On-Board Monocular Vision System Pose Estimation through a Dense Optical Flow*

Naveen Onkarappa and Angel D. Sappa

Computer Vision Center, Edifici O, Campus UAB
08193 Bellaterra, Barcelona, Spain
{naveen,asappa}@cvc.uab.es

Abstract. This paper presents a robust technique for estimating on-board monocular vision system pose. The proposed approach is based on a dense optical flow that is robust against shadows, reflections and illumination changes. A RANSAC based scheme is used to cope with the outliers in the optical flow. The proposed technique is intended to be used in driver assistance systems for applications such as obstacle or pedestrian detection. Experimental results on different scenarios, both from synthetic and real sequences, shows usefulness of the proposed approach.

1 Introduction

During the last decade on-board vision has gained popularity in the automotive applications due to the increase of traffic accidents in modern age. According to the World Health Organization, every year almost 1.2 million people are killed and 50 million are injured in traffic accidents worldwide [1]. A key solution to this is the use of intelligent vision systems that are able to predict dangerous situations and anticipate accidents; these systems are usually referred in the literature as advanced driver assistance systems (ADAS). They help the driver by providing warnings, assisting to take decisions and even taking automatic evasive actions in extreme cases. Some common examples are *lane departure warning*, *pedestrian protection systems* and *adaptive cruise control*.

On-board vision systems can be classified into two different categories: *monocular* or *stereo*. Although each one of them has its own advantages and disadvantages both approaches have a common problem: real-time estimation of on-board vision system pose—position and orientation—, which is a difficult task since: (a) the sensor undergoes motion due to the vehicle dynamics, and (b) the scene is unknown and continuously changing.

In general, monocular based approaches tackle the camera pose problem by using the prior knowledge of the environment as an extra source of information. For instance, Coulombeau and Laurgeau [2] assume that the road observed on

* This work has been partially supported by the Spanish Government under project TRA2007-62526/AUT; research programme Consolider-Ingenio 2010: MIPRCV (CSD2007-00018); and Catalan Government under project CTP 2008ITT 00001.

A. Campilho and M. Kamel (Eds.): ICIAR 2010, Part I, LNCS 6111, pp. 230–239, 2010.

images has a constant known width; Liang et al. [3] assume that the vehicle is driven along two parallel lane markings, which are projected to the left and to the right of the image; Bertozzi et al. [4] assume that the camera's position and orientation remain constant through the time. Obviously the performance of these methods depends on fulfillment of assumptions, which in general cannot be taken for granted.

On the other hand, stereo based approaches have also used prior knowledge of the scene to simplify the problem and to speed up the whole process by reducing the amount of information to be handled. For instance, [5] proposes to reduce the processing time by computing 3D information only on edge points (e.g., lane markings on the image). Similarly, the edge based v-disparity approach proposed in [6], for an automatic estimation of horizon lines and later used for applications such as obstacle or pedestrian detection (e.g., [7],[8]), only computes 3D information over local maxima of the image gradient. A different stereo vision based approach has been proposed in [9]. It uses dense depth maps and is based on the extraction of a dominant 3D plane that is assumed to be the road plane. Camera's position and orientation are directly computed, referred to that plane. A recent work [10] proposes a novel paradigm that is based on raw stereo images provided by a stereo head. This paradigm includes a stochastic technique to track vehicle pose parameters given stereo pairs arriving in a sequential fashion. In [10], the assumption is that the selected region only contains road points, as well as the road surface is assumed to be a plane.

The current work proposes a novel approach for estimating camera's position and orientation for monocular vision systems, which are finally represented as a single value. It is based on a dense optical flow estimated by means of the TV-L^1 formulation. Previous approaches rely on local formulations: a technique based on optical flow with template matching scheme was used in [11], while a maximum likelihood formulation over small patches was introduced in [12].

The main advantage of the proposed approach with respect to other monocular based approaches is that it does not require feature extraction neither imposes restrictive assumptions. The advantage with respect to the previous optical flow based approaches is that the current one is based on an accurate variational dense optical flow formulation. Finally, since it is based on a monocular vision, a system cheaper than stereo based solutions can be reached.

The remainder of this paper is organized as follows. Section 2 briefly introduces the TV-L^1 formulation used to compute dense optical flow, together with the proposed adaptation to reduce the processing time or to increase the accuracy of the flow estimation. The proposed approach is presented in Section 3. Experimental results on different sequences/scenarios are presented in Section 4. Finally, conclusions are given in Section 5.

2 TV-L^1 Optical Flow

State of the art in optical flow techniques unveil that varational techniques give dense estimation with more accuracy as compared to other approaches. TV-L^1

is a variational optical flow technique proposed in [13] that gives dense flow field. In the current work, an improved version [14] is used, which is briefly presented in this section. As the initial formulation of the variational method proposed by Horn and Schunck [15], the formulation in [14] also involves an optical flow constraint and a regularization term but both of them with L^1 norm. The TV-L^1 optical flow is obtained by minimizing the following energy function:

$$E = \int_{\Omega} \{ \alpha \underbrace{|I_1(\boldsymbol{x}+\boldsymbol{u}(\boldsymbol{x})) - I_0(\boldsymbol{x})|}_{Data\ Term} + \underbrace{|\nabla \boldsymbol{u}|}_{Regularization} \} \, d\boldsymbol{x}, \tag{1}$$

where I_0 and I_1 are two images; $\boldsymbol{x} = (x_1, x_2)$ is the pixel location within a rectangular image domain $\Omega \subseteq \mathbf{R}^2$; and $\boldsymbol{u} = (u_1(\boldsymbol{x}), u_2(\boldsymbol{x}))$ is the two dimensional displacement field. The parameter α weighs between data term and regularization term. The objective is to find the displacement field $\boldsymbol{u}$ that minimizes the energy function in (1). The regularization term $|\nabla \boldsymbol{u}|$ with L^1 norm is called total variation regularization. Replacing these data and regularization terms with L^2 norm lead us to the original Horn and Schunck formulation [15]. Since the terms in (1) are not continuously differentiable, the energy function can be minimized using dual formulation for minimizing total variation as proposed in [16] and adapted to optical flow in [13]. Linearizing I_1 near to $(\boldsymbol{x} + \boldsymbol{u}_0)$, where $\boldsymbol{u}_0$ is a given flow field, the whole data term is denoted as an image residual $\rho(\boldsymbol{u}) = I_1(\boldsymbol{x} + \boldsymbol{u}_0) + \langle \nabla I_1, \boldsymbol{u} - \boldsymbol{u}_0 \rangle - I_0(\boldsymbol{x})$. Then, by introducing an auxiliary variable $\boldsymbol{v}$, the data term and regularization term in (1) can be rewritten as indicated in (2), making easier the minimization process. Without loss of generality, in the two-dimensional case, the resulting energy can be expressed as:

$$E = \int_{\Omega} \left\{ \alpha \, |\rho(\boldsymbol{v})| + \sum_{d=1,2} (1/2\theta)(u_d - v_d)^2 + \sum_{d=1,2} |\nabla u_d| \right\} d\boldsymbol{x}, \tag{2}$$

where θ is a small constant, such that $\boldsymbol{v}$ is a close approximation of $\boldsymbol{u}$; and d indicating the dimension takes value as 1 and 2. This convex energy function is optimized by alternative updating steps 1 and 2 for $\boldsymbol{u}$ and $\boldsymbol{v}$:

Step 1. By keeping $\boldsymbol{u}$ fixed, $\boldsymbol{v}$ is computed as:

$$\min_{\boldsymbol{v}} \{\alpha \, |\rho(\boldsymbol{v})| + \sum_{d=1,2} (1/2\theta) \, (u_d - v_d)^2\}, \tag{3}$$

Step 2. Then, by keeping v_d fixed for every d, u_d is computed as:

$$\min_{u_d} \int_{\Omega} \{1/2\theta (u_d - v_d)^2 + |\nabla u_d|\} d\boldsymbol{x}. \tag{4}$$

Equation (4) can be solved for each dimension using the dual formulation. The solution is given by:

$$u_d = v_d - \theta \, \mathbf{div} \boldsymbol{p}_d, \tag{5}$$

where the dual variable $\boldsymbol{p} = [p_1, p_2]$ for a dimension d is iteratively defined by

$$\tilde{\boldsymbol{p}}^{n+1} = \boldsymbol{p} + \tau/\theta(\nabla(v_d + \theta \, \mathbf{div} \boldsymbol{p}^n)), \tag{6}$$

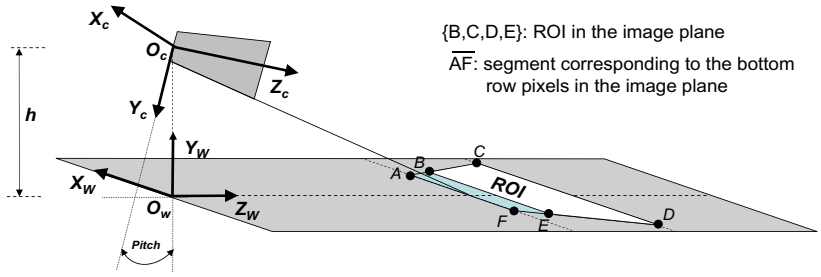

Fig. 1. Camera coordinate system (X_C, Y_C, Z_C) and world coordinate system (X_W, Y_W, Z_W)

$$\boldsymbol{p}^{n+1} = \tilde{\boldsymbol{p}}^{n+1} / \max(1,\ |\tilde{\boldsymbol{p}}^{n+1}|), \tag{7}$$

where $\boldsymbol{p}^0 = \mathbf{0}$ and the time step $\tau \leq 1/4$.

The solution of equation (3) is a simple thresholding step since it does not involve derivative of $\boldsymbol{v}$, and is given by:

$$\boldsymbol{v} = \boldsymbol{u} + \begin{cases} \alpha\theta\nabla I_1 & \text{if} \quad \rho(\boldsymbol{u}) < -\alpha\theta|\nabla I_1|^2 \\ -\alpha\theta\nabla I_1 & \text{if} \quad \rho(\boldsymbol{u}) > \alpha\theta|\nabla I_1|^2 \\ -\rho(\boldsymbol{u})\nabla I_1/|\nabla I_1|^2 & \text{if} \quad |\rho(\boldsymbol{u})| \leq \alpha\theta|\nabla I_1|^2 \end{cases} \tag{8}$$

In this optical flow method, the structure-texture blended image that is robust against sensor noise, illumination changes, reflections and shadows as explained in [14] is used. Additionally, in the current implementation an initialization step is proposed for reducing the CPU time or increasing the accuracy. This step consists in using the optical flow computed between the previous couple of frames as initial values for the current couple instead of initializing by zero.

3 Proposed Approach

Before detailing the approach proposed to estimate the monocular vision system pose, the relationships between the coordinate systems (world and camera) and the camera parameters, assuming a flat road are presented.

3.1 Model Formulation

Camera pose parameters are computed relative to a world coordinate system (X_W, Y_W, Z_W), defined for every frame, in such a way that: the $X_W Z_W$ plane is co-planar with the current road plane. Figure 1 depicts the camera coordinate system (X_C, Y_C, Z_C) referred to the road plane. The origin of the camera coordinate system O_C is contained in the Y_W axis—it implies a $(0, t_y, 0)$ translation of the camera w.r.t. world coordinate system. Hence, since yaw angle is not considered in the current work (i.e., it is assumed to be zero), the six camera pose parameters[1] $(t_x, t_y, t_z, yaw, roll, pitch)$ reduce to just three $(0, t_y, 0, 0, roll, pitch)$,

[1] A 3D translation and a 3D rotation that relates O_C with O_W.

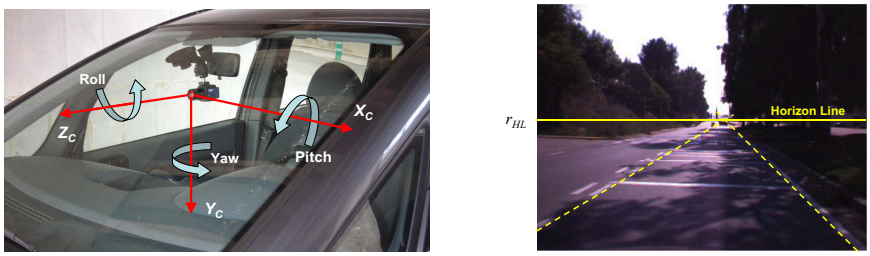

Fig. 2. (*left*) On-board camera with its corresponding coordinate system. (*right*) Horizon line (r_{HL}) estimated by the intersection of projected lane markings.

denoted in the following as (h, Φ, Θ) (i.e., camera height, roll and pitch). Figure $2(left)$ shows the onboard camera used for testing the proposed approach.

Among the parameters (h, Φ, Θ), the value of the roll angle (Φ) will be very close to zero in most situations, since when the camera is rigidly mounted on the car, a specific procedure is followed to ensure an angle at rest within a given range, ideally zero, and in regular driving conditions this value scarcely varies (more details can be found in [9]). Finally, the variables (h, Θ) that represents the camera pose parameters are encoded as a single value, which is the *horizon line* position in the image plane (e.g., [17],[18]). The horizon line corresponds to the back-projection of a point, lying over the road at an infinite depth. Assuming the road can be modelled as a plane, let $ax + by + cz + h = 0$ be the road plane equation and h the camera height, see Fig. 1 (since $(h \neq 0)$ the plane equation can be simplified dividing by $(-h)$). Let $P_i(0, y, z)$ be a point lying over the road plane at an infinite depth z from the camera reference frame with $x = 0$; from the plane equation the y_i coordinates of P_i corresponds to $y_i = \frac{1-cz_i}{b}$. The backprojection of y_i into the image plane when $z_i \to \infty$ defines the row coordinate of the horizon line r_{HL} in the image. It results into:

$$r_{HL} = r_0 + f\frac{y_i}{z_i} = r_0 + \frac{f}{z_i b} - f\frac{c}{b} \ , \tag{9}$$

where f denotes the focal length in pixels, r_0 represents the vertical coordinate of the camera principal point, and z_i is the depth value of P_i. Since $(z_i \to \infty)$, the row coordinate of the horizon line in the image is finally computed as $r_{HL} = r_0 - f\frac{c}{b}$. Additionally, when lane markings are present in the scene, the horizon line position in the image plane can be easily obtained by finding the intersection of these two parallel lines, see Fig. $2(right)$.

3.2 Horizon Line Estimation

In the current work, a RANSAC based approach is proposed to estimate the horizon line position. It works directly in the image plane by using the optical flow vectors computed between two consecutive frames. The TV-L^1 optical flow [14] with a minor modification as explained in the previous section is used. The flow vectors within a rectangular region centered in the bottom part of the

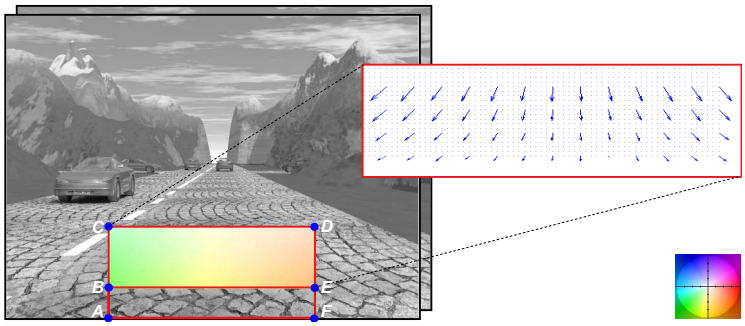

Fig. 3. A couple of consecutive synthetic frames illustrating the rectangular free space {A,C,D,F}, containing the ROI {B,C,D,E} from which computed flow vectors are used for estimating horizon line position. (*top* − *right*) Enlarged and sub-sampled vector field from the ROI. (*bottom* − *right*) Color map used for depicting the vector field in the ROI.

image are used instead of considering the flow vectors through the whole image. The specified region is a rough estimation of the minimum free space needed for a vehicle moving at 30km/h to avoid collisions—rectangle defined by the points $\{A, C, D, F\}$, in Fig. 3. Note that, at a higher speed this region should be enlarged. Actually from this rectangular free space only the top part is used (rectangular ROI defined by the points $\{B, C, D, E\}$ in Fig. 3), since the flow vectors at the bottom part (image boundary) may not be as accurate as required. Figure 3 presents a couple of synthetic frames with the optical flow computed over that ROI; an enlarged and sub-sampled illustration of these flow vectors is given in the top-right part.

Let $\boldsymbol{u}$ be the computed flow field corresponding to a given ROI $\{B, C, D, E\}$. This vector field can be used for recovering the camera motion parameters through a closed form formulation (e.g., [11] and [12]). However, since it could be noisy and contains outliers, a robust RANSAC based technique [19] is proposed for computing the horizon line position. It works as follow:

Random sampling: *Repeat the following three steps K times*

1. Draw a couple of vectors, $(\boldsymbol{u}^1, \boldsymbol{u}^2)$ from the given ROI where $\boldsymbol{u}^1 = (u_1^1, u_2^1)$ and $\boldsymbol{u}^2 = (u_1^2, u_2^2)$.
2. Compute the point (S_x, S_y) where these two vectors intersect.
3. Vote into the cell $\boldsymbol{C}_{(i,j)}$, where $i = \lfloor S_y \rfloor$ and $j = \lfloor S_x \rfloor$ and (i, j) lie within the image boundary.

Solution:

1. Choose the cell that has the highest number of votes in the voting matrix $\boldsymbol{C}$. Let $\boldsymbol{C}_{(i,j)}$ be this solution.
2. Set the sought horizon line position r_{HL} as the row i.

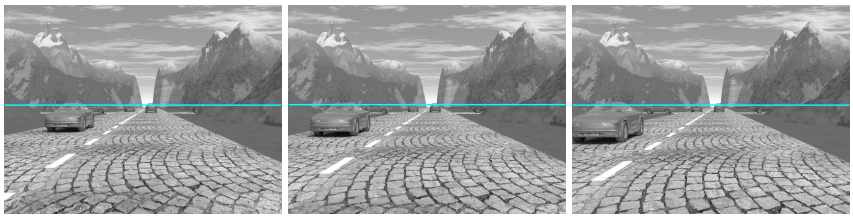

Fig. 4. Horizon line computed by the proposed approach on a synthetic sequence

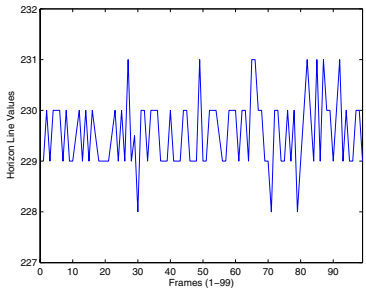

Fig. 5. Plot of variations in horizon line in a sequence of 100 frames

4 Experimental Results

The proposed technique has been tested on several synthetic and real video sequences. Firstly, a synthetic sequence (gray scale sequence-1 in set 2 of *enpeda* [20]) was used for validating the proposed approach. Figure 4 shows some frames with the horizon line computed by the proposed technique. Note, that in this case, since a perfect flat road without any vehicle dynamics, camera pose almost remains constant (horizon line variation through this synthetic sequence is presented in Fig. 5). On the contrary, horizon line undergoes large variations in Fig. 6. This synthetic sequence (gray scale sequence-2 in Set 2 of *enpeda* [20]) contains uphill, downhill and flat road scenarios. Figure 7(*left*) presents the variations of horizon line for the whole sequence. Figure 7(*right*) depicts the pitch angle variation from the ground-truth data. The similarity between these two plots confirms the effectiveness of the presented approach. The sequences in Fig.4, and Fig.6 are of a resolution of 480×640 pixels, and the ROI contains 96×320 pixels placed above 48 pixels from the bottom of the image.

Figure 8 shows a frame from a real sequence (Intern-On-Bike-left sequence in set 1 of *enpeda* sequences [20]) with the horizon line estimated by the proposed approach. The variation of the horizon line over a set of 25 frames of that sequence is presented in Fig. 8(*right*). Additionally, few different real frames, with horizon line estimated by the proposed approach, are shown in Fig.2(*right*) and Fig. 9. Notice that the horizon lines estimated by intersecting the projected lane markings (dotted lines) also coincide with those obtained by the proposed

Fig. 6. Horizon lines computed by the proposed approach on a synthetic video sequence illustrating different situations: uphill, downhill and flat roads

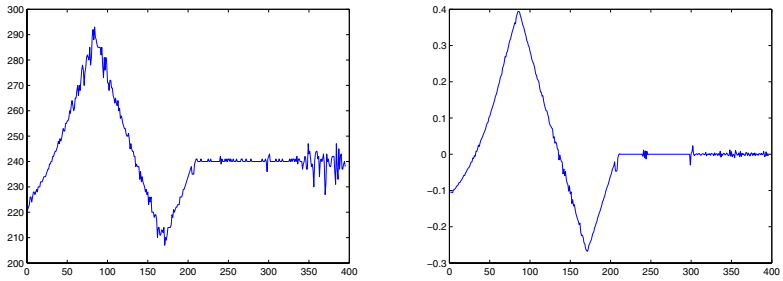

Fig. 7. (*left*) Variations in horizon line position over a sequence of 396 frames. (*right*) Pitch angle variations from the ground-truth.

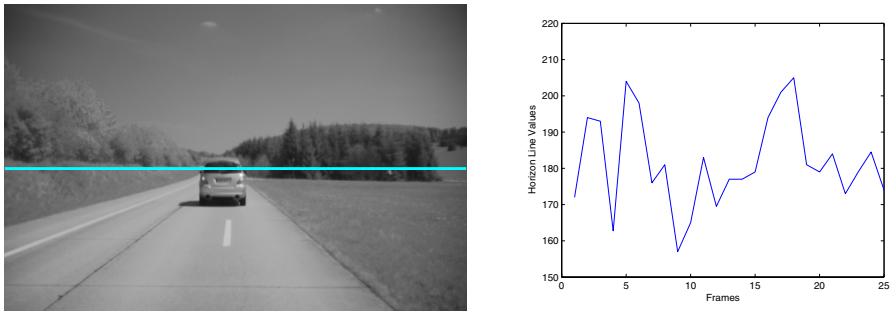

Fig. 8. Horizon line for a real sequence and its variations for 25 frames

approach, in spite of the fact that some frames contain outliers (see lane barriers in the top-left frame in Fig. 9). The video frames in Fig. 9 are captured at a resolution of 480×752 pixels at about 30fps. The value of K is empirically determined and the better value is about half of the total number of flow vectors in the specified ROI. The specified ROI contains 96×376 pixels and is placed above 48 pixels from the bottom of the image.

Fig. 9. Real video frames with the horizon lines estimated by the proposed approach (note that they correspond with the intersections of the projected lane markings)

5 Conclusions

A robust technique for pose estimation of an on-board monocular vision system has been presented. It uses dense flow field from a state of the art variational optical flow technique that is robust against common obstructions in real traffic such as shadows, reflections and illumination changes. The proposed modified initialization step to the optical flow estimation has the advantage to be more accurate or less computation time. The camera pose parameters estimation is modelled as a horizon line estimation problem and has been solved using a RANSAC based approach that is robust against outliers in the flow field. The proposed approach is validated on both synthetic and real sequences. With the advance in the real-time implementation of optical flow algorithms and particularly, for our problem of estimating the flow vectors only in the specified region instead of the whole image, the proposed approach can be implemented on real applications with real-time performance.

References

1. Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A., Jarawan, E., Mathers, C.: World Report on road traffic injury prevention. World Health Organization, Geneva (2004)
2. 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 (2002)

3. Liang, Y., Tyan, H., Liao, H., Chen, S.: Stabilizing image sequences taken by the camcorder mounted on a moving vehicle. In: Proc. IEEE Int. Conf. on Intelligent Transportation Systems, Shangai, China, pp. 90–95 (2003)
4. Bertozzi, M., Broggi, A., Carletti, M., Fascioli, A., Graf, T., Grisleri, P., Meinecke, M.: IR pedestrian detection for advaned driver assistance systems. In: Proc. 25th. Pattern Recognition Symposium, Magdeburg, Germany, pp. 582–590 (2003)
5. Nedevschi, S., Vancea, C., Marita, T., Graf, T.: Online extrinsic parameters calibration for stereovision systems used in far-range detection vehicle applications. IEEE Trans. on Intelligent Transportation Systems 8(4), 651–660 (2007)
6. 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 (2002)
7. Bertozzi, M., Binelli, E., Broggi, A., Del Rose, M.: Stereo vision-based approaches for pedestrian detection. In: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, San Diego, USA (2005)
8. Labayrade, R., Aubert, D.: A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereovision. In: Proc. IEEE Intelligent Vehicles Symposium, Columbus, OH, USA, pp. 31–36 (2003)
9. Sappa, A., Dornaika, F., Ponsa, D., Gerónimo, D., López, A.: An efficient approach to on-board stereo vision system pose estimation. IEEE Trans. on Intelligent Transportation Systems 9(3), 476–490 (2008)
10. Dornaika, F., Sappa, A.: A featureless and stochastic approach to on-board stereo vision system pose. Image and Vision Computing 27(9), 1382–1393 (2009)
11. Suzuki, T., Kanade, T.: Measurement of vehicle motion and orientation using optical flow. In: Proc. IEEE Int. Conf. on Intelligent Transportation Systems, Tokyo, Japan, pp. 25–30 (1999)
12. Stein, G., Mano, O., Shashua, A.: A robust method for computing vehicle egomotion. In: IEEE Intelligent Vehicles Symposium, Dearborn Michigan, USA, pp. 362–368 (2000)
13. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L^1 optical flow. In: Proc. 29th Annual Symposium of the German Association for Pattern Recognition, Heidelberg, Germany, pp. 214–223 (2007)
14. Wedel, A., Pock, T., Zach, C., Cremers, D., Bischof, H.: An improved algorithm for TV-L1 optical flow. In: Proc. of the Dagstuhl Motion Workshop, Dagstuhl Castle, Germany, pp. 23–45 (2008)
15. Horn, B.K.P., Schunk, B.G.: Determining optical flow. Artificial Intelligence 17, 185–203 (1981)
16. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. 20(1-2), 89–97 (2004)
17. Zhaoxue, C., Pengfei, S.: Efficient method for camera calibration in traffic scenes. Electronics Letters 40(6), 368–369 (2004)
18. Rasmussen, C.: Grouping dominant orientations for ill-structured road following. In: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Washington, USA, pp. 470–477 (2004)
19. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Graphics and Image Processing 24(6), 381–395 (1981)
20. Vaudrey, T., Rabe, C., Klette, R., Milburn, J.: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In: Proc. Image and Vision Computing New Zealand, Christchurch, New Zealand, pp. 1–6 (2008)