Autonomous Robots 19, 141–157, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

A Stereovision Method for Obstacle Detection and Tracking in Non-Flat Urban Environments

QIAN YU
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China
yuqian97@mails.tsinghua.edu.cn

HELDER ARAÚJO
Department of Elect. and Computer Eng.–Polo II, Institute of Systems and Robotics, University of Coimbra, 3030-290 Coimbra, Portugal
helder@isr.uc.pt

HONG WANG
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China
wanghong@mail.tsinghua.edu.cn

Abstract. Obstacle detection is an essential capability for the safe guidance of autonomous vehicles, especially in urban environments. This paper presents an efficient method to integrate spatial and temporal constraints for detecting and tracking obstacles in urban environments. In order to enhance the reliability of the obstacle detection task, we do not consider the urban roads as rigid planes, but as quasi-planes, whose normal vectors have orientation constraints. Under this flexible road model, we propose a fast, robust stereovision based obstacle detection method. A watershed transformation is employed for obstacle segmentation in dense traffic conditions, even with partial occlusions, in urban environments. Finally a UKF (Unscented Kalman filter) is applied to estimate the obstacles parameters under a nonlinear observation model. To avoid the difficulty of the computation in metric space, the whole detection process is performed in the disparity image. Various experimental results are presented, showing the advantages of this method.

Keywords: obstacle detection, quasi-plane road assumption, stereo vision, plane normal, Unscented Kalman filter tracking

1. Introduction

Obstacle detection and tracking for autonomous vehicle navigation has been extensively studied. Most of the obstacle detection systems are based on stereo vision (Bertozzi et al., 1997, 1998, 2000; Bertozzi and Broggi, 1998) and use an inverse perspective model applied to a planar model (static or dynamic) of the road. Optical flow based approaches (Enkelmann, 1997) and correlation based stereo systems (Saneyoshi, 1994) have also been used. A system developed at Daimler-benz Research also uses stereo vision (Franke et al., 1999). At Carnegie-Mellon University considerable work on obstacle detection was also done within the framework of Navlab (Williamson, 1998; Thorpe et al., 1998).

Several difficulties arise from the planar assumption. First high-speed operation requires a strict real time performance and also a large range of distances. These requirements imply that the use of a strictly planar model of the road is inadequate. Second many roads in urban environments are not completely planar in a small range of distances, often with hills and valleys

(Labayrade et al., 2002), and even with a slight curvature across the road. Moreover typical urban roads lack texture and reliable landmarks. The existence of texture and/or patterns is essential for the dynamic estimation of the plane parameters. Small errors in the plane parameters severely affect obstacle detection.

To improve the reliability of obstacle detection, we first introduce a quasi-plane model assumption and propose a fast and robust approach, for obstacle detection without depending on the planar road assumption. This method is able to cope with the common road gradients in urban environments and does not require any rigid restriction on the placement of the stereo head. In order to estimate the parameters of the ground plane offline, a RANSAC plane-fitting algorithm is employed.

In order to detect obstacles in clustered traffic conditions, even with partial occlusion, a watershed transformation is applied to segment isolated obstacles. The last step is to estimate the obstacle parameters based on the detected obstacle regions in image plane and the object dynamics. The UKF (Julier and Uhlmann, 1997) is used to approximate the nonlinear system up to the third order for Gaussian distributions, while usually EKF is a first order approximation. In UKF, a small number of carefully chosen sample points are propagated in each estimation step, which provides a compact parameterization of the underlying distribution in contrast to the random sampling methods. To avoid the difficulties of dealing with the metric space, the whole process is performed in the disparity image. An overview flowchart of the system is shown in Fig. 1.

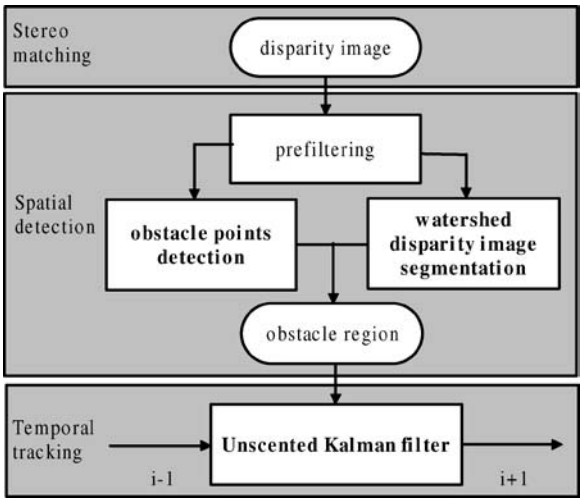

Figure 1. The architecture of the algorithm.

The rest of this paper is organized as follows: Section 2 presents the quasi-planar assumption and details the spatial detection approach to identify and segment isolated obstacles. Section 3 introduces a UKF to track obstacles. Section 4 gives experimental results and describes the performance; finally Section 5 contains the conclusions and the description of future work.

2. Spatial Obstacle Detection

2.1. The Coordinate Systems Used

In the remaining of this paper we will use the two coordinate systems shown in Fig. 2: R_w (world) and R_c (left camera). The angle between the optic axis and the road plane is α; the angle between the X_c axis (which coincides with the baseline) and the X_w axis is β. In the left camera coordinate system, the image coordinates (using a pinhole camera model) can be expressed as in Eq. (1). In a stereo system, we can make the approximation that the pixels are square and therefore the scale factors t_u and t_v are equal: $t = t_u = t_v$. Then if f is the focal length in pixels: $f = f'/t$, where f' is focal length in metric units.

$$u = f\frac{X}{Z}$$
$$v = f\frac{Y}{Z} \tag{1}$$

2.2. Stereo Analysis

For an efficient stereo analysis, it is preferable that the epipolar lines are parallel to the scan lines of the camera. This configuration is obtained with both cameras having the same focal length and with parallel image

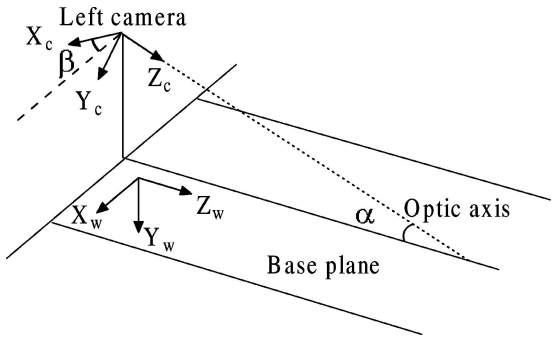

Figure 2. The coordinate systems used.

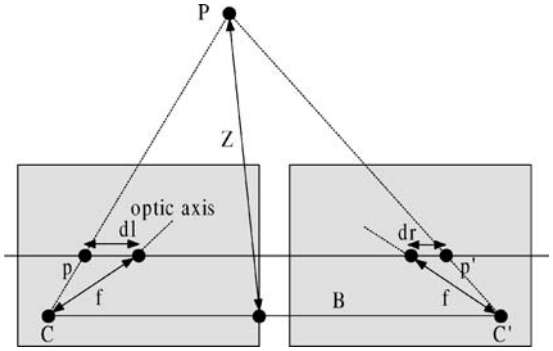

Figure 3. Relationship between disparity and depth.

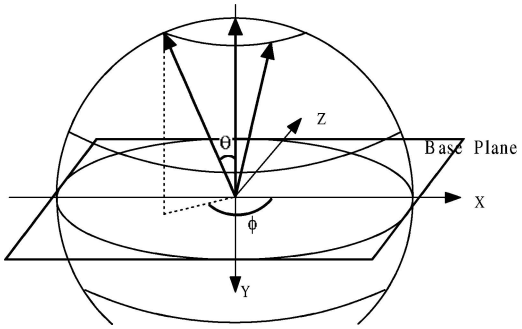

Figure 4. The field of the unit normal in world coordinates.

planes. In these conditions, stereo matching can be performed along the scan lines.

The offset of the image location of an object in the left image and in the right image is called disparity. The disparity is directly related to the distance of the object from the cameras (measured along the normal to the image planes). With the stereo baseline B and the focus in pixels f, the relationship can be represented as follows, where Z, d represent depth and disparity respectively (Fig. 3):

$$Z = \frac{Bf}{d}, \quad \text{where } d = dl - dr \tag{2}$$

To acquire dense depth information, we employ an area correlation stereo matching method proposed by Konolige (1997), which correlates not raw intensity image, but the $L1$ norm (absolute difference) of *LOG* transformation. A left/right check is introduced as a filtering to eliminate bad matches.

2.3. *Quasi-Planar Scene Assumption*

Many roads in urban environments are not completely planar, often with longitudinal (along the road) curvatures, even with slight latitudinal (across the road) curvatures. Instead of a rigid plane, we model the road as a smooth surface, so called a quasi-plane, which is a plane in general and whose normal vector is constrained within a range around the normal of the base plane. Under this assumption, we define obstacles as objects above the road surface, whose normal vector is almost perpendicular to that of local road surface. According to the quasi-planar scene assumption, the unit normal vector of the road surface (a_w, b_w, c_w) can be represented in spherical coordinates as follows (see Fig. 4):

$$\{N(R, \theta, \phi) \mid R = 1, 0 < \theta < \theta_0\} \tag{3}$$

where θ is the latitude angle relative to the negative Y_w axis. The unit normal vector of the obstacle surface can be represented as follows:

$$\left\{N(R, \theta, \phi) \mid R = 1, \frac{\pi}{2} - \theta_1 < \theta < \frac{\pi}{2} + \theta_1\right\} \tag{4}$$

θ_0 and θ_1 are the predetermined thresholds for the road surface and for the obstacle surface, which depend on the road conditions.

2.4. *Obstacle Detection*

The various expressions of the normal in the different coordinate systems are the following:

(a, b, c)	normal in left camera coordinates;
(a_c, b_c, c_c)	unit normal in left camera coordinates;
(a_w, b_w, c_w)	unit normal in world coordinates;

In the left camera coordinate system, the equation of a plane can be expressed as in Eq. (5), since in our application we do not have to deal with planes passing through the origin, (the left optical center).

$$aX + bY + cZ = 1 \tag{5}$$

On the basis of Eqs. (1), (3) and (5), we obtain the equation of a plane in disparity space (d, u, v) as follows:

$$d = B(au + bv) + Bfc \tag{6}$$

From Eq. (6), a plane in 3D space is also a plane in disparity space. Note that the partial derivatives of d with respect to u and v are functions of the first two components (a, b) of the plane normal:

$$\left(\frac{\partial d}{\partial u}, \frac{\partial d}{\partial v}\right) = B(a, b) \tag{7}$$

Using the assumption of a quasi-planar scene it will be shown that the knowledge of both a and b is sufficient to discriminate the road surface from the obstacles. This also avoids the estimation of the third component of the normal vector.

Due to the rotation between the left camera and world coordinate system, the relation between the unit normal in left camera coordinates $(a_c, b_c, c_c,)$ and in world coordinates can be represented as follows:

$$[a_c, b_c, c_c]^T = R(\beta)R(\alpha)\,[a_w, b_w, c_w]^T \tag{8}$$

where

$$R(\beta) = \begin{bmatrix} \cos\beta & -\sin\beta & 0 \\ \sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$

The first two components of the normal vector, namely (a_c, b_c) are shown in Fig. 5(b), which can be also regarded as the projection of the normal vector (a_c, b_c, c_c) into the image plane, i.e. the $X_c - Y_c$ plane.

With the knowledge of the angle $\angle TON$ denoted by θ_0' and the length of OT (see Fig. 5(b)), we can find out whether a point belongs to an obstacle or to the road surface. In approximate terms, if the angle between a vector and OT is larger than θ_0', or the length of a vector is less than length of OT, the point associated with the vector can be considered as belonging to an obstacle; otherwise it can be considered as belonging to the road surface. The OT and θ_0' can be represented respectively as follows (see the Appendix):

$$\tan\theta_0' = \left((\cos\alpha/\tan\theta_0)^2 - (\sin\alpha)^2\right)^{-1/2} \tag{9}$$

$$OT = \cos\alpha\cos\theta_0 - \sin\alpha\sin\theta_0 \tag{10}$$

From Fig. 5, the obstacle detection task can be regarded as a procedure to classify obstacle points and road surface points. The region *MTN* corresponds to points on the road surface and the region *RSQP* corresponds to the obstacle surface.

If (Eq. (4)) $\theta_1 < \alpha$, T' is below the X_c axis. The application of conditions defined in Eq. (3) and in Eq. (4) is simpler since the direction of the normal is sufficient to discriminate the regions of the road surface from the obstacle region.

We introduce a depth radio, defined in Eq. (11), to scale the magnitude of the vector (a_c, b_c), where Z is the depth of point P on the obstacle or road surface; dis is the distance between the optical center and the surface.

$$ratio = \frac{Z}{dis} \tag{11}$$

In the sector *OMTN*, the imaging process can be represented approximately as in Fig. 6. H is the height

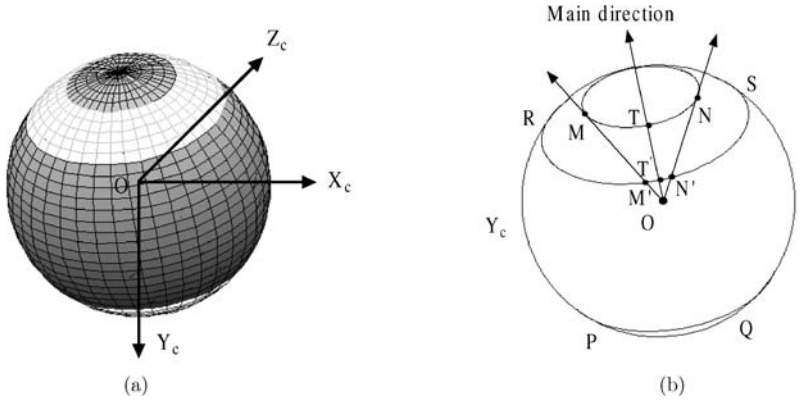

Figure 5. Classification in the image plane. (a) The unit normal vector in left camera coordinates, and (b) The first two components of the normal vector.

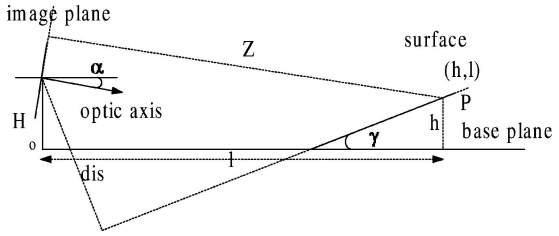

Figure 6. Illustration of depth ratio.

of the camera relative to the base plane; h the height of point P relative to the base plane and l is the distance between point P and the optical center measured on the base plane. γ is the angle between the surface and the road plane; α is the angle between the optical axis and the base plane. The depth ratio can be expressed in terms of these parameters as follows:

$$ratio = \left| \frac{\left(\frac{h-H}{l} - \tan^{-1}\alpha\right)(1+\tan^2\gamma)^{1/2}}{\left(\frac{H-h}{l} + \tan\gamma\right)(1+\tan^{-2}\alpha)^{1/2}} \right| \quad (12)$$

Since $(h - H)$ is usually much smaller than l, $(h - H)/l$ is negligible. Thus we get

$$ratio = \frac{\cos\alpha}{\sin\gamma} \quad (13)$$

Therefore the depth ratio for regions of the road surface will be larger than the depth ratio for the obstacle surface. For that reason vector (a_c, b_c) is scaled by the depth ratio defined in Eq. (11). As a result, the difference between the normal vector corresponding to the road surface region *OMTN* (Fig. 5) and the normal vector corresponding to the obstacles surface region $OM'N'$ (Fig. 5) is increased.

Corresponding to a planar patch defined as in Eq. (5), the relationship between the unit normal vector (a_c, b_c, c_c) and the normal (a, b, c) can be expressed as follows:

$$(a_c, b_c, c_c) = (a, b, c)/(a^2 + b^2 + c^2)^{1/2} \quad (14)$$

And the distance of the planar patch to the origin is

$$dis = 1/(a^2 + b^2 + c^2)^{1/2} \quad (15)$$

From Eqs. (2), (1), (14) and (15), we get:

$$ratio(a_c, b_c) = \frac{Bf(a, b)}{d} \quad (16)$$

After taking into account the derivatives of the disparity in the image plane Eq. (7), we acquire the vector field in Eq. (17) for classification. A surface whose direction of the vector is close to that of the main direction, and whose vector magnitude is large enough, will be considered a road surface, otherwise as an obstacle surface.

$$ratio(a_c, b_c) = \frac{f}{d}\left(\frac{\partial d}{\partial u}, \frac{\partial d}{\partial v}\right) \quad (17)$$

2.5. *Discussion of Eq. (17)*

Based on Eqs. (1) and (2), the scaled vector in Eq. (17) can be rewritten as follows,

$$ratio(a_c, b_c) \propto \left(\frac{\partial Z}{\partial X}, \frac{\partial Z}{\partial Y}\right) \quad (18)$$

which reveals the most important cue for obstacle detection under the quasi-planar road assumption.

In order to obtain reliable gradient estimates, we employ an adaptive gradient estimation at different scale levels. Since different disparity values imply different object scales/sizes in the image plane (i.e. an object with larger disparities should have a bigger scale in the disparity image), an adaptive gradient operator is defined as follows. Let $d(x, y)$ denote the disparity image prefiltered by median filters:

$$\begin{aligned} &\nabla(d_k) = \nabla_{\sigma_k}(d_k) \quad \text{where} \\ &d = \bigcup_k d_k; d_k > d_{k-1}, \sigma_k > \sigma_{k-1} \end{aligned} \quad (19)$$

2.6. *RANSAC Plane Fitting*

In order to obtain the main direction of the base plane of a certain configuration, we propose the use of RANSAC (Fischler and Bolles, 1981; Se, 2002), to estimate offline the parameters of the plane in disparity space defined in Eq. (6). The procedure is the following:

1. In disparity space, randomly select three points to fit a plane. Check, for each and all the points (by computing the distance of a point to the plane), whether they satisfy the estimated plane equation and count the number of fitting points.
2. Repeat step (1) m times, select the triple (a, b, c) with maximum support and do least-squares base

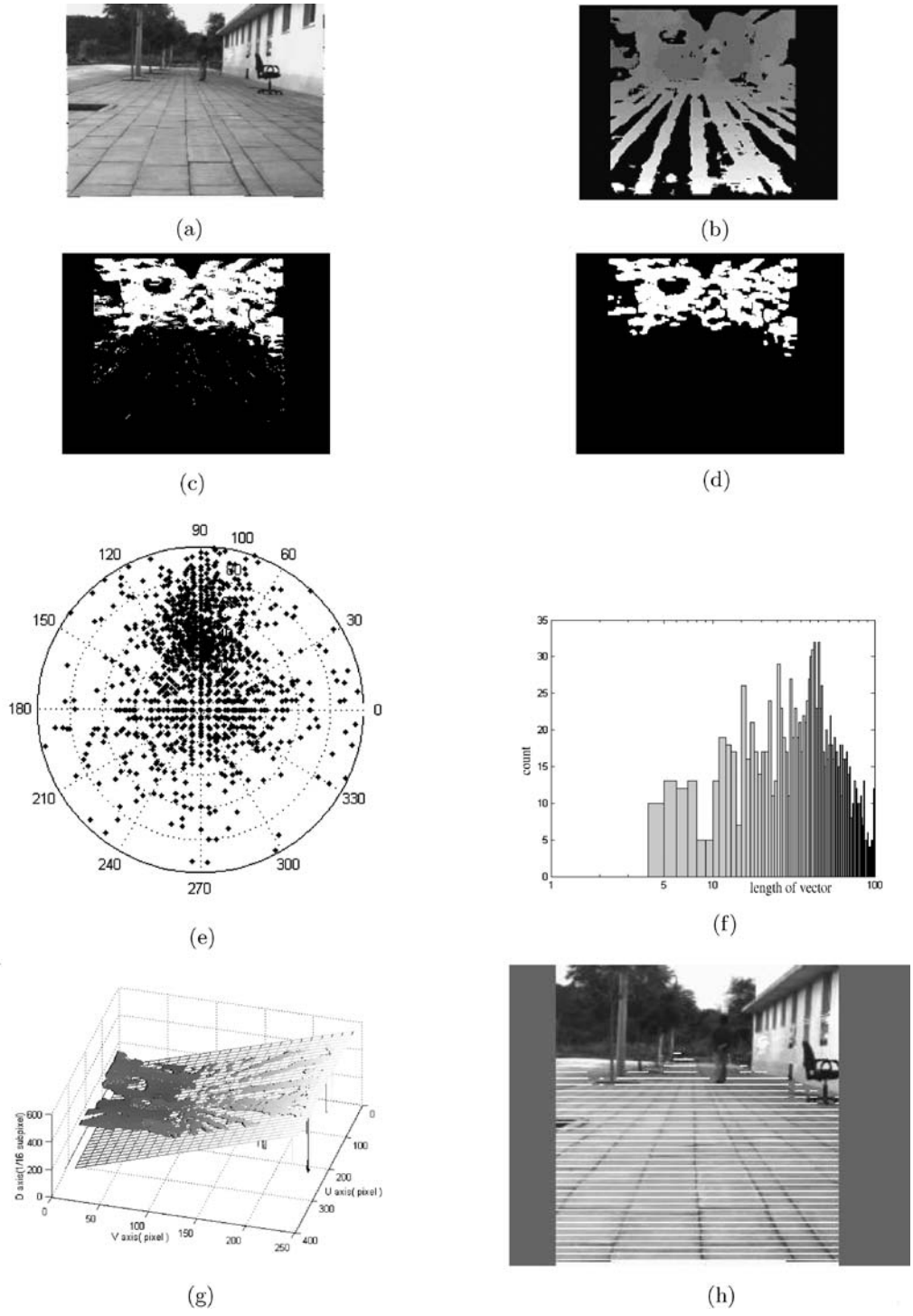

Figure 7. Obstacle detection in the disparity space. (a) The original left image, (b) the disparity image, (c) obstacle points, (d) result after opening, (e) the scaled vector field, (f) the histogram of vector lengths, (g) RANSAC plane-fitting in disparity space, and (h) free space after obstacle detection.

plane fitting to this triple using all its supporting points. Provided that there are sufficient road plane points in the disparity space, the plane thus estimated will correspond to the road plane. Given that the outlier proportion is e, the probability that the algorithm will exit without finding a good fit is p_f, and the sample size is N, the number of samples L can be represented as Eq. (20).

$$L = \frac{\log(p_f)}{\log(1-(1-e)^N)} \tag{20}$$

3. Assuming that the percentage of contamination for an offline scene is 75%, a probability of 99% for a good sampling requires a number of approximately 300 samples.

2.7. Comparison of the Detection Results

Assuming that the ground is planar, obstacles are usually defined as objects rising out from the ground plane. The homography-based method, which is popular for obstacle detection, is applied to warp the left image into the right image. Comparing the right image with the warped left image, points in non-free-space (i.e. obstacles area) will show disparities due to the obstacles' heights, while those on the free-space will not.

Let H denote the homography matrix between two image planes with respect to the ground plane. For a point P on the ground plane, the homography relates its projection p_l in left image to its projection p_r in the right image as follows:

$$p_\mathrm{r} \cong H p_\mathrm{l} \tag{21}$$

where $\cong$ represents equality up to a scale factor.

Many methods have been proposed to estimate the homography matrix. Since there are 8 unknowns in $H_{3\times3}$, at least four non-collinear correspondences are needed to compute $H_{3\times3}$. Although four correspondences are enough for a solution, in practice tens of correspondences are used to reduce the impact of noise with a least-squares method. If the parameters of the ground plane can be estimated (for example by the RANSAC algorithm proposed above), the homography matrix can be estimated as follows:

$$H = A_\mathrm{l}(R + tN^T){A_\mathrm{r}}^{-1} \tag{22}$$

where A_l and A_r are 3×3 the intrinsic matrices of the two cameras; R, t are the rotation matrix and the translation vector between the two camera coordinates and n is the unit normal of the plane.

Furthermore, and independently of the type of homography based method, updating the homography matrix is important for robust obstacle detection in dynamic scenes. Unfortunately typical urban roads lack texture and reliable landmarks (as the scene in Fig. 7). The existence of texture and/or patterns is essential for the dynamic estimation of the parameters related to the road plane.

In addition, and as a result of lack of information, it is difficult to cluster the obstacle points in the case of the homography-based methods, especially in dense traffic conditions (Figs. 8 and 9).

2.8. Obstacle Segmentation

To extract high-level attributes of the obstacles (such as width, height, etc.) in clustered traffic conditions, we introduce a watershed-based algorithm to identify and label isolated obstacle regions in the disparity image. Watershed transformation is known as a powerful tool for image segmentation (Vincent and Soille, 1991). Usually it is performed on the gradient of the image to be segmented. A proper definition of the gradient is crucial to the segmentation result and to the performance of the algorithm. To group points in 3D space, a reasonable gradient operator in the left camera coordinate system is defined as follows:

$$\nabla(\cdot) = \mathrm{Max}\left\{Z_{\pi_{(x,y)}}\right\} - \mathrm{Min}\left\{Z_{\pi_{(x,y)}}\right\} \tag{23}$$

where $\pi_{(x,y)}$ is the unit area in $X-Y$ plane. The gradient operator defined in Eq. (23) estimates the depth range in the $X - Y$ plane. However projecting image pixels from the disparity image to the 3-D world generates a non-uniform point set, which causes difficulties in the gradient computation. Usually a grid based space representation or interpolation method is applied to solve this problem. To avoid the difficulty of estimating in 3D, we directly estimate the gradient in the disparity image. Based on Eqs. (1) and (2), Eq. (23) can be rewritten as follows:

$$\nabla(\cdot) \propto \frac{1}{d}\left(\mathrm{Max}\left\{d_{n(u,v)}\right\} - \mathrm{Min}\left\{d_{n(u,v)}\right\}\right) \tag{24}$$

where $d_{n(u,v)}$ denotes a $n \times n$ square centered at a point(u, v) in the image plane. Equation (24) estimates the gradient by taking into account the disparities that

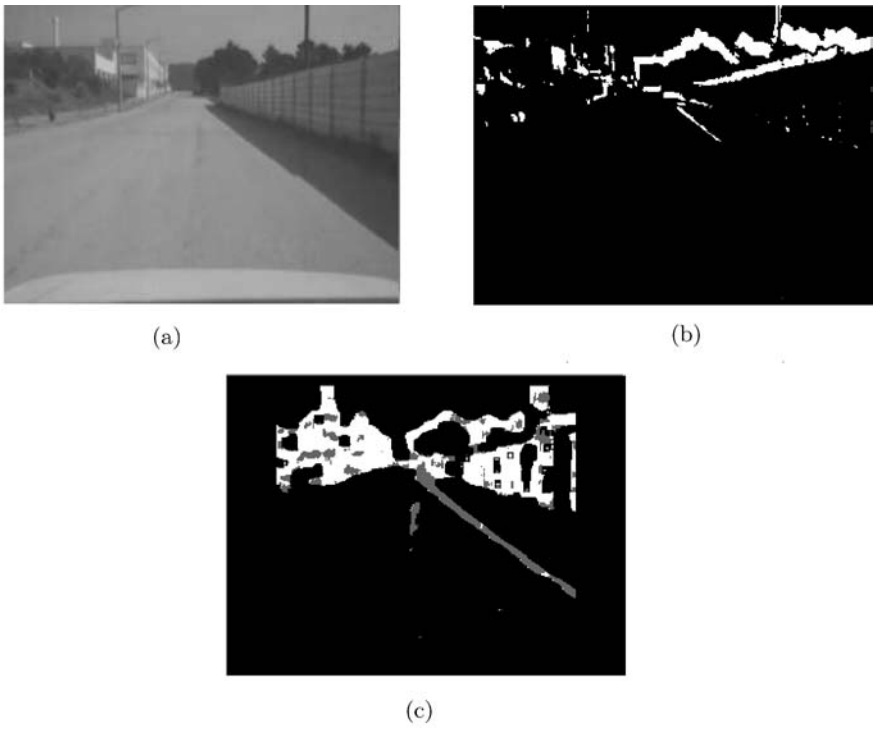

Figure 8. Comparison with results based on the planar assumption. (a) image of a road with a non-uniform slope, (b) obstacle points detected using a planar assumption, and (c) obstacle points detected using the quasi-planar assumption (white points are obstacle points, gray points belong to the road).

Figure 9. Segmentation results in a clustered scene and bounded obstacle objects.

result from the different depths, thus avoiding the unbalance that would result if the disparity gradient were to be directly estimated.

In practice, a morphology gradient operator and a non-linear transformation (Beucher and Bilodeau, 1994) are used to estimate the gradient as follows:

$$\nabla(\cdot) = \log_2\left(\frac{sign(grad(\cdot)+1)}{2} grad(\cdot) + 1\right) \quad (25)$$

where

$$grad(\cdot) = \frac{K_1}{d}\nabla_{\text{morph}}\left(d_{n(u,v)}\right) - K_2$$

$$sign(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

K_1, and K_2 are predetermined scale factors to adjust the gradient values into a range of integers. Moreover by adjusting K_2, we can remove small gradient values to avoid over-segmentation. The non-linear transformation leaves the low values practically unchanged and decreases significantly the high values. This transformation is used to reduce the number of quantization levels of the gradient values, and thus reduce the computational cost.

The procedure for obstacle region segmentation is realized in the following steps:

1. Adaptive multiscale morphological gradient: a multiscale morphological gradient (Wang, 1997) controlled by depth is defined. B_i is a group of square structure elements of dimensions $(2i+1)\times(2i+1)$, where $0 \le i \le 3$. $\oplus$ and $\ominus$ denote the dilation and erosion respectively. The multiscale morphological gradient is expressed as follows:

$$\nabla_{\text{morph}}(\cdot) = \frac{1}{k}\sum_{i=1}^{k}(((d_k \oplus B_i) - (d_k \ominus B_i)) \ominus B_{i-1}) \quad (26)$$

   where $d_k < d_{k-1}; d = \bigcup d_k; k = 1, \ldots, L$
2. A nonlinear transformation is applied to convert the gradient image into an integer image.
3. A watershed transformation is performed in the gradient image after the non-linear transformation.
4. Classify a region according to the ratio between obstacle points and road surface points.
5. Merge obstacle regions using as a criterion the common boundary sharpness; the common boundary sharpness is computed as the average edge magnitude along the common boundary.

3. Temporal Tracking

As a result of the application of the obstacle detection method described, estimates of the positions and of the dimensions (height and width) of the obstacles in left camera coordinates can be obtained. These are, in general, difficult to estimate by other sensors. The dimensional information can be used to identify obstacles, such as pedestrians, automobiles or background buildings. The relevant obstacles can be tracked by using temporal information.

As a result of the application of the detection process some obstacles may be lost or merged with other obstacles at one instant, or one obstacle may be detected as two or more obstacles close to each other. Temporal tracking can be applied to establish a correspondences between obstacles and to filter out the wrong results of the detection process in dynamic scenes. Furthermore, and besides the information relative to the dimensions, motion provides another important cue to identify obstacles qualitatively.

In typical traffic conditions most of the movement is linear. As a result we use a constant velocity motion model as the state transition model. Let $G_w = [x_w, y_w, z_w]^T$ denote the geometry center of an obstacle and W, H denote the width and height in world coordinates. Then the state vector can be written as $S_k = [W, H, G_w, G'_w]^T$. For one time instant, the measurement vector can be represented as $O_k = [w, h, g]^T$, where w, h denote the width and height of the obstacle in the image plane; $g = [u, v, d]^T$ denotes the center of an obstacle in disparity space.

Since tracking is carried out in world coordinates, the vehicle's position information is required. At the time instant k, given the rotation $R(\alpha, \beta, \gamma_k)$ and translation $t(x_k, 0, z_k)$ between the left camera coordinate system and the absolute world coordinate system (supposing that the rotation around X_w, Z_w is constant and the translation along Y_w is negligible), the relationship between the obstacle's position in left camera coordinates $G_c(x_c, y_c, z_c)$ and in world coordinates $G_w(x_w, y_w, z_w)$ can be described as follows:

$$G_c = M_1(G_w) = RG_w + t; \quad (27)$$

The relation between the measurement vector and the center of an obstacle in left camera coordinates can be represented as follows:

$$[w\ h\ u\ v\ d]^T = M_2(W, H, G_c) = \left[f\frac{W}{z_c}\ f\frac{H}{z_c}\ f\frac{x_c}{z_c}\ f\frac{y_c}{z_c}\ f\frac{b}{z_c}\right]^T. \quad (28)$$

Thus the state equation and measurement equation of a Kalman filter can be described as follows:

$$\begin{cases} s_k = D \cdot s_{k-1} + v \\ o_k = M \cdot s_k \quad + n \end{cases} \quad (29)$$

where

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t_k \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M(\cdot) = M_2(W, H, M_1(G_w))$$

The subscript k denotes the time instant. The random variables v and n represent the state error and the measurement error, respectively, which are assumed to be white, independent of each other, and with normal probability distribution. $v = [0, 0, 0, e_{ac}]^T$. The e_{ac} is the noise term used to absorb the error made by the constant velocity assumption and n can be estimated using some off-line sample measurements.

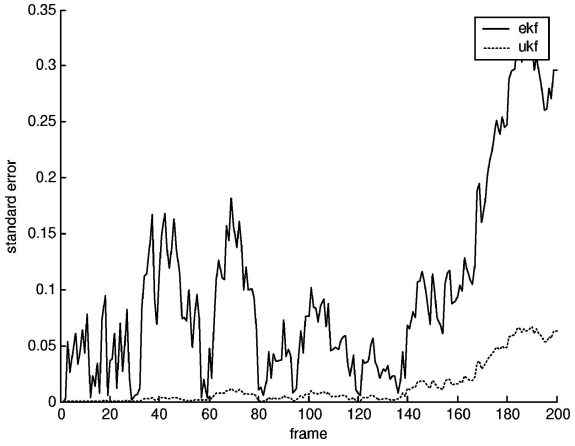

Figure 10. Simulation result of Z between UKF and EKF.

Due to the nonlinear and time variant observation model in Eq. (29), we resort to a nonlinear estimation technique, the UKF, which was proposed in Julier and Uhlmann (1997) and provably superior to the traditional EKF (Fig. 10). In UKF, a minimal set of carefully chosen sample points, which capture the mean and covariance of Gaussian random variable, propagate through the true nonlinear system instead of the linear approximation. Furthermore for any nonlinear system calculation of *Jacobian* is not required in this algorithm. The UKF can be summarized as follows:

1. Initialize with $k = 1$:

$$\begin{aligned}
\hat{s}_{k-1} &= E[s_{k-1}] \\
P_{k-1} &= E[(s_{k-1} - \hat{s}_{k-1})(s_{k-1} - \hat{s}_{k-1})^T] \\
\hat{s}^a_{k-1} &= E\big[s^a_{k-1}\big] = \big[\hat{s}^T_{k-1}, 0, 0\big]^T \\
P^a_{k-1} &= E\big[\big(s^a_{k-1} - \hat{s}^a_{k-1}\big)\big(s^a_{k-1} - \hat{s}^a_{k-1}\big)^T\big] \\
&= \mathrm{Diag}(P_{k-1}, P_v, P_n)
\end{aligned} \tag{30}$$

   where the $s^a_k = [s^T_k v^T n^T]^T$ is augmented state vector; $P_v =$ process noise covariance, $P_n =$ measure noise covariance.
2. $L =$ dimension of the augmented state s^a_k. We form a matrix $\mathcal{S}^a_{k-1}$ of $2L + 1$ augmented sigma vectors s^a_{k-1} (with corresponding weights W_i), according to the following:

$$\begin{aligned}
\mathcal{S}^a_{0,k-1} &= \hat{s}^a_{k-1} \\
\mathcal{S}^a_{i,k-1} &= \hat{s}^a_{k-1} + \Big(\sqrt{(L+\lambda)P^a_{k-1}}\Big)_i \; i \in [1, L] \\
\mathcal{S}^a_{i,k-1} &= \hat{s}^a_{k-1} - \Big(\sqrt{(L+\lambda)P^a_{k-1}}\Big)_{i-L} i \in [L+1, 2L]
\end{aligned}$$

$$\begin{aligned}
W^{(m)}_0 &= \lambda/(L+\lambda) \\
W^{(c)}_0 &= \lambda/(L+\lambda) + (1 - \alpha^2 + \beta) \\
W^{(m)}_i &= W_i(c) = 1/\{2(L+\lambda)\} \; i \in [1, 2L]
\end{aligned} \tag{31}$$

   where $\mathcal{S}^a = [(\mathcal{S}^s)^T(\mathcal{S}^v)^T(\mathcal{S}^n)^T]^T$. $\lambda = \alpha^2(L+k) - L$ is a scaling parameter. α determines the spread of the sigma points around $\hat{s}^a_{k-1}$ and is usually set to a small positive value. k is a secondary scaling parameter which is usually set to 0, and β is used to incorporate prior knowledge of the distribution (for Gaussian distributions, $\beta = 2$ is optimal (Julier and Uhlmann, 1997)). $\mathcal{S}^a_{i,k-1}$ is the ith row of the $\mathcal{S}^a_{k-1}$. $(\sqrt{(L+\lambda)P^a_{k-1}})_i$ is the ith row of the matrix square root.
3. Time update:

$$\begin{aligned}
\mathcal{S}^x_{k|k-1} &= D\mathcal{S}^s_{k-1} + \mathcal{S}^v_{k-1} \text{ (see Eq. (29))} \\
\hat{s}^-_k &= \sum_{i=0}^{2L} W^{(m)}_i \mathcal{S}^s_{i,k|k-1} \\
P^-_k &= \sum_{i=0}^{2L} W^{(c)}_i \big[\mathcal{S}^s_{i,k|k-1} - \hat{s}^-_k\big]\big[\mathcal{S}^s_{i,k|k-1} - \hat{s}^-_k\big]^T \\
\mathcal{O}_{k|k-1} &= M\big(\mathcal{S}^s_{k|k-1}\big) + \mathcal{S}^n_{k-1} \quad \text{(see Eq. (29))} \\
\hat{o}^-_k &= \sum_{i=0}^{2L} W^{(m)}_i \mathcal{O}_{i,k|k-1}
\end{aligned} \tag{32}$$

   where $\hat{s}^-_k$ is the a priori state estimate at step k, P^-_k is priori estimate error covariance.
4. Measurement update:

$$\begin{aligned}
P_{\tilde{o}_k\tilde{o}_k} &= \sum_{i=0}^{2L} W^{(c)}_i [\mathcal{O}_{i,k|k-1} - \hat{o}^-_k][\mathcal{O}_{i,k|k-1} - \hat{o}^-_k]^T \\
P_{s_k o_k} &= \sum_{i=0}^{2L} W^{(c)}_i [\mathcal{S}_{i,k|k-1} - \hat{s}^-_k][\mathcal{O}_{i,k|k-1} - \hat{o}^-_k]^T \\
\mathcal{K} &= P_{s_k o_k} P^{-1}_{\tilde{o}_k\tilde{o}_k} \\
\hat{s}_k &= \hat{s}^-_k + \mathcal{K}(o_k - \hat{o}^-_k) \\
P_k &= P^-_k - \mathcal{K} P_{\tilde{o}_k\tilde{o}_k} \mathcal{K}^T
\end{aligned} \tag{33}$$

   $\hat{s}_k$ is the posteriori state estimate at step k; P_k is the posteriori error covariance estimate.

By using the tracking based on the UKF, we get an estimate of the obstacle's position according to the past states, which enables us to check the current detection result namely in what concerns obstacles that are lost (see Fig. 11). Since multiple obstacles can exist in one

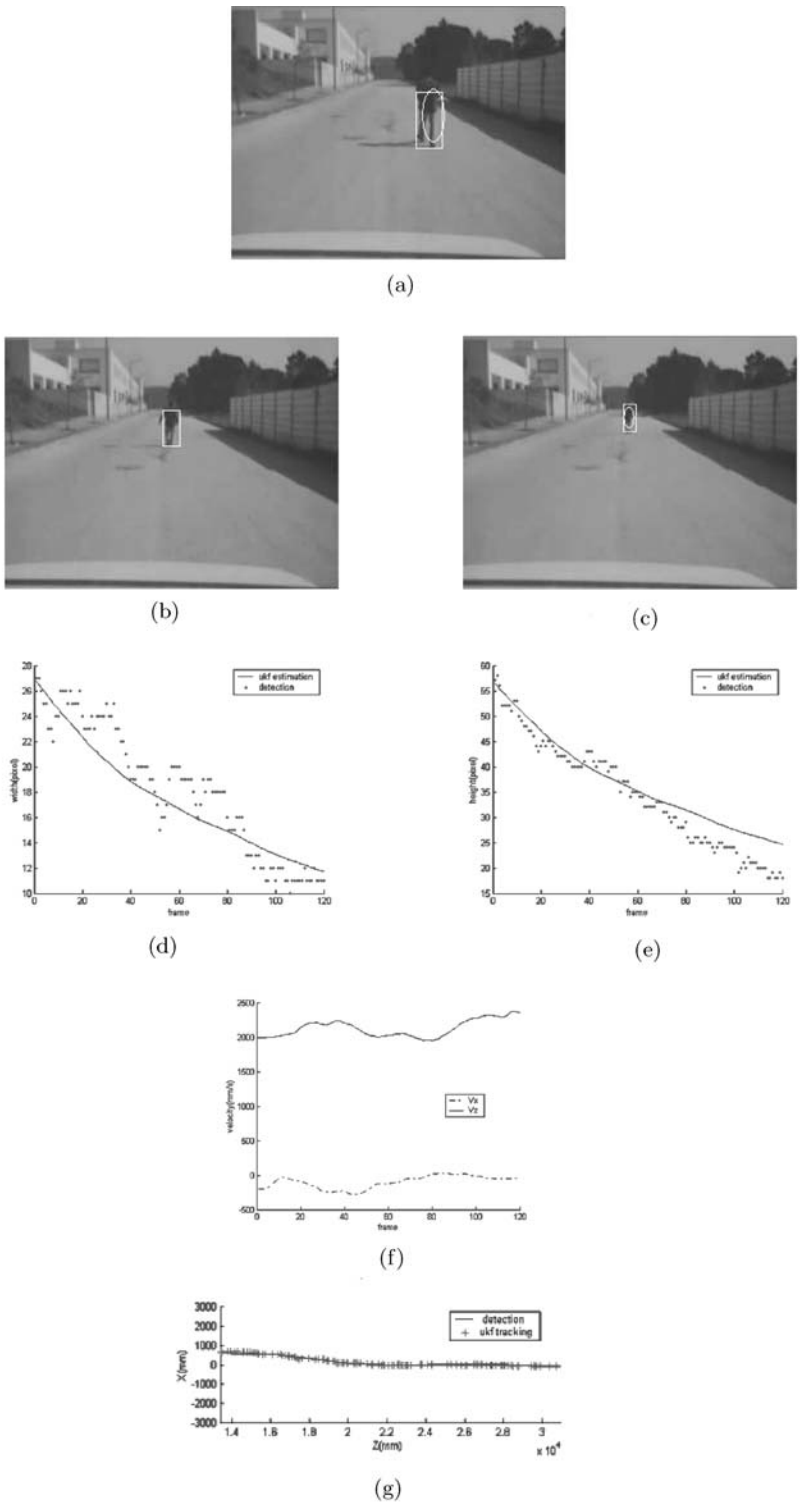

Figure 11. Results from the tracking process. (a) 35th frame (the ellipse and the rectangle bound the regions that result from the detection and tracking processes respectively, (b) 84th frame (tracking after a detection miss), (c) 118th frame, (d) tracking of the obstacle width, (e) tracking of the obstacle height, (f) the velocities along the X axis and the Z axis, and (g) the trajectory of the tracked obstacle.

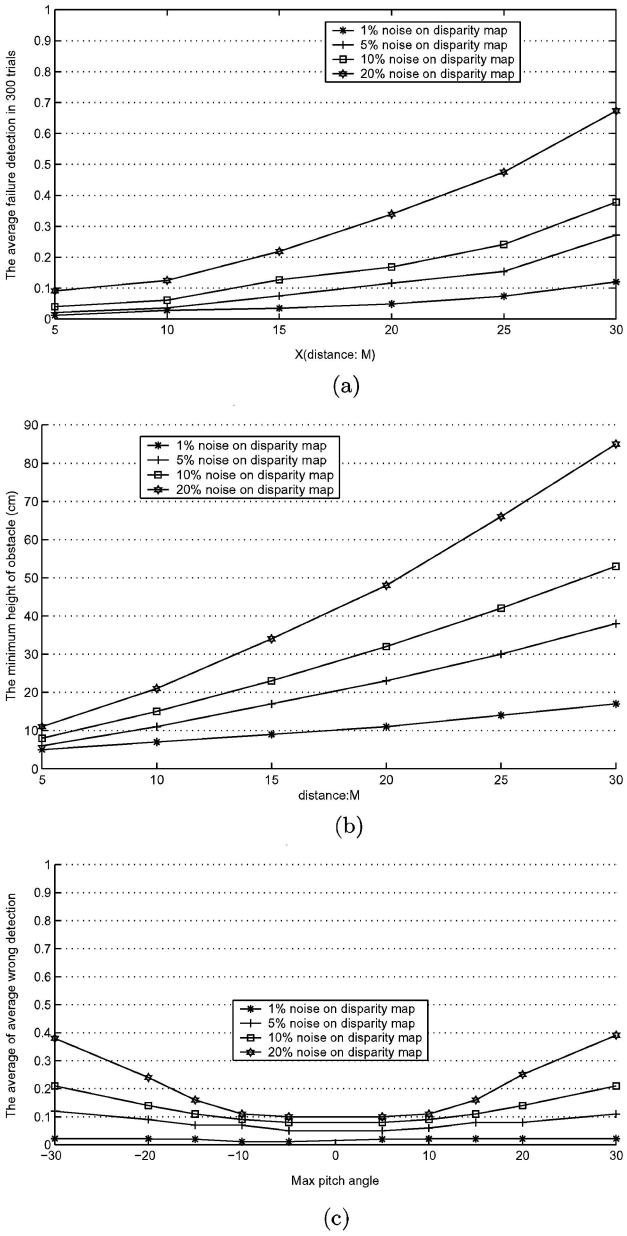

Figure 12. Simulation results for the case of a square obstacle with 40 cm width at different distances, for different levels of noise. (a) average failure detection ratio (points wrongly classified as road points) in 300 trials, (b) the minimum height of an obstacle that can be reliably detected, i.e. with a ratio of failure detection of less than 30%, and (c) average ratio of points wrongly classified as obstacle points in 300 trials.

typical scene, the mean and variance of the intensity are used to evaluate the correspondence between the obstacle regions from successive frames. From the results shown below, one concludes that the UKF tracking provides more reliable results and also improves the accuracy.

4. Experiments

In this section we present results that characterize the performance of the method described in this paper. Results from both real experiments and simulation are presented. Two issues are specifically addressed:

the smallest obstacle that can be reliably detected at a specific distance, for a certain noise level, and the level of variation in the road slope that is acceptable.

To perform the simulation experiments we used a Monte Carlo method. For that purpose a specific set of parameters for the stereo pair of cameras was considered. The pitch of the ground plane was varied between certain values. Given the ground plane and the parameters of the stereo cameras setup, the disparity map was then computed. The disparity values were corrupted by zero-mean Gaussian noise with different values of standard deviation. The simulation was run 300 times for each set of noise parameters (for each set of noise parameters different seed values were used). The obstacle considered was a simulated square, placed vertically at a variable distance from the stereo pair. The width of the square was varied between 10 cm and 90 cm. For each simulation run, the height of the camera was also affected by 5% noise, which can be interpreted as the fluctuations caused by jolting when a real vehicle is moving.

The whole disparity map results mostly from two main geometric elements: the ground plane (with pitch variation) and the obstacle area. To evaluate the algorithm two indicators are used:

- The ratio between the number of points wrongly classified as road points and the total number of obstacle points. This ratio is considered as an indicator of detection failure.
- The ratio between the number of points wrongly classified as obstacle points and the total number of road points. This ratio is considered as an indicator of wrong detection.

Figure 12(a) shows the average ratio between the number of points wrongly classified as road points and the total number of obstacle points in the disparity map. If the indicator of failure detection is greater than 50% the decision involves a high risk. Let us restrict the indicator of failure detection to a maximum of 30%. Using this value we can estimate the minimum obstacle size that can be detected. That result is shown in Fig. 12(b).

Figure 12(c) shows the simulation results obtained for roads with smoothly varying slopes, for different levels of noise in the disparity map. The axis with the label "Max pitch angle" indicates the angular difference between the tangents at the beginning and at the end of the road segment considered in the simulation. From Fig. 12(c), we can see that the values of the indicator of wrong detection are not sensitive to the variations on the road gradient.

From the data presented in Fig. 12, one can conclude that the algorithm based on the quasi-planar assumption is capable to detect obstacles on a smooth surface with a range of different slopes. Figure13 shows a sample of a stereo sequence of a real scene with large gradients along the road. Images with a resolution of 320×240 are grabbed with an IEEE 1394 interface card by a digital stereo head, with baseline of 20 cm. The two stereo cameras are mounted with parallel axes and carefully aligned for high quality disparity images. With this setup of two stereo cameras with focal length of 12 mm, the designed look-ahead distance range is from 5 m to 20 m. Before testing the algorithm, a calibration scene like the scene shown in Fig. 7 was used to estimate parameters of the base plane. From the Fig. 13, it is clear that the algorithm performed well in the scene with a road with a large gradient. The green pixels correspond to the pixels belonging to the road markings and the gray pixels correspond to the areas considered to be too far from the vehicle. However the resolution of the disparity computed from the two real images did not allow the precise detection of the road curb.

Figure 14 shows the experiment in a dense traffic scene. The indicator of failure detection for the closest obstacles in front of vehicle is very low, for the common traffic condition. However for conditions of dense traffic, obstacles which are close to each other are often merged into one. Since the quasi-planar road assumption is applied, the rate of wrong detection usually caused by lane markings is almost zero.

The method proposed in this paper also presents good time performance. It has been implemented in C++ on a commercial Pentium IV 1.4 GHz. The whole process for detecting obstacle points and segmentation of isolated obstacles is performed within 40 ms. The performance of each step is shown in Table 1.

Table 1. Performance of the algorithm.

Step	Time (ms)
Stereo analysis	<15
Pre-filtering	<2
Obstacle detection	<5
Watershed Segmentation	<15
Tracking	<1

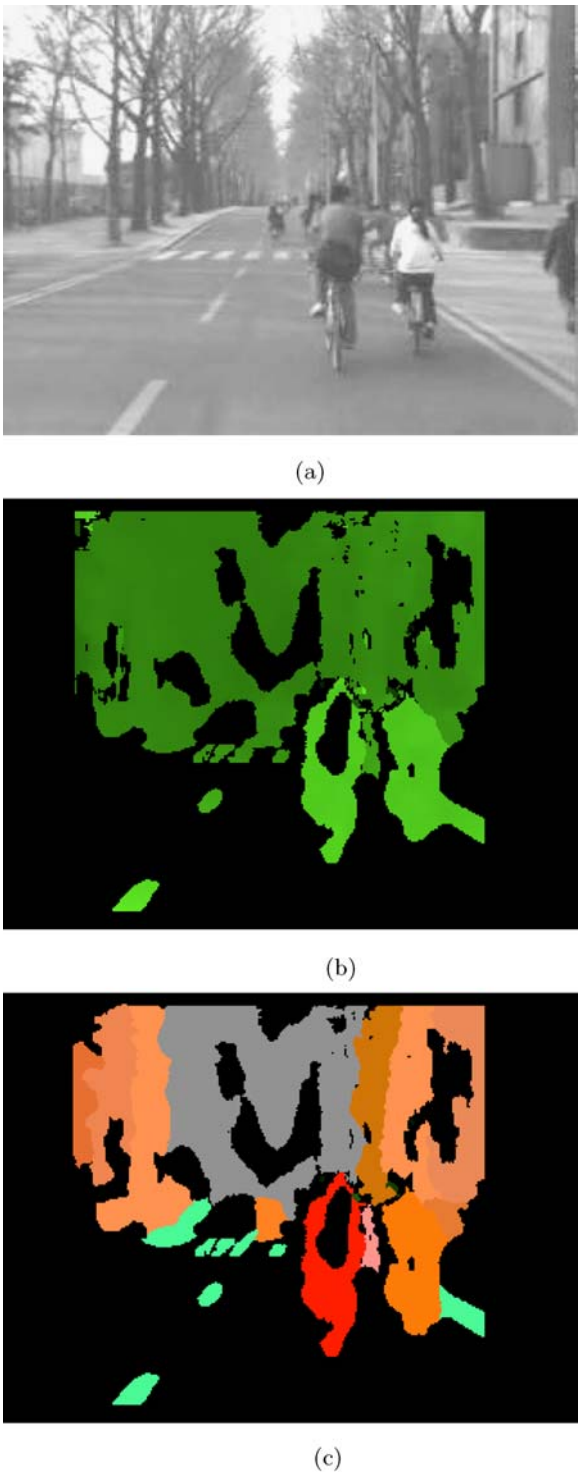

Figure 13. An image from a real scene with a road with gradient. (a) left image from a stereo sequence, (b) the disparity image, and (c) the segmentation result.

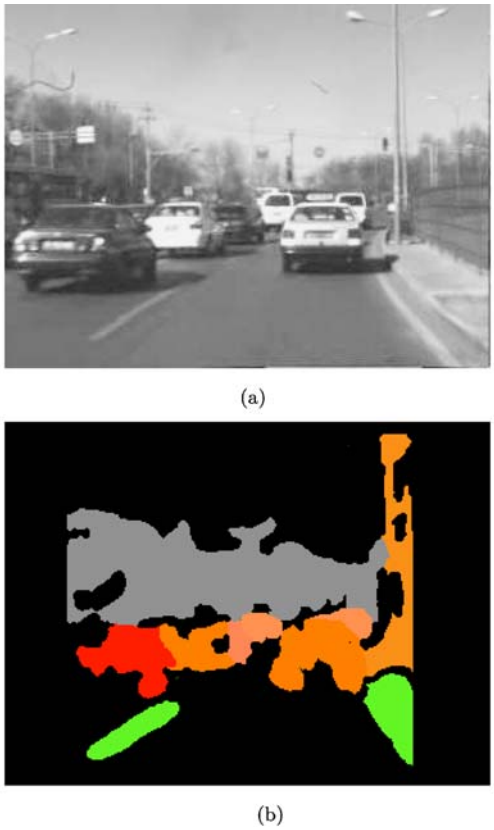

Figure 14. An image from a dense traffic scene. (a) Left image from a stereo sequence obtained with dense traffic, and (b) the segmentation image.

5. Conclusions

In this paper we first extend the usual planar road model by proposing the flexible quasi-planar road assumption. Then we describe a fast and robust method for obstacle detection on flat roads or non-flat roads. This method uses the surface normal vectors in 3D space. The normal vectors enable the distinction between the obstacles and the road surface. The method is implemented in disparity space. With the results from the detection, obstacle segmentation can be performed by a clustering procedure. This clustering procedure is performed by means of a robust watershed transformation. Better than EKF, UKF approximates the nonlinear system up to third order. This approach has several advantages namely:

1. We do not use the restriction of a planar model. Note that in a dynamic environment, and with noisy

data, it is difficult to reliably estimate a fully planar model. As a result of this approach there is no need for patterns on the road (be it planar or non-planar). On the other hand shadows and/or landmarks on the road do not affect the procedure.
2. This method does not impose any specific strict restriction, for example, that the baseline of the cameras must be parallel to the road profile.
3. This method is entirely implemented on disparity space. In addition the algorithm can be easily implemented on a standard microprocessor or on a DSP.

This system will be fully integrated in our experimental vehicle used in the Cybercars project.

Further improvements will include fusion of multiple kinetic models of maneuvering obstacles.

Appendix A: Proof of the Result of Eq. (10)

In Fig. 5, the angle θ_0', $\angle TON$ (in image plane) is the maximum angle between road surface normal and the main direction of the road normal measured in the image plane. OT is the minimum distance of road surface to origin. They are used to determine the thresholds for the road surface. Since the angle β will not affect the difference between the maximum normal angle and the angle relative to the main direction (since both are rotated by β), we have:

$$\begin{bmatrix} a_c \\ b_c \\ c_c \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} a_w \\ b_w \\ c_w \end{bmatrix} \tag{34}$$

Since the maximum angle of the road surface normal relative to the main direction happens on the circle where $\theta = \theta_0$, in the spherical coordinate system, a point on the arc can be represented as follows (see Fig. 15):

$$\begin{aligned} a_w^0 &= \sin\theta_0 \cos\phi \\ b_w^0 &= \cos\theta_0 \\ c_w^0 &= \sin\theta_0 \sin\phi \end{aligned} \tag{35}$$

Based on definition of θ_0', we have:

$$\begin{aligned} \theta_0' &= \max\left(\arctan\frac{a_c}{b_c}\right) \\ &= 90 - \min\left(\arctan\frac{b_c}{a_c}\right) \end{aligned} \tag{36}$$

We define

$$f(\phi) = \frac{b_c}{a_c} \tag{37}$$

On the basis of Eqs. (34) and (36), Eq. (37) can be rewritten as follows:

$$\begin{aligned} f(\phi) &= \frac{\cos\alpha}{\tan\theta_0 \cos\phi} - \sin\alpha \tan\phi \\ &= \frac{t_a}{\cos\phi} - t_b \tan\phi \end{aligned} \tag{38}$$

where, t_a and t_b are constants. From the Fig. 16, we can conclude that there is only one minimum in the region $[-\pi/2, \pi/2]$. We let: $\frac{\partial f(\phi)}{\partial \phi} = 0$. Thus:

$$\sin\phi^* = \frac{t_b}{t_a} = \tan\alpha \tan\theta_0 \quad \text{if } \tan\alpha \tan\theta_0 \le 1 \tag{39}$$

On the basis of Eqs. (37) and (39), we have

$$\tan\theta_0' = ((\cos\alpha/\tan\theta_0)^2 - (\sin\alpha)^2)^{-\frac{1}{2}} \tag{40}$$

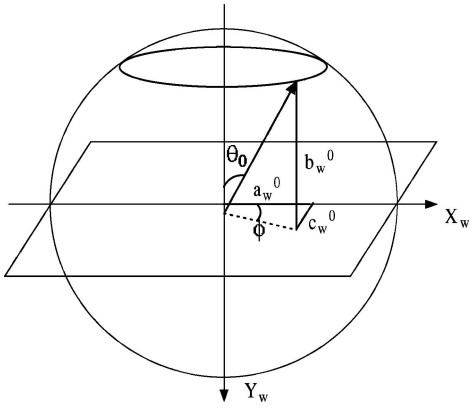

Figure 15. Demonstration of Eq. (10).

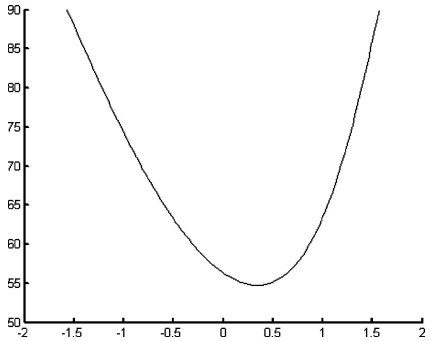

Figure 16. Plot of $f(\phi)$ for $-\pi/2 \le \phi \le \pi/2$.

From the definition of OT

$$OT = \text{Min}(b_c) = \cos\alpha\cos\theta_0 - \sin\alpha\sin\theta_0 \quad (41)$$

Acknowledgments

This work was partially supported by the Chinese High Technology Development Program, by a Portugal-China Science and Technology Cooperation Project, and by the European project Cybercars. The authors would like to thank Luis Conde for his help in the experiments, and Dr. Ming Yang's valuable comments and discussion.

References

Bertozzi, M. and Broggi, A. 1998. GOLD: A parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE Transactions on Image Processing*, 7(1):62–81.

Bertozzi, M., Broggi, A., and Fascioli, A. 1997. Obstacle and lane detection on ARGO. In *Proc. IEEE Int. Transportation Systems Conference'97*, Boston, USA, pp. 1010–1015.

Bertozzi, M., Broggi, A., and Fascioli, A. 1998. Stereo inverse perspective mapping: Theory and applications. *Image and Vision Computing*, 16:585–590.

Bertozzi, M., Broggi, A., and Fascioli, A. 2000. Vision-based intelligent vehicles: State of the art and perspective. *Robotics and Autonomous Systems*, 32:1–16.

Beucher, S. and Bilodeau, M. 1994. Road segmentation and obstacle detection by a fast watershed transformation. *Intelligent Vehicles Symposium'94*, Paris.

Enkelmann, W. 1997. Robust obstacle detection and tracking by motion analysis. *IEEE Conf. on Intelligent Transportation Systems ITSC '97*, Boston, pp. 9–12.

Fischler, M.A. and Bolles, R.C. 1981. Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the Association and Computing Machine*, 24(6):381–395.

Franke, U., Gavrila, D., Görzig, S., Lindner, F., Paetzold, F., and Wöhler, C. 1999. Autonomous driving approaches downtown. *IEEE Intelligent Systems*, 13(6).

Julier, S.J. and Uhlmann, J.K. 1997. A new extension of the kalman filter to nolinear systems. In *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Controls*.

Konolige, K. 1997. Small vision systems: Hardware and implementation. In *8th International Symposium on Robotics Research*, Hayama, Japan.

Labayrade, R., Aubert, D., and Tarel, J.-P. 2002. Real time obstacle detection on non flat road geometry through'V-Disparity' representation. In *Proceedings of IEEE Intelligent Vehicle Symposium*.

Onoguchi, K., Takeda, N., and Watanabe, M. 1998. Planar projection stereopsis method for road extraction. *IEICE Transaction on Information and Systems*, E81-D:1006–1018.

Saneyoshi, K. 1994. 3-D Image recognition system by means of stereoscopy combined with ordinary image processing. *Intelligent Vehicles '94*, Paris, pp. 24–26.

Se, S. 2002. Ground plane estimation, error analysis and applications. *Robotics and Autonomous Systems*, 39(2):59–71.

Shashua, A. and Navab, N. 1996. Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:873–883.

Thorpe, C., Herbert, M., Kanade, T., and Shafer, S.A. 1988. Vision and navigation for the Carnegie-Mellon Navlab. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:362–373.

Vincent, L. and Soille, P. 1991. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Machine Intell.*, 13:583–598.

Wang, D. 1997. A multiscale gradient algorithm for image segmentation using watersheds. *Pattern Recognition*, 678(12):2043–2052.

Wang, D. 1998. Unsupervised video segmentation based on watersheds and temporal tracking. *IEEE Trans. Circuits Syst. Video Technol.*, 8(5):539–546.

Williamson, T. 1998. A high-performance stereo vision system for obstacle detection. Ph.D. Dissertation, Robotics Institute, Carnegie-Mellon University, CMU-RI-TR-98-24.

Qian Yu received the B.E. degree in Computer Science from Tsinghua University, Beijing, China, in 2001, and the Master degree in Computer Science also from Tsinghua University in 2004, working at the Artificial Intelligence Laboratory. From October 2002 to April 2003, he was a visiting student at the Institute of System and Robotics (ISR), University of Coimbra, Portugal. His current research interests are in computer vision and robotics.

Helder Araujo is currently Associate Professor in the Department of Electrical and Computer Engineering, University of Coimbra, Portugal. He is co-founder of the Portuguese Institute for Systems and Robotics (ISR), where he is now a Researcher and Vice-Director of the Coimbra pole. His primary research interests are in computer vision and mobile robotics.

Hong Wang received his Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University in 1993. He is currently an associate professor at Department of Computer Science and Technology, Tsinghua University. He worked as a visiting researcher at the Department of Intelligent Assistant Driving, Daimler-Benz Research, Stuttgart, Germany, from August 1996 to August 1997. His main research interests include Artificial Intelligence, Mobile Robotics, Vision Navigation, Multi-sensor Data Fusion. He has published over 40 papers in international conference and journals. He is a member of Special Committee of Machine Perception and Virtual Reality of the Chinese Association of Artificial Intelligence and a member of Scientific Committee of the Olympiad in Informatics of the Chinese Computer Association. He has served as an Associated Director of the Central Laboratory of the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University.