outdoor scenes may change within seconds during cloudy days with heavy winds: the sunlight, for example, can rapidly become occluded under such conditions, changing the illumination from directed to diffuse. When extensive recordings of traffic scenes have to be evaluated, it becomes necessary to automatically detect the actually prevailing illumination in order to select the appropriate initialisation process.

This problem is handled here by a three-step approach: (i) automatic determination of scene structures which cast detectable and simply interpretable shadows, (ii) image sequences are recorded together with data about the time and geographical location of the recorded scene, and (iii) development of a decision procedure which exploits this information in order to distinguish between directed and diffuse illumination.

2.1 Time-Indexed Recording of Traffic

In order to determine the direction of shadows in the image cast due to sunlight, the exact time of recording and the exact location of the recorded scene must be known. A timestamp with GMT is imprinted, therefore, into each digitized image of the sequence. In addition, the geographical location of the camera – and thereby the scene – is determined using, e.g., a Global Positioning System (GPS).

Figure 2. Edges compatible with the expected direction of shadows, accumulated over time and backprojected onto the ground (road) plane from Figure 1 as viewed from above.

Based on this information, the relative position of the sun with respect to the scene location at a certain time can easily be calculated (see, e.g., [5]). A few simple coordinate transformations provide the direction of sunrays in scene coordinates and thus the direction of shadows cast onto the ground plane by known straight long structures such as masts or tree trunks. The direction of corresponding shadow images can then be calculated based on the external camera calibration – see Figure 1.

2.2 Automatic Localisation of Shadow-Casting Masts in the Scene

The approach to be described in the sequel uses a calibrated camera. One thus might be tempted to determine suitable shadow casting structures in the scene during interactions necessitated by camera calibration. Since we expect, however, that camera calibration will be automatized in the future, we decided to develop an automatic determination of shadow casting structures in the scene. The idea is to detect image regions where Edge Elements (EEs) extracted from an image sequence ‘move’ in the image plane as a function of time consistently with the expected direction of shadows. Such ‘moving shadow edges’ are treated as cues to the position of long, thin objects raising perpendiculary from the – assumed planar – ground of the scene, for
Figure 3. White patches represent significant clusters of intersections of edge segments compatible with the expected direction of moving shadows. Black crosses mark likely ground plane positions of automatically detected shadow-casting objects like masts.

Figure 4. Edge elements (EEs, white pixels) in an image region around expected shadow of a mast (thin white lines in road plane).

example lamp posts or masts carrying traffic signs. We hypothesize, moreover, that shadows are cast onto the ground plane and not onto other objects which raise above ground, for example vehicles. Figure 2 shows accumulated EEs for selected times during a day, backprojected onto the ground plane as viewed from above, provided they are compatible with the expected direction of shadows.

Ideally, edge segments derived from elongated clusters of such EEs should intersect in the same spot — the image of the ground plane location of a shadow-casting object. In practice, however, even selected edge segments do not intersect exactly in the same point, for example due to distortions in pixel grayvalues, effects associated with image digitisation, or artefacts of EE extraction. The cluster centroids of neighboring intersection points associated with a ‘fan’ of edge segments are assumed to represent good approximations for the image of the foot of a mast raising perpendicularly from the ground plane. Figure 3 shows positions detected in the described manner for a typical traffic scene.

Assuming a shadow to be evaluated is cast by a long, thin object raising perpendicularly from the ground plane, a (generic) geometrical model of a mast is placed hypothetically at the detected position in the scene. Based on calculations corresponding to those mentioned in Section 2.1, an image region can be determined where shadows cast by the object in question are expected — see Figure 4. Within this region, all EEs are evaluated with respect to their distance and their difference in orientation from expected shadow edges. If the corresponding image region for an expected shadow image in the training image sequence is not covered significantly enough by EEs, the hypothesized mast position has to be rejected.

Since shadows cast by sunlight onto a ground plane move at a rate of 15 degrees per hour, it suffices to record about one image every twenty minutes. It is necessary, however, to record images for a few hours in order to cover a significant change in the direction of shadows. In our experiments, we recorded a training image sequence digitising one image every 20 minutes over a period of approximately 10 hours on a sunny, cloudless day.

2.3 Determination of Illumination Conditions

In order to distinguish between directed and diffuse illumination during initialisation and tracking, we rely on the automatically determined positions of shadow-casting objects to estimate the illumination conditions prevailing in the depicted scene at a given time (compare preceding section and Figure 5). The higher the score, the more likely it is that the scene is directly illuminated by sunlight.

Figure 6 shows the accumulated lengths of all expected projected shadow edge segments and the sum of scores of all extracted EEs as a function of time for an image sequence recorded on a day with rapidly changing illumination conditions. EEs have been estimated at 6 sec intervals between about 10:00 a.m. and 12:20 p.m., i.e. for a period of more than 2 hours.
Figure 5. Brighter pixels represent higher scores, darker pixels lower scores for automatically evaluated EE.

Parts of the curve with high scores correspond to time slices when the sunlight caused strong shadows on the lane, curve parts with low scores correspond to diffuse illumination because a cloud temporarily covered the sun. A threshold depending on the lengths of expected shadows in the image can be used to distinguish between directed and diffuse illumination, in the case of Figure 6 for example at 95.

3. Model-Based Initialisation

Model-based tracking can be initialized by search for a suitable fit between the projection of a given vehicle model into the image plane and EE extracted from the frame in question [8]. Testing different vehicle models and accepting the one with the best fit allows in principle to even determine the vehicle type. Alternatively, either a difference image ([6]) or an optical flow field ([2, 1]) is estimated and segmented in order to obtain an initial vehicle pose. The authors of [2, 1] append a transition phase to the purely data-driven initialisation: they use an Iterated Extended Kalman Filter (IEKF) – by exploitation of knowledge encoded into an interactively selected polyhedral vehicle model – in order to improve the state estimation further before entering the model-based tracking phase proper. Whereas [6] do not exploit shadow casting during tracking, the IEKF used by [2, 1] for state estimation takes shadow-casting into account. The latter authors do not, however, exploit illumination effects and shadow casting during the purely data-driven initialisation phase.

Since a vehicle model has to be provided anyway for the tracking phase, it may as well be exploited already during the initialisation phase. A recently outlined idea [7] allows to operationalize this suggestion. Given a 3D vehicle model together with knowledge about the incoming light direction, one may compute the vector between the centroids of the image plane area due to the projection of the vehicle model with shadow and without shadow. In case the scene is illuminated by oriented rather than diffuse light, the region corresponding to a moving vehicle both in the difference image as well as in the optical flow field will comprise moving shadows. One thus could subtract the difference vector from the segmented region centroid in order to obtain an estimate of the centroid corresponding to the vehicle image area without shadow. This should yield a better initial estimate for the vehicle’s position in the scene.

4. Comparison of Initialisation Alternatives

The utilization of a vehicle model during initialisation as outlined above has been combined with the estimation of the actual illumination condition. This combination has been integrated into a model-based tracking system reported elsewhere [4]. The combined approach has been studied by evaluation of three image sequences recorded under different illumination conditions: constant diffuse illumination, constant directed illumination, as well as repeated – comparatively sharp – changes between diffuse and directed ill-
lumination. Since we thus have three different image sequences, three alternatives in modelling the prevailing illumination conditions (constant diffuse, constantly directed, or automatically determined), and two alternative initialisation modi (purely data-driven or using the vehicle model), 18 different experiments have been performed, retaining all other initialisation and tracking parameters unchanged.

To remain consistent with our goal of improving a fully automatic initialisation for model-based vehicle tracking, we did not intervene to select the actually most likely vehicle model in each case. Instead, we always applied the same vehicle model (a fastback) without interactive adaptation of size or shape parameters. The consequences of using an inappropriate vehicle model, however, have to be separated from those related to the automatic determination of the illumination model to be used and its effects on continuous automatic initialisation. Table 1 thus comprises results only for vehicles assigned to the category ‘fastback’ during the interactive assessment of tracking results. These have been rated either as good (+), acceptable (o), or failed (−). Vehicles which could not be initialised for tracking were assigned to a category (−−). Each image sequence comprises about 2000 or more evaluated frames although the number of visible vehicles varies depending on the actual traffic at the time of recording.

The frame-incremental ‘continuous’ initialisation turned out to be partly an advantage, partly a disadvantage. In order not to re-initialize the same vehicle over and over again, initialisation has been blocked if the corresponding image area overlapped the projection of a vehicle for which a model-based tracking was in progress. In case the quality of a model-based tracking process deteriorated or if it failed entirely, the incorrectly positioned model could block other vehicles from being properly initialized. On the other hand, we observed examples where an early initialisation failed, but the vehicle in question could be picked up automatically a few frames later – in which case we counted it as a success if the subsequent tracking process proceeded without further difficulties.

A priori, one expects that tracking results derived from a model-based initialisation should be rated superior to those derived from a purely data-driven initialisation, provided the appropriate illumination model has been specified or it has been correctly determined automatically. The only study ([1]) known to us which reports results from a systematic evaluation of a larger sample recommends caution, however: the authors of [1] mention a multitude of potential causes for the roughly 15-20 % failure rate reported by them. Since the experiments reported here essentially address only one among these failure categories, one should be prepared for an improvement of at most 5-10 % in the success rate if the chosen assumptions match the actually prevailing conditions. In our experiments, initialisation (−−) and tracking (−) failures combined amount to between 1–4 out of 16 (≈ 7–25 %), 2 out of 8 (≈ 25 %), and 1–2 out of 15 (≈ 7–14 %). These rates are compatible with those reported by [1] despite the fact that they have been obtained using a continuous automatic initialisation.

Using a diffuse illumination model in the case of a constantly diffuse scene illumination is clearly advantageous compared to using a constantly directed illumination model, both for the data-driven and the model-based initialisation. For a directed scene illumination, the results obtained on the basis of a directed illumination model are equal to or better than those obtained based on the diffuse illumination model. In the case of rapidly changing illumination conditions, the results obtained with an automatic selection of the illumination model are slightly better than those obtained by enforcing a diffuse illumination and about as good as those obtained by enforcing a constantly directed illumination model.

Regarding the model-based vs. the purely data-driven initialisation, the advantage of using the former is not pronounced although the data overall indicate a (very) slight preference for the model-based approach.

5. Summary and Conclusions

We exploit the capability to directly digitize video recordings of traffic scenes under varying illumination conditions in combination with time-stamping each digitized video-frame and GPS-based camera position information. This allows to study the effect of an automatic selection between diffuse and directed scene illumination models. The selected illumination model is exploited in order to investigate the difference between data-driven versus model-based initialisation processes as well as the results obtained during a subsequent tracking phase based on the selected illumination model. The total number of tracking steps (a single vehicle tracked for a single frame interval) performed during the experiments reported above exceeds a million.

Different assumptions underlying the entire evaluation process have been combined and compared, keeping all other parameters constant. Quantitative results indicate that the choice of the appropriate illumination model (slightly) improves the initialisation and tracking even without interactive tuning of a model to each vehicle, provided the model type is correct. In the case of rapidly changing illumination conditions, results obtained by the automatic selection of the illumination model to be used are at least comparable with those based on a fixed illumination model. Given the added degrees of freedom provided by an automatic selection of the illumination model, this result is considered to be a success already. Larger sample sizes, however, are necessary in order to arrive at statistically significant conclusions. The resources required for even larger samples than those
<table>
<thead>
<tr>
<th>Actual illumination</th>
<th>Initialisation</th>
<th>Illumination model</th>
<th>+</th>
<th>0</th>
<th>−</th>
<th>−−</th>
<th># of vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>constantly</td>
<td>data-driven</td>
<td>diffuse</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>16 (30)</td>
</tr>
<tr>
<td></td>
<td>data-driven</td>
<td>directed</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>16 (30)</td>
</tr>
<tr>
<td></td>
<td>data-driven</td>
<td>automatic</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>16 (30)</td>
</tr>
<tr>
<td>diffuse</td>
<td>using model</td>
<td>diffuse</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>16 (30)</td>
</tr>
<tr>
<td></td>
<td>using model</td>
<td>directed</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>16 (30)</td>
</tr>
<tr>
<td></td>
<td>using model</td>
<td>automatic</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>16 (30)</td>
</tr>
<tr>
<td>constantly</td>
<td>data-driven</td>
<td>diffuse</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>8 (17)</td>
</tr>
<tr>
<td></td>
<td>data-driven</td>
<td>directed</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>8 (17)</td>
</tr>
<tr>
<td></td>
<td>data-driven</td>
<td>automatic</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>8 (17)</td>
</tr>
<tr>
<td>directed</td>
<td>using model</td>
<td>diffuse</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>8 (17)</td>
</tr>
<tr>
<td></td>
<td>using model</td>
<td>directed</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>8 (17)</td>
</tr>
<tr>
<td></td>
<td>using model</td>
<td>automatic</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>8 (17)</td>
</tr>
<tr>
<td>rapidly</td>
<td>data-driven</td>
<td>diffuse</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>15 (34)</td>
</tr>
<tr>
<td></td>
<td>data-driven</td>
<td>directed</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>15 (34)</td>
</tr>
<tr>
<td></td>
<td>data-driven</td>
<td>automatic</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>15 (34)</td>
</tr>
<tr>
<td>changing</td>
<td>using model</td>
<td>diffuse</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>15 (34)</td>
</tr>
<tr>
<td></td>
<td>using model</td>
<td>directed</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>15 (34)</td>
</tr>
<tr>
<td></td>
<td>using model</td>
<td>automatic</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>15 (34)</td>
</tr>
</tbody>
</table>

Table 1. Comparison between different initialisation alternatives, restricted to vehicles judged interactively as fastback: their number is given in the last column (in parentheses: total number of vehicles observed in the respective image sequences). "+" denotes good and "0" acceptable tracking results. The tracking of vehicles rated with "−" failed, whereas vehicle images rated "−−" could not be initialised with the chosen parameter set.

used here prevented such an extension so far. A careful analysis of the results obtained so far suggests that small, but possibly systematic differences between results based on different assumptions may be masked by the fact that the vehicle model has not been optimized. Obviously, a parallel initialisation for all admissible models or an automatic adaptation of a generic vehicle model appears desirable.

The important conclusion consists in the statement that the automatic estimation of illumination conditions in combination with the exploitation of model-based knowledge about vehicles for initialisation allows the transition from interactive initialisation at selected image frames to a continuous automatic initialisation for each frame.

Acknowledgments

We thank G. Eichberger and S. Bertsch for the installation of a remote video camera and Professor W. Jüling for permission to use a dedicated fibre from the broadband network of the Universität Karlsruhe to transmit video signals from this camera into our laboratory. Partial support of this research by the ‘Deutsche Forschungsgemeinschaft (DFG)’ is gratefully acknowledged.

References