Real-Time Vehicle Ego-Motion Using Stereo Pairs and Particle Filters*

Fadi Dornaika[1] and Angel D. Sappa[2]

[1] Institut Géographique National
94165 Saint Mandé, France
fadi.dornaika@ign.fr
[2] Computer Vision Center
08193 Bellaterra, Barcelona, Spain
sappa@cvc.uab.es

Abstract. This paper presents a direct and stochastic technique for real time estimation of on board camera position and orientation—the ego-motion problem. An on board stereo vision system is used. Unlike existing works, which rely on feature extraction either in the image domain or in 3D space, our proposed approach directly estimates the unknown parameters from the brightness of a stream of stereo pairs. The pose parameters are tracked using the particle filtering framework which implicitly enforces the smoothness constraints on the dynamics. The proposed technique can be used in driving assistance applications as well as in augmented reality applications. Experimental results and comparisons on urban environments with different road geometries are presented.

1 Introduction

In recent years, several vision based techniques were proposed for advanced driver assistance systems [1,2,3,4]. They can be broadly classified into two different categories: highways and urban. Most of the techniques proposed for highways environments are focused on lane and car detection, looking for an efficient driving assistance system. On the other hand, in general, techniques for urban environments are focused on collision avoidance or pedestrian detection.

Of particular interest is the estimation of on board camera position and orientation related to the current 3D road plane parameters—the ego-motion problem. Note that since the 3D plane parameters are expressed in the camera coordinate system, the camera position and orientation are equivalent to the plane parameters. Algorithms for fast road plane estimation are very useful for driver assistance applications as well as for augmented reality applications.

* This work was supported in part by the MEC project TRA2004-06702/AUT and The Ramón y Cajal Program.

M. Kamel and A. Campilho (Eds.): ICIAR 2007, LNCS 4633, pp. 469–480, 2007.

The prior knowledge of the environment is a source of information generally involved in the proposed solutions. For instance, highway driver assistance systems are based on assumptions such as: i) the vehicle is driven along two parallel lane markings [5], ii) lane markings, or the road itself, have a constant width [6], and iii) the camera position as well as its pitch angle (camera orientation with respect to the road plane) are constant values [7].

Similarly, vision-based urban driver assistance systems, also propose to use the prior knowledge of the environment to simplify the problem. Some of the aforementioned assumptions are also used on urban environment, together with additional assumptions related to urban scenes. In summary, scene prior knowledge has been extensively used. However, making assumptions cannot always solve problems. It can sometimes provide erroneous results. For instance, constant camera position and orientation, which is a generally used assumption on highways, is not so valid in an urban scenario. In the latter, camera position and orientation are continuously modified by factors such as: road imperfections or artifacts (e.g., rough road, speed bumpers), car accelerations, uphill/downhill driving, among others. [6] introduces a technique for estimating vehicle yaw, pitch and roll. It is based on the assumption that some parts of the road have a constant width (e.g., lane markings). A different approach was presented in [8]. The authors propose an efficient technique able to cope with uphill/downhill driving, as well as dynamic pitching of the vehicle. It is based on a v-disparity representation and Hough transform. The authors propose to model not only a single plane road—a straight line—but also a non-flat road geometry—a piecewise linear curve. This method is also limited since a longitudinal profile of the road should be extracted for computing the v-disparity representation.

In this paper, a new approach based on raw stereo images provided by a stereo vision system is presented. It aims to compute camera position and orientation, avoiding most of the assumptions mentioned above. Since the aim is to estimate the pose of an on board stereo camera from stereo pairs arriving in a sequential fashion, the particle filtering framework seems very useful. In other words, we track the pose of the vehicle (stereo camera) given the sequence of stereo pairs. The proposed technique could be indistinctly used for urban or highway environments, since it is not based on a specific visual traffic feature extraction neither in 2D nor in 3D. Our proposed method has a significant advantage over existing methods since it does not require road segmentation nor dense matching—two difficult tasks. Moreover, to the best of our knowledge, the work presented in this paper is the first work estimating road parameters directly from the rawbrightness images using a particle filter.

The rest of the paper is organized as follows. Section 2 describes the problem we are focusing on. Section 3 briefly describes a 3D data based method. Section 4 presents the proposed stochastic technique. Section 5 gives some experimental results and method comparisons. In the sequel, the "road plane parameters" and the"camera pose parameters" will refer to the same entity.

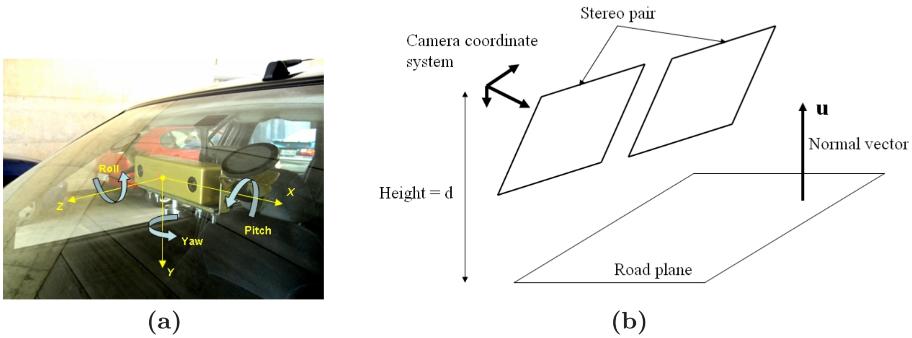


Fig. 1. **(a)** On board stereo vision sensor with its corresponding coordinate system. **(b)** The time-varying road plane parameters d and $\mathbf{u}$.

2 Problem Formulation

Experimental setup. A commercial stereo vision system (Bumblebee from Point Grey[1]) was used. It consists of two Sony ICX084 color CCDs with 6mm focal length lenses. Bumblebee is a pre-calibrated system that does not require in-field calibration. The baseline of the stereo head is 12cm and it is connected to the computer by a IEEE-1394 connector. Right and left color images can be captured at a resolution of 640×480 pixels and a frame rate near to 30 fps. This vision system includes a software able to provide the 3D data. Figure 1**(a)** shows an illustration of the on board stereo vision system as well as its mounting device.

The problem we are focusing on can be stated as follows. Given a stream of stereo pairs provided by the on board stereo head we like to recover the parameters of the road plane for every captured stereo pair. Since we do not use any feature that is associated with road structure, the computed plane parameters will completely define the pose of the on board vision sensor. This pose is represented by the height d and the plane normal $\mathbf{u} = (u_x, u_y, u_z)^T$ (See Figure 1**(b)**). Due to the reasons mentioned above, these parameters are not constant and should be estimated online for every time instant.

Image transfer function. It is well-known [9] that the projections of 3D points belonging to the same plane onto two different images are related by a 2D projective transform having 8 independent parameters—*homography*. In our setup, the right and left images are horizontally rectified[2]. Let $p_r(x_r, y_r)$ and $p_l(x_l, y_l)$ be the right and left projection of an arbitrary 3D point P belonging to the road plane (d, u_x, u_y, u_z) (see Figure 2). In the case of a rectified stereo pair where the left and right images have the same intrinsic parameters, the right and left coordinates of corresponding pixels belonging to the road plane

[1] `[www.ptgrey.com]`

[2] The use of non-rectified images will not have any theoretical impact on our developed method. However, the image transfer function will be given by a general homography.

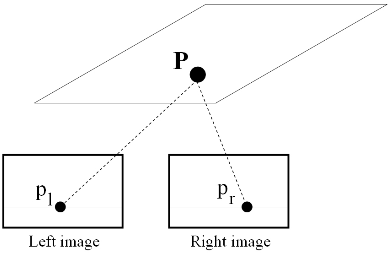


Fig. 2. The mapping between corresponding left and right road pixels is given by a linear transform

are related by the following linear transform, i.e., the homography reduces to a linear mapping:

$$x_l = h_1\, x_r + h_2\, y_r + h_3 \tag{1}$$

$$y_l = y_r \tag{2}$$

where h_1, h_2, and h_3 are function of the intrinsic and extrinsic parameters of the stereo head and of the plane parameters. For our setup (rectified images with the same intrinsic parameters), those parameters are given by:

$$h_1 = 1 + b\frac{u_x}{d} \tag{3}$$

$$h_2 = b\,\frac{u_y}{d} \tag{4}$$

$$h_3 = -b\,u_0\frac{u_x}{d} - b\,v_0\frac{u_y}{d} + \alpha\, b\frac{u_z}{d} \tag{5}$$

where b is the baseline of the stereo head, α is the focal length in pixels, and (u_0, v_0) is the image center—the principal point.

3 3D Data Based Approach

In [10], we proposed an approach for on-line vehicle pose estimation using the above commercial stereo head. The camera position and orientation—the road plane parameters—are estimated from raw 3D data. The proposed technique consists of two stages. First, a dense depth map of the scene is computed by the provided dense matching technique. Second, the parameters of a plane fitting to the road are estimated using a RANSAC based least squares fitting. Moreover, the second stage includes a filtering step that aims at reducing the number of 3D points that are processed by the RANSAC technique. The proposed technique could be indistinctly used for urban or highway environments, since it is not based on a specific visual traffic feature extraction but on raw 3D data points. This technique has been tested on different urban environments. The proposed algorithm took, on average, 350 ms per frame. The main drawback of the proposed 3D data technique is its high CPU time. Moreover, it requires a dense 3D reconstruction of the captured images. In the current study, this method is used for comparisons with the proposed stochastic technique. It can be used to initialize the proposed approach—providing the solution for the first video frame.

4 A Direct and Stochastic Approach

Our aim is to estimate the pose parameters from the stream of stereo pairs. In other words, we track the pose over time. In this section, we propose a novel approach that directly infers the plane parameters from the stereo pair using the particle filtering framework. The idea of a particle filter (also known as Sequential Monte Carlo (SMC) algorithm) was independently proposed and used by several research groups. These algorithms provide flexible tracking frameworks as they are neither limited to linear systems nor require the noise to be Gaussian and proved to be more robust to distracting clutter as the randomly sampled particles allow to maintain several competing hypotheses of the hidden state. Therefore, the main advantage of particle filtering methods is the fact that any loose of track will not lead to permanent loss of the object. Note that when the noise can be modelled as Gaussian and the observation model is linear then the solution will be given by the Kalman filter.

Particle filtering is an inference process which aims at estimating the unknown time-t state $\mathbf{b}_t$ from a set of noisy observations (images), $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \cdots, \mathbf{z}_t\}$ arriving in a sequential fashion [11,12]. Two important components of this approach are the state transition and observation models. The particle filter approximates the posterior distribution $p(\mathbf{b}_t|\mathbf{z}_{1:t})$ by a set of weighted particles or samples $\{\mathbf{b}_t^{(j)}, \pi_t^{(j)}\}_{j=1}^N$. Each element $\mathbf{b}_t^{(j)}$ represents the hypothetical state of the object and $\pi_t^{(j)}$ is the corresponding discrete probability. Then, the state estimate can be set for example to the minimum mean square error or to the maximum *a posteriori* (MAP): $\arg\max_{\mathbf{b}_t} p(\mathbf{b}_t|\mathbf{z}_{1:t})$.

Based on such generative models, the particle filtering method is a Bayesian filtering method that recursively evaluates the posterior density of the target state at each time step conditionally to the history of observations until the current time.

4.1 Algorithm

Dynamics model. At any given time, the road plane parameters are given by the plane normal $\mathbf{u}_t$, a unit vector, and the distance d_t between the camera center and the plane. These parameters can be encapsulated into a 3-vector $\frac{\mathbf{u}_t}{d_t}$. Therefore, the state vector $\mathbf{b}_t$ representing the plane parameters will be given by

$$\mathbf{b}_t = (b_{x(t)}, b_{y(t)}, b_{z(t)})^T = (\frac{u_{x(t)}}{d_t}, \frac{u_{y(t)}}{d_t}, \frac{u_{z(t)}}{d_t})^T \tag{6}$$

Note that the vector $\mathbf{b}_t$ fully describes the current road plane parameters since the normal vector is a unit vector. Since the camera height and orientation, the

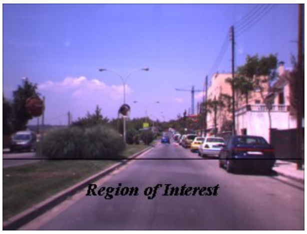


Fig. 3. The region of interest associated with the right image. In this example, its height is set to one third of the image height.

plane parameters, are ideally constant, the dynamics of the state vector $\mathbf{b}_t$ can be well modelled by a Gaussian noise:

$$b_{x(t)} = b_{x(t-1)} + \epsilon_t \tag{7}$$
$$b_{y(t)} = b_{y(t-1)} + \epsilon_t \tag{8}$$
$$b_{z(t)} = b_{z(t-1)} + \epsilon_t \tag{9}$$

where ϵ is a noise (scalar) drawn from a centered Gaussian distribution $\mathcal{N}(0, \sigma)$. The standard deviation of the noise can be computed from previously recorded camera pose variations. However, we believe that fixed standard deviations or context-based standard deviations are more appropriate since they are directly related to the kind of perturbations and to the video rate.

Observation model. The observation model should relate the state $\mathbf{b}_t$ (plane parameters) to the measurement $\mathbf{z}_t$ (stereo pair). We use the following fact: *if the state vector encodes the actual values of the plane distance and of its normal, then the registration error between corresponding road pixels in the right and left images should correspond to a minimum.* In our case, the measurement $\mathbf{z}_t$ is given by the current stereo pair. The registration error is simply the Sum of Squared Differences between the right image and the corresponding left image computed over a given region of interest. The registration error is given by:

$$e(\mathbf{b}) = \frac{1}{N_p} \sum_{(x_r, y_r) \in ROI} (I_{r(x_r, y_r)} - I_{l(h_1\, x_r + h_2\, y_r + h_3, y_r)})^2 \tag{10}$$

where N_p is the number of pixels contained in the region of interest. The above summation is carried out over the right region of interest. The corresponding left pixels are computed according to the affine transform (1). The computed $x_l = h_1\, x_r + h_2\, y_r + h_3$ is a non-integer value. Therefore, $I_l(x_l)$ is set to a linear interpolation between two neighboring pixels.

Note that the region of interest is a user-defined region. Ideally, this region should not include non-road objects but as will be seen in the experiments this is not a hard constraint because we use a stochastic tracking technique.

In our study, the ROI is set to a rectangular window that roughly covers the lower part of the original image (one third). Figure 3 illustrates a typical region of interest.

The observation likelihood is given by

$$p\left(\mathbf{z}_t|\mathbf{b}_t\right) = \frac{1}{\sqrt{2\pi}\,\sigma_e}\exp\left(-\frac{e(\mathbf{b}_t)}{2\sigma_e^2}\right) \tag{11}$$

where σ_e is a parameter controlling the aperture of the Gaussian distribution.

Computing the state $\mathbf{b}_t$ from the previous posterior distribution $p(\mathbf{b}_{t-1}|\mathbf{z}_{1:t-1})$ is carried out using the particle filtering framework described in Figure 4.

1. Initialization $t = 0$: Generate N state samples $\mathbf{a}_0^{(1)}, \ldots, \mathbf{a}_0^{(N)}$ according to some prior density $p(\mathbf{b}_0)$ and assign them identical weights, $w_0^{(1)} = \ldots = w_0^{(N)} = 1/N$
2. At time step t, we have N weighted particles $(\mathbf{a}_{t-1}^{(N)}, w_{t-1}^{(N)})$ that approximate the posterior distribution of the state $p(\mathbf{b}_{t-1}|\mathbf{z}_{1:(t-1)})$ at previous time step
   - (a) Resample the particles proportionally to their weights, *i.e.* keep only particles with high weights and remove particles with small ones. Resampled particles have the same weights
   - (b) Draw N particles according to the dynamic model $p(\mathbf{b}_t|\mathbf{b}_{t-1} = \mathbf{a}_{t-1}^{(j)})$ (7), (8), and (9). These particles approximate the predicted distribution $p(\mathbf{b}_t|\mathbf{z}_{1:(t-1)})$
   - (c) Weight each new particle proportionally to its likelihood:

     $$w_t^{(j)} = \frac{p(\mathbf{z}_t|\mathbf{b}_t = \mathbf{a}_t^{(j)})}{\sum_{m=1}^{N} p(\mathbf{z}_t|\mathbf{b}_t = \mathbf{a}_t^{(m)})}$$

     where $p(\mathbf{z}_t|\mathbf{b}_t)$ is given by (11). The set of weighted particles approximates the posterior $p(\mathbf{b}_t|\mathbf{z}_{1:t})$
   - (d) Give an estimate of the state $\hat{\mathbf{b}}_t$ as the MAP:

     $$\hat{\mathbf{b}}_t = \arg\max_{\mathbf{b}_t} p(\mathbf{b}_t|\mathbf{z}_{1:t}) \approx \arg\max_{\mathbf{a}_t^{(j)}} w_t^{(j)}$$

Fig. 4. Particle filter algorithm

Initialization. Note that the initial distribution $p(\mathbf{b}_0)$ can be either a Dirac or Gaussian distribution centered on a provided solution. We have used two methods for estimating this solution: i) the 3D data-based algorithm, and ii) the differential evolution algorithm [13] which aims at minimizing the registration error (10).

5 Experiments

The proposed technique has been tested on different urban environments.

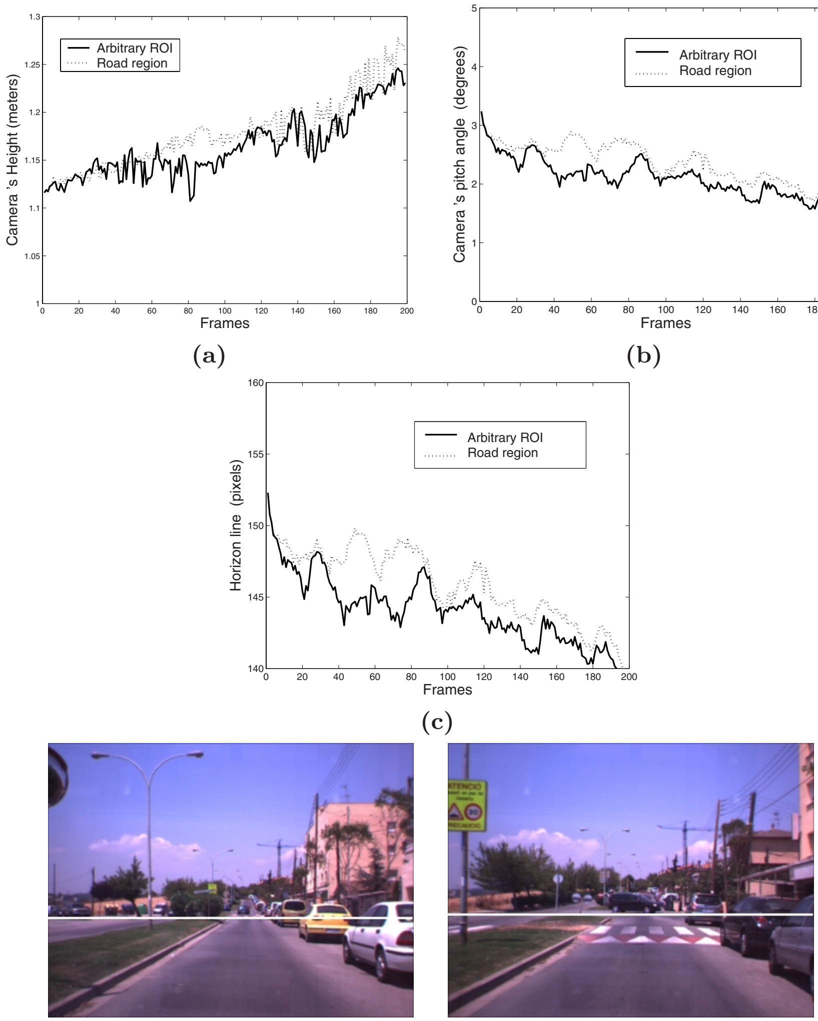
First experiment. The first experiment has been conducted on a short sequence of stereo pairs corresponding to a typical urban environment (see Figure 3). The stereo pairs are of resolution 320 × 240. Here the road is almost flat and the perturbations are due to accelerations and decelerations of the car. Figures 5(**a**) and 5(**b**) depict the estimated camera height and orientation as a function of the sequence frames, respectively. The plotted solutions correspond to the Maximum a Posteriori solution. The solid curves corresponds to an arbitrary ROI of size 270×80 pixels centered at the bottom of the image. The dotted curves correspond to a ROI covering the road region only (here the ROI is manually set to 200 × 80 pixels centered at the bottom of the image). The arbitrary ROI includes some objects that do not belong to the road plane—the vehicles on the right bound. As can be seen, the estimated solutions associated with the two cases are almost the same, which suggests that the obstacles will not have a big impact on the solution. In this experiment the number of particles N was set to 200 and the parameters were as follows $\sigma = 0.002$ and $\sigma_e = 1$.

In the literature, the pose parameters (road plane parameters) can be represented by the horizon line. Figure 5(**c**) depict the vertical position of the horizon line as a function of the sequence frames. Figure 5(**d**) illustrates the computed horizon line for frames 55 and 182.

In order to study the algorithm behavior in the presence of significant occlusions or camera obstructions, we conducted the following experiment. We used the same sequence of Figure 3. We run the proposed technique described in Section 4 twice. In the first run the stereo images were used as they are. In the second run, the right images were modified to simulate a significant occlusion. To this end, we set the vertical half of a set of 20 right frames to a constant color. The left frames were not modified[3]. Figure 6 compares the pose parameters obtained in the two runs. The solid curves were obtained with the non-occluded images (first run). The dotted curves were obtained in the second run. The occlusion starts at frame 40 and ends at frame 60. The upper part of the figure illustrates the stereo pair 40. As can be seen, the discrepancy that occurs at the occlusion is considerably reduced when the occlusion disappears.

Second experiment and method comparison. The second experiment has been conducted on another short sequence corresponding to an uphill driving. The stereo pairs are of resolution 160 × 120. Figures 7(**a**) and 7(**b**) depict the estimated camera height and orientation as a function of the sequence frames, respectively. The solid curves correspond to the developed stochastic approach.The dashed curves correspond to the 3D data based approach

[3] Therefore, there is a sudden increase in the registration error.



(a)

(b)

(c)

Fig. 5. Camera position and orientation, estimated by the particle filter. The plotted solutions correspond to the Maximum a Posteriori (MAP) solution. The solid curves corresponds to a ROI having an arbitrary width. The dotted curves correspond to a ROI containing the road image only.

obtained with full resolution images, i.e., 640 × 480 [10]. Figure 7**(c)** depicts the estimated position of the horizon line obtained by three methods: the above two methods (solid and dashed curves) and a manual method (dotted curve) based on the intersection of two parallel lines. As can be seen, the horizon

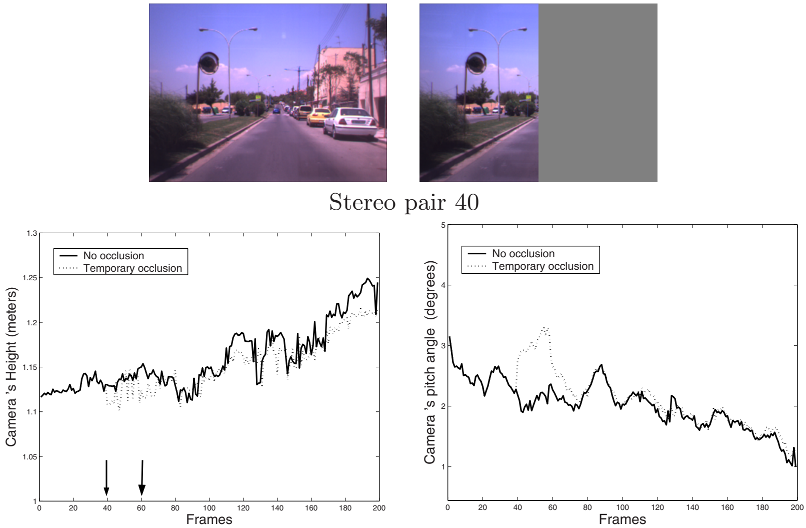

Stereo pair 40

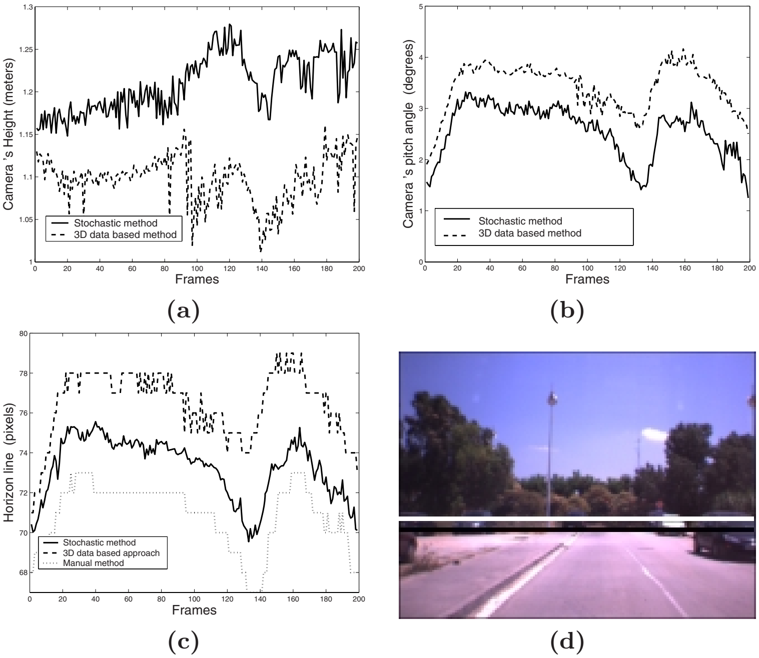
Fig. 6. Comparing the pose parameters when a significant occlusion or camera obstruction occurs. This occlusion starts at frame 40 and ends at frame 60.

line estimated by the proposed featureless approach is closer to the manually estimated horizon line—assumed to be very close to the ground-truth data. Figure **7(d)** displays the horizon line associated with frame 160. The white line corresponds to the proposed technique and the black one to the 3D data based technique.

Note that even the first stereo frame was initialized with the 3D data-based approach, the solid and dashed curves (the two methods) do not coincide at the first frame (see Figure **7(a)**). This is because only the MAP solution is plotted. According to the obtained results, the average discrepancy in the height was about 10 cm[4] and in the pitch angle was smaller than one degree.

Figure 8 displays the estimated camera height associated with the sequence of Figure 7 when the number of particles was set to 300, 200, 100, and 50. As can be seen, the estimated parameters were very consistent. A similar behavior was obtained with the pitch angle. A non-optimized C code processes one stereo pair in 30 ms assuming the size of the ROI is 6000 pixels and the number of particles is 100. The proposed approach runs almost twelve times faster than the 3D data-based approach (Section 3). Moreover, our stochastic approach is faster than many approaches based on elaborated road segmentation and detection.

[4] By assuming that this discrepancy is an upper bound of the camera height error, the latter can be considered small given the fact that the camera height was estimated with a small focal length (200 pixels) and with a small baseline.



(a) **(b)**

(c) **(d)**

Fig. 7. Method comparison for on board camera pose estimation. The solid curves correspond to the developed stochastic approach and the dashed curves to the 3D data based approach obtained with full resolution images, i.e., 640 × 480. **(d)** displays the horizon line associated with frame 160 obtained with two automatic methods: the proposed technique (white) and the 3D data based technique (black).

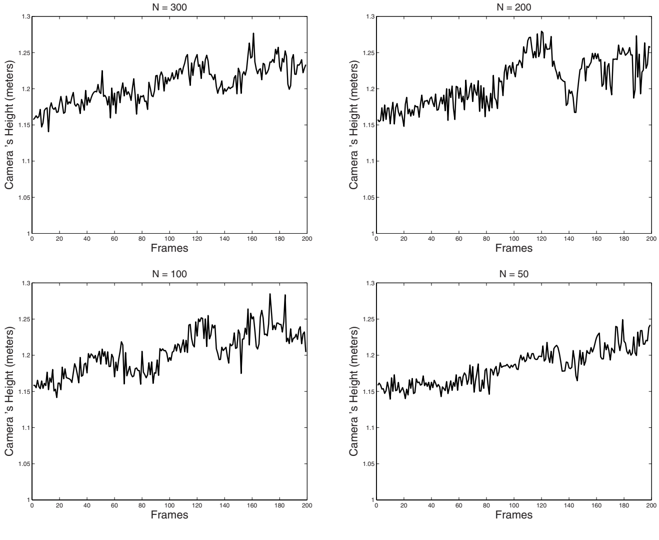


Fig. 8. The estimated camera height obtained by the proposed stochastic approach for different numbers of particles. From up to bottom $N = 300, 200, 100$, and 50.

6 Conclusion

A featureless and stochastic technique for real time ego-motion estimation of on board vision system has been presented. The method adopts a particle filtering scheme that uses images' brightness in its observation likelihood. The advantages of the proposed technique are as follows. First, the technique does need any feature extraction neither in the image domain nor in 3D space. Second, the technique inherits the strengths of stochastic tracking approaches. A good performance has been shown in several scenarios—uphill, downhill and a flat road. Furthermore, the technique can handle significant occlusions. Although it has been tested on urban environments, it could be also useful on highways scenarios.

References

1. Broggi, A., Bertozzi, M., Fascioli, A., Sechi, M.: Shape-based pedestrian detection. In: Procs. IEEE Intelligent Vehicles Symposium, Dearborn, pp. 215–220. IEEE Computer Society Press, Los Alamitos (2000)
2. Labayarde, R., Aubert, D.: A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereovision. In: IEEE Intelligent Vehicles Symposium, IEEE Computer Society Press, Los Alamitos (2003)
3. Lefée, D., Mousset, S., Bensrhair, A., Bertozzi, M.: Cooperation of passive vision systems in detection and tracking of pedestrians. In: Proc. IEEE Intelligent Vehicles Symposium, Parma, Italy, pp. 768–773. IEEE Computer Society Press, Los Alamitos (2004)
4. Liu, X., Fujimura, K.: Pedestrian detection using stereo night vision. IEEE Trans. on Vehicular Technology 53(6), 1657–1665 (2004)
5. Liang, Y., Tyan, H., Liao, H., Chen, S.: Stabilizing image sequences taken by the camcorder mounted on a moving vehicle. In: Procs. IEEE Intl. Conf. on Intelligent Transportation Systems, Shangai, China, pp. 90–95. IEEE Computer Society Press, Los Alamitos (2003)
6. Coulombeau, P., Laurgeau, C.: Vehicle yaw, pitch, roll and 3D lane shape recovery by vision. In: Proc. IEEE Intelligent Vehicles Symposium, Versailles, France, pp. 619–625. IEEE Computer Society Press, Los Alamitos (2002)
7. Bertozzi, M., Broggi, A.: GOLD: A parallel real-time stereo vision system for generic obstacle and lane detection. IEEE Trans. on Image Processing 7(1), 62–81 (1998)
8. Labayrade, R., Aubert, D., Tarel, J.: Real time obstacle detection in stereovision on non flat road geometry through "V-disparity" representation. In: Proc. IEEE Intelligent Vehicles Symposium, Versailles, France, pp. 646–651. IEEE Computer Society Press, Los Alamitos (2002)
9. Faugeras, O.: Three-Dimensional Computer Vision: a Geometric Viewpoint. MIT Press, Cambridge (1993)
10. Sappa, A., Gerónimo, D., Dornaika, F., López, A.: On-board camera extrinsic parameter estimation. Electonics Letters 42(13) (2006)
11. Blake, A., Isard, M.: Active Contours. Springer, Heidelberg (2000)
12. Doucet, A., Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York (2001)
13. Storn, R., Price, K.: Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. Journal od Global Optimization 11, 341–359 (1997)