Stereo-Based All-Terrain Obstacle Detection Using Visual Saliency

Pedro Santana
LabMAg, University of Lisbon
Pedro.Santana@di.fc.ul.pt

Magno Guedes
R&D Division, IntRoSys, S.A.

Luís Correia
LabMAg, University of Lisbon

José Barata
UNINOVA, New University of Lisbon

Last revision: January 4, 2012

Abstract
This paper proposes a hybridisation of two well-known stereo-based obstacle detection techniques for all-terrain environments. While one of the techniques is employed for the detection of large obstacles, the other is used for the detection of small ones. This combination of techniques opportunistically exploits their complementary properties to reduce computation and improve detection accuracy. Being particularly computation intensive and prone to generate a high false positive rate in the face of noisy three-dimensional point clouds, the technique for small obstacle detection is further extended in two directions. The goal of the first extension is to reduce both problems by focussing the detection on those regions of the visual field that detach more from the background and, consequently, are more likely to contain an obstacle. This is attained by means of spatially varying the data density of the input images according to their visual saliency. The second extension refers to the use of a novel voting mechanism, which further improves robustness. Extensive experimental results confirm the ability of the proposed method to robustly detect obstacles up to a range of 20 m on uneven terrain. Moreover, the model runs at 5 Hz on 640 × 480 stereo images.
1 Introduction

The unconstrained appearance of obstacles in all-terrain environments results in the essential use of volumetric information for their detection. Recent developments have made practical the use of laser scanners and stereoscopic vision sensors for the acquisition of volumetric data, i.e., dense three-dimensional (3-D) point clouds. A 3-D point cloud is said to be dense when it is not confined to the representation of notable locations in the environment. In the case of stereoscopic vision, this virtually means that the 3-D position of every patch of the environment that is imaged by the sensor needs to be known. With a high resolution vision sensor, one can obtain denser 3-D point clouds than those produced by typical laser scanners. When compared to laser scanners, stereoscopic vision sensors also tend to be less power consuming and to be lighter, smaller, and cheaper. In contrast to laser scanners, stereoscopic vision sensors provide 3-D data whose acquisition is exactly registered and synchronised with visual data. This is an important asset to facilitate both object and place recognition. However, stereoscopic sensors are usually less accurate than laser scanners, more computationally intensive, more dependent on lighting conditions, and less prepared to reconstruct the 3-D structure of textureless surfaces.

Motivated by the previously mentioned benefits of stereoscopic vision, this paper proposes a set of developments to tackle the major challenges associated to its use for the problem of all-terrain obstacle detection. These challenges are mostly related with the ability of efficiently and robustly processing the considerable amount of noisy and inaccurate sensory data that is generated by the sensor. Nonetheless, we believe that the concepts developed in this paper may also be applied to volumetric data generated by 3-D laser scanners, provided that the point cloud is associated with visual data. This can be done by registering a camera to the laser scanner.

Hard assumptions on the environment’s topology are usually exploited to enable fast stereo-based terrain assessment. The most common assumption states that the terrain can be reasonably approximated by planar models (Batavia and Singh, 2001; Labayrade et al., 2002; Dang and Hoffmann, 2005; Soquet et al., 2007; Vernaza et al., 2008; Konolige et al., 2009; Poppinga et al., 2008; Hadsell et al., 2009), which is often not the case on all-terrain. A well known alternative defines obstacles according to geometric relationships among neighbour 3-D points (Bellutta et al., 2000; Talukder et al., 2002; Manduchi et al., 2005; Dubbelman et al., 2007; van der Mark et al., 2007), thus relaxing the planar terrain assumption. However, this alternative technique is computationally intensive and suffers from a considerable sensitivity to both noise and sparsity in 3-D point clouds. This paper proposes and validates a hybrid model for obstacle detection, in which the strengths of both techniques are brought together to attain higher levels of robustness, accuracy, and computational efficiency. In this sense, two key aspects of this work are the use of the planar assumption to detect large obstacles and the use of the geometric relationships among neighbour 3-D points to detect small ones.

Additionally, variable data density, also known as space-variant resolution (SVR), is used in the computations to improve the overall efficiency of the algorithms. Visual saliency is used to determine regions of particular interest where a higher density of data must be processed. This saliency-based modulation of the SVR allows computation to be focused on the most promising regions of the environment while reducing the false positive rate. Finally, a novel voting
mechanism is also introduced to augment the robustness of obstacle detection in the presence of noisy 3-D point clouds. Preliminary versions of this saliency-based obstacle detection model can be found in previous publications (Santana et al., 2008, 2009, 2010).

This paper is organised as follows. Section 2 relates this proposed method to previous work. Section 3 overviews the proposed system architecture. Sections 4 and 5 describe the novel way visual saliency is computed and the ground-plane is estimated, respectively. Then, in Section 6, particular focus is given to the small obstacle detector. Parameter selection details are subsequently provided in Section 7. Finally, experimental results are described in Section 8, followed by conclusions and future work in Section 9.

2 Related Work

This section describes related work on terrain classification based on volumetric data provided by laser scanners or stereoscopic vision sensors, with particular emphasis on work validated on all-terrain environments.

Only a few assumptions can be made regarding the structure of all-terrain environments. A typical assumption states that the environment’s ground can be modelled by planes. In this case, obstacles are considered to be these 3-D points standing out of the estimated ground-planes (Dang and Hoffmann, 2005; Vernaza et al., 2008; Konolige et al., 2009). Some local statistics regarding the distance 3-D points have to the ground-planes can be used to reduce the method’s sensitivity to false positives (Hadsell et al., 2009). This plane-based approach can also be applied indirectly, such as through an estimated homography (Batavia and Singh, 2001) or through a Hough space of planes (Poppinga et al., 2008). Also under the planar assumption, the v-disparity space approach (Labayrade et al., 2002; Soquet et al., 2007) is usually employed for on-road obstacle detection. These plane-based approaches are highly appealing due mostly to their computational efficiency. However, the unevenness of typical all-terrain environments breaks down the planar assumption. When that happens, a terrain’s surface variations can be erroneously characterised as obstacles. Nevertheless, compelled by its computational parsimony, we propose in this work to make a contextualised use of the planar assumption, i.e., only for large obstacle detection.

A way of relaxing the planar terrain assumption is through the use of heuristics applied locally to the disparity/range image (Broggi et al., 2005; Schafer et al., 2005; Caraffi et al., 2007; Konolige et al., 2009). Heuristics in the form of local point statistics obtained directly from the 3-D point cloud can also be used to produce accurate results (Wellington and Stentz, 2004; Lalonde et al., 2006). However, this technique is impracticable for stereoscopic vision due mostly to its noisy nature. Heuristics on the residual resulting from a line fitting process can also be applied on a scan-by-scan basis to data generated by two-dimensional (2-D) laser scanners (Moorehead et al., 1999; Batavia and Singh, 2002; Castano and Matthies, 2003; Urmson et al., 2006; Andersen et al., 2008).

Alternative heuristics can be applied when a dense 3-D point cloud is available. One example is traversability estimation in terms of the residuals resulting from several local plane-fitting
processes applied to the 3-D point cloud (Moorehead et al., 1999; Gennery, 1999; Singh et al., 2000; Goldberg et al., 2002; Biesiadecki and Maimone, 2006; Ye, 2007). More recently, octree decomposition has been employed to create grid representations of the 3-D point cloud (Rusu et al., 2009). Polygonal models are then fitted to each cell and heuristically labelled as ground, level, vertical obstacles, stairs, or unknown. To determine their traversability, plane-fitting processes are then applied to the models falling in the latter category. In addition to its heuristic definition of obstacle, this method still lacks experimental validation on all-terrain.

Heuristics can ultimately be learned (Wellington and Stentz, 2004; Bajracharya et al., 2008). Linguistic heuristic rules can also be applied to blend several cues when computing traversability indexes (Seraji, 1999, 2006). The major limitation of heuristic-based solutions is the difficulty in defining obstacles in terms of the robot’s mobility. A way of circumventing this limitation is through the construction of Digital Elevation Maps (DEM) of the environment, upon which a detailed kinematic and/or dynamic model of the robot can be used for safe motion planning (Kelly and Stentz, 1998; Lacroix et al., 2002; Plagemann et al., 2008; Kolter et al., 2009). However, these solutions tend to be too computationally demanding.

To take into account key mobility properties, such as the clearance height under the robot, obstacles can also be defined in terms of geometrical relationships between neighbour 3-D points. In the case of 2-D laser scanners these relationships can be computed in scan-by-scan fashion (Chang et al., 1999). A similar process can also be applied on a column-by-column way in stereo-based systems (Bellutta et al., 2000; Dubbelman et al., 2007). Probabilistic models have been successfully applied for improved robustness in the integration of evidence across 2-D laser scans (Thrun et al., 2006). Talukder et al. (2002) and Manduchi et al. (2005) proposed and validated on a stereo-based system a more general geometric approach that can be applied to 3-D point clouds. However, this method is computationally intensive and sensitive to artefacts induced by the 3-D reconstruction process. These limitations have been partially mitigated with the use of look-up tables and explicit handling of uncertainty (van der Mark et al., 2007). In this work, a higher level of robustness is attained with a novel voting filter and an improved posture compensation mechanism. To speed up computation we propose the synergistic use of visual saliency and SVR.

The robot’s focus of attention and respective resolution can be defined in terms of the robot’s speed so as to avoid holes/overlaps in the analysis of consecutive frames (Kelly and Stentz, 1998). This process is in line with the active vision approach (Bajcsy, 1988; Aloimonos et al., 1988; Ballard, 1991) and it can also be used to reduce the risk of not detecting obstacles (Grandjean and Matthies, 1993). Our method dynamically adapts the resolution and focus of attention based on the contents of the visual input. Hence, despite sharing the same goal of reducing the cost of perception to its minimum, the two approaches show complementary properties. As will be shown, our saliency-based SVR mechanism has the additional advantage of reducing the false positive rate.

Visual saliency has been thoroughly employed for object detection in indoor environments (Vijayakumar et al., 2001; Orabona et al., 2005; Moren et al., 2008; Meger et al., 2008; Yu et al., 2007). It has also been used to select strong landmarks for visual simultaneous localisation and mapping in urban environments (Newman and Ho, 2005; Frintrop et al., 2007). Laser-based
range images have been used to focus the analysis of registered colour images on the task of searching for information placards along dirt roads (Hong et al., 2002). In this case one might say that salient regions in the laser data are used to focus the analysis of a colour-based detector. However, to the best of our knowledge, our proposed method is the first application of visual saliency to modulate all-terrain obstacle detection.

3 System Overview

The proposed model (see Fig. 1) is characterised in part by the novel integration of two complementary obstacle detection techniques. The two techniques differ mostly in their definition of an obstacle. The first considers as obstacles those 3-D points that stand out from an estimated ground-plane (Dang and Hoffmann, 2005; Vernaza et al., 2008; Konolige et al., 2009). Because most all-terrain environments are not perfectly planar, this detector can search robustly only for large obstacles. Small variations in height on uneven terrains would often be confused with small obstacles. Hence, the obstacle detector is configured in a way that only obstacles above a large distance \( h \) off the ground plane are detected. Smaller obstacles are instead considered by the second technique, which relaxes the planar assumption by defining obstacles according to geometrical relationships between neighbour 3-D points. Although this second technique is more robust in uneven terrains, it requires dense 3-D point clouds. The level of noise and sparseness of the 3-D point cloud in large homogeneous objects, due to failure in the stereo reconstruction process, makes this technique less capable of detecting large obstacles. As confirmed by experimental results (see Section 8.3), from the strengths of both techniques emerges a more robust all-terrain obstacle detector.

Another relevant contribution of the model is the small obstacle detector itself (see Section 6), which departs from previous related work (Talukder et al., 2002; Manduchi et al., 2005; van der Mark et al., 2007) by improving its robustness and computational efficiency. The latter is partially due to the innovative use of visual saliency (see Section 4) to modulate all-terrain obstacle detection. The challenges posed by this new domain required some adaptations to the standard way of computing saliency (e.g., (Itti et al., 1998; Frintrop et al., 2005)). The model also exploits for the first time visual saliency on the task of ground-plane estimation (see Section 5). The estimated ground-plane supports the large obstacle detector while it permits the rotation of the 3-D point cloud to compensate for the robot’s posture in the small obstacle detector computations (see Section 6.5).

To avoid unnecessary and consequently inefficient applications of the small obstacle detector, the output of the large obstacle detector is used to mask the saliency map with negative values. Using the masked saliency map instead of the original saliency map, the small obstacles detector will ignore regions with negative saliency that have already been labelled by the other detector. Finally, the outputs of both detectors are merged in order to produce the final obstacle map.

We now characterise which types of obstacle can be detected with the proposed model. For this purpose we assume that an environment can be generally classified in terms of its
Figure 1: Proposed model building blocks.
slipperiness, compressibility, permeability, and morphology. With the exception of the fourth property, which can be assessed with volumetric data, these properties require an appearance-based analysis. Consequently, the proposed model is constrained to detect obstacles that can be segmented by the background based solely on their 3-D morphology. Both positive (e.g., rocks/trees) and negative (e.g., ditches/holes) obstacles are detected by the proposed model provided that they are associated to sufficient 3-D information.

From this analysis it follows that the proposed model is not capable of detecting water bodies nor is it capable of distinguishing between compressible (e.g., wall of tall grass) and noncompressible obstacles. Nevertheless, due to its fragile morphological characteristics, spurious tall grass is filtered out by the system. It follows also that the model is unable to assess the slipperiness of the terrain, as it typically requires the identification of the materials composing it.

4 Saliency Computation

The goal of using visual saliency is to determine which regions of the visual field stand out more significantly from the background. The basic idea is that the higher its associated saliency, the more prone a given region of the visual field is to contain an object. Thus, perceptual processes exploiting this input are better fit to handle the speed-accuracy trade-off. The following describes the proposed saliency model, which is an adaptation for all-terrain environments of the biologically-inspired model proposed by Itti et al. (1998).

Let \( L \) be the left image of an image pair obtained from a stereoscopic vision sensor. To reduce computational cost, saliency is computed on a region of interest (ROI) of \( L \). The ROI is a horizontal strip between the bottom row and row \( u \). Row \( u \) corresponds to the uppermost row containing at least a given percentage \( \zeta \) of pixels with an associated depth within the minimum and maximum considered ranges for obstacle detection, \( r_{\text{min}} \) and \( r_{\text{max}} \), respectively. With this process we guarantee that the definition of the ROI is not affected by spurious pixels whose associated 3-D points are erroneously beyond \( r_{\text{max}} \). To further reduce computational cost, all image operators are performed over 8-bit pixels whose magnitude is clamped to \([0,255]\). All experiments use \( 640 \times 480 \) input images.

A dyadic Gaussian pyramid \( I(\sigma) \) with six levels \( \sigma \in \{0,\ldots,5\} \) is created from the intensity channel of the ROI. The resolution scale of level \( \sigma \) is \( 1/2^\sigma \) times the ROI resolution scale. Intensity is obtained by averaging the three colour channels. Then, four on-off centre-surround intensity feature maps are created to promote bright objects on dark backgrounds. Four off-on centre-surround intensity feature maps are also created to promote dark objects on bright backgrounds. On-off centre-surround operations are performed by across-scale, point-by-point subtraction between level \( c \), with a finer scale, and level \( s \), with a coarser scale linearly interpolated to the finer scale, with \((c,s) \in \Omega = \{(2,4),(2,5),(3,4),(3,5)\}\). Off-on maps are computed the other way around, that is, by subtracting the coarser level from the finer level. On-off, \( I_{\text{on-off}}(c,s) \), and off-on, \( I_{\text{off-on}}(c,s) \), centre-surround maps are then combined to generate the intensity conspicuity map.
Figure 2: Examples of typical saliency maps. ROI with $r_{\text{max}} = 10$ m. The brighter the pixels in the saliency map, the higher their saliency. From these examples, it is possible to see the concentration of salient pixels on obstacle regions.

$$C_I = \sum_{i \in \{\text{on-off, off-on}\}} \left( \frac{1}{2} \bigoplus_{(c,s) \in \Omega} I^i(c,s) \right),$$  \hspace{1cm} (1)

where the across-scale addition $\bigoplus$ is performed with point-by-point addition of the maps after being scaled to the resolution of level $\sigma = 3$. Sixteen orientation feature maps, $O(\sigma, \theta)$, are created by convolving levels $\sigma \in \{1, \ldots, 4\}$ with Gabor filters tuned to orientations $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Gabor filters are themselves centre-surround operators and therefore require no across-scale subtraction procedure (Frintrop, 2006). As before, all orientation feature maps are combined at the resolution of level $\sigma = 3$ in order to create the orientation conspicuity map

$$C_O = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \left( \frac{1}{4} \bigoplus_{\sigma \in \{1, \ldots, 4\}} O(\sigma, \theta) \right).$$  \hspace{1cm} (2)

The saliency map $S$ (e.g., Fig. 2) is then obtained by modulating the intensity conspicuity map $C_I$ with the orientation conspicuity map $C_O$:

$$S = \mathcal{M} \left( \frac{1}{2} \cdot \mathcal{N}(C_I), \frac{1}{2} \cdot \mathcal{N}(C_O) \right),$$  \hspace{1cm} (3)

where $\mathcal{M}(A, B) = A \cdot \text{sigm}(B)$, $\text{sigm}(\cdot)$ is the sigmoid operator and $\mathcal{N}(\cdot)$ scales the provided image amplitude between $[0, 255]$.

This saliency model is essentially the one proposed by Itti et al. (1998) with one general improvement and two adaptations for all-terrain obstacle detection. The improvement was proposed by Frintrop (2006), and it refers to the use of both on-off and off-on feature channels separately. The first adaptation for all-terrain obstacle detection is in the normalisation operator
\( \mathcal{N}(\cdot) \), which does not try to promote maps according to their number of activity peaks, as is typically done. This is because the frequency at which objects appear in the maps is not necessarily correlated to their proneness of being obstacles to the robot.

The second adaptation for all-terrain obstacle detection is in the way conspicuity maps are blended. Rather than their typical addition, conspicuity maps are non-linearly multiplied. This results in making salient only those regions of the environment that are simultaneously conspicuous in both orientation and intensity channels. The basic idea is that obstacles on all-terrain environments are normally highly textured and at the same time conspicuous in the intensity channel. The non-linearity introduced by the sigmoid operator in Equation 3 aims at inhibiting the saliency map in the presence of low orientation conspicuity in order to remove the background noise inherent in textured outdoor terrains. When orientation conspicuity is strong, the sigmoid operator amplifies the saliency map and promotes the orientation conspicuity map over the intensity one.

Although not focused on all-terrain obstacle detection, a parallel and independent study (Hwang et al., 2009) has also reported the benefits of using a weighted product of the conspicuity maps as an alternative to their standard summation. Fig. 3 compares a saliency map computed by the proposed model with a saliency map obtained with the model proposed by Itti et al. (1998).

5 Ground Plane Estimation

Obstacles can be in part characterised as disturbances occurring at the ground surface. Thus, estimating how the robot relates to the ground-plane based on a dense 3-D point cloud is of extreme importance to support obstacle detection. However, the roughness of outdoor terrain renders highly unlikely the existence of a clear-cut ground-plane. Thus, robust estimation methods, such as random sample consensus (RANSAC) (Fischler and Bolles, 1981), must be employed. In a nutshell, a typical RANSAC procedure for ground-plane estimation in the ROI is composed of the following seven steps:

Step 1. Pick randomly a set \( R \) of three non-collinear 3-D points within range \([r_{\text{min}}, r_{\text{max}}]\) and
generate its corresponding ground plane hypothesis, $h_R$, with some straightforward geometry.

**Step 2.** The score of the plane hypothesis is the cardinality of the set of its inliers, $score(h_R) = |P_{h_R}|$. An inlier $p \in P_{h_R}$, where $p = (p_x, p_y, p_z)^T$, is a 3-D point whose distance to plane $h_R$, $d(p, h_R)$, is smaller than a given threshold $d_{plane}$. $P_{h_R}$ is thus the set of all 3-D points that cope with this condition and that are at the same time located in the ROI.

**Step 3.** Repeat steps 1 and 2 until $n_{hypo}$ hypotheses, composing a set $H$, have been generated.

**Step 4.** Select from $H$ the hypothesis with highest score for refinement, $b = \arg\max_{h \in H} score(h)$.

**Step 5.** Compute a refined version of $b$, $b'$, by fitting the inliers set of $b$, $P_b$. This fitting is done with weighted least-squares orthogonal regression via the well known Singular Valued Decomposition (SVD) technique. The weight $w_q$ of an inlier $q \in P_b$, where $q = (q_x, q_y, q_z)^T$, is given by $w_q = 1 - \frac{d(q, b)}{d_{plane}}$. That is, the farther $q$ is from $b$, the less it weights in the fitting process. Compute the inliers set of $b'$, $P_{b'}$, and substitute the current best ground-plane estimate with the refined one. That is, make $b = b'$ and $P_b = P_{b'}$. $P_{b'}$ is computed with the procedure described in Step 2.

**Step 6.** Iterate step 5 until $|P_b|$ becomes constant across iterations or a maximum number of iterations, $m_{refit}$, is reached.

**Step 7.** Take $b$ as the ground-plane estimate.

Points belonging to obstacles will inevitably generate poor ground-plane hypotheses in Step 1. This means that in environments cluttered with obstacles, a larger number of hypotheses must be generated to guarantee that a good hypothesis is found. This additional computation can be reduced by rejecting selected points if they are likely to belong to an obstacle. This likelihood is here defined in terms of visual saliency level (see Fig. 4(b-c)).

Formally, in Step 1, a randomly selected 3-D point $p$ is promptly rejected and so not considered to build a plane hypothesis, if $s_{p'} > \frac{\bar{s}}{p \cdot n_l}$, where $x \in [0, 255]$ is a number sampled from a uniform distribution each time the inequality is tested; $n_l \in [0, 1]$ is the ratio of pixels with saliency below a given threshold $l$; local saliency $s_{p'} \in [0, 255]$ is the maximum saliency within a given sub-sampled chess-like squared neighbourhood of $p'$, with size $q \cdot n_l$, $q$ being the empirically defined maximum size; and $\bar{s}$ is an empirically defined scaling factor. The goal of using the ratio of pixels with a saliency value under a given threshold is to allow the system to progressively fall-back to a non-modulated procedure as saliency reduces its discriminative power. This happens, for instance, in highly textured terrains, in which the sampling procedure is too constrained as a result of the saliency map’s cluttering. See Fig. 4(d) for an example of the output generated by the ground-plane estimation process.

Complementary mechanisms could be exploited for improved performance and robustness. In terms of saliency modulation, the score of each ground-plane hypothesis could be weighted
Figure 4: Typical example illustrating the advantages of using visual saliency to modulate the hypothesis generation step of a RANSAC procedure for ground-plane estimation. (b) Overlaid pixels (in red) correspond to 5000 3-D points randomly sampled without saliency modulation. (c) Same as (b) but with saliency modulation. (d) Overlaid pixels (in green) whose corresponding 3-D points lie on the estimated ground-plane.

according to the saliency of its supporting inliers. Additionally, the weight assigned to each 3-D point in the weighted least-squares orthogonal regression process could also be modulated by its associated saliency. Spatio-temporal prior information could also be applied to modulate the RANSAC process (Mufti et al., 2008; Chumerin and Hulle, 2008) with the purpose of increasing the chances of detecting the actual plane given a fixed number of RANSAC iterations. The 3-D points of obstacles detected in the previous frame could also be directly removed from consideration in the estimation process (van der Mark et al., 2007). If ground-plane estimation and detected obstacles removal are done sequentially and in an iterative way per frame, multiple planes could also be detected (Hadsell et al., 2009).
6 Small Obstacle Detection

Detecting obstacles based uniquely on the distance between 3-D points and the ground-plane is a brittle procedure on all-terrain. This is mostly because small variations in height on uneven terrains would often be confused with small obstacles. In those situations obstacles are better defined in terms of geometrical relationships between their composing 3-D points. Aligned with this idea, the model proposed by Talukder et al. (2002) and Manduchi et al. (2005), which is summarised in Section 6.1 and from now on designated Original Obstacle Detector (OOD), is taken as the starting point for our small obstacle detector. To reduce its computational cost and sensitivity to noise, the OOD is here extended with a SVR mechanism and a voting filter, as described in Sections 6.2 and 6.3, respectively.

6.1 Obstacle Definition

Let \( \{x, y, z\} \) be the basis that defines three axes relative to the centre of the left camera and with the \( z \)-axis aligned with the sensor’s optical axis (see Fig. 5(b)). This basis is the sensor’s local reference frame, \( \mathcal{F} \). Let \( P \) be the set of 3-D points, defined with respect to \( \mathcal{F} \), computed by the stereo-based 3-D reconstruction process and provided to the obstacle detector. As proposed by Talukder et al. (2002) and Manduchi et al. (2005), a 3-D point is considered to belong to an obstacle if it is compatible with any other 3-D point in the point cloud. Two 3-D points obtained from \( P, p_a = (x_a, y_a, z_a)^T \) and \( p_b = (x_b, y_b, z_b)^T \), are said to be compatible if

\[
H_{\text{min}} < |y_b - y_a| < H_{\text{max}} \land \frac{|y_b - y_a|}{\|p_b - p_a\|} > \sin \theta,
\]

where \( \theta \) is the maximum slant angle that the vehicle can safely negotiate, \( H_{\text{min}} \) is the clearance height under the vehicle, and \( H_{\text{max}} \) is the maximum height to be considered by the detector.

Fig. 5 provides an intuitive geometrical interpretation of this definition of compatibility. That is, all 3-D points that are compatible with a given 3-D point \( p \in P \) are encompassed by two truncated cones, \( U_p \) and \( L_p \). Although pointing in opposite directions, both upper and lower truncated cones are normal to the \( xz \) plane and have their vertexes located in \( p \). Additionally, both cones have an aperture angle of \( (\pi - 2\theta) \) and are limited by the planes \( y = H_{\text{min}} \) and \( y = H_{\text{max}} \).

Checking the compatibility between all possible 3-D points is limited by real-time requirements. Fortunately, according to Talukder, Manduchi et al., only a reduced subset of all combinations needs to be assessed if the process is carried out in the image space. For this purpose, \( p' = (p'_x, p'_y)^T \) is defined as the projection of the 3-D point \( p \in P \) onto the left camera’s image plane. This projection is defined with respect to the local reference frame \( \{x', y'\} \) and refers basically to the pixel in the left camera that is the image of the world point \( p \) (see Fig. 5(a)). Similarly, the two truncated cones of \( p \) project onto two truncated triangles in the image plane, \( U_{p'} \) and \( L_{p'} \), whose vertexes are both located at pixel \( p' \) (see Fig. 5(a)). The height of these truncated triangles is given by \( H_{\text{max}}f/p_z \), and their width can be approximated by

\[
\frac{2H_{\text{max}}f}{\tan \theta_{\max}p_z \cos \eta},
\]
Figure 5: Geometric interpretation of the compatibility test (Talukder et al., 2002; Manduchi et al., 2005). (a) Filled and unfilled circles represent 3-D points that are compatible and incompa-
tible, respectively, with points $p_1$ and $p_2$. It is also possible to depict the projections of 3-D points $p_1$ and $p_2$ onto the image plane, i.e., pixels $p'_1$ and $p'_2$, respectively. Note that the farther the points are from the sensor, the smaller the projections of their truncated cones are. For read-
ability reasons, $L_{p_1}$, $L_{p'_1}$, $L_{p_2}$ and $L_{p'_2}$ are not represented. (b) Diagram illustrating a situation in which the compatibility test allows the detection of an obstacle in front of the robot. The white circles represent the points falling in the upper truncated cone and consequently considered compatible, that is, pertaining to the same obstacle.

where $f$ is the camera’s focal length and $\eta = \arctan \frac{p_x}{p_z}$. In the image space, the set of pixels that may be compatible with $p'$, and consequently with $p$, are now constrained to those encompassed in both truncated triangles. Furthermore, if the image is scanned from bottom to top and from left to right, it suffices to consider only the upper truncated triangles to efficiently detect and label all pixels.

Finally, two points that are compatible to each other are said to pertain to the same obstacle. By transitivity, two points that are linked by a chain of compatible points are also said to belong to the same obstacle. This property will enable the segmentation of obstacles according to their 3-D relationships.
6.2 Space-Variant Resolution

Despite the advantages of using a truncated triangle in order to focus the application of the compatibility test, the computation time of the method is still on the order of seconds for $640 \times 480$ image-pairs. With quantisation and extensive use of look-up tables it is possible to attain faster processing rates (van der Mark et al., 2007). However, the loss of accuracy is unavoidable with such approximations. This section proposes the use of SVR as a complementary way to reduce the computational cost of the method.

For fast detection, pixels are *coarsely analysed* in a first phase according to the scanning procedure previously described, that is, from bottom to top and from left to right. However, this time the analysis is done with steps of $n$ pixels, which may be skipped based on their visual saliency and on an additional constraint (see below). Furthermore, compatibility is tested only against a sub-sampled set of the pixels falling inside the upper truncated triangle of each analysed pixel. This sub-sampling is done in a chess-like pattern with $1/m$ of the image resolution.

Whenever an obstacle is detected, it is foveated by performing a *finer analysis* of the region. Concretely, the pixels within the upper truncated triangle of the pixel just labelled obstacle are re-sampled from $1/n$ of the image full resolution, with $n < m$. This aims at improving the representation of any obstacle that has been detected.

As soon as the scanning procedure is completed, the detector enters in a second phase for *full resolution recovery*. This final phase is important as most morphological filters that might be applied afterwards perform better in high resolution. This second phase is implemented with the following region growing mechanism. Let $p_1'$ be a pixel labelled obstacle and $p_1$ its corresponding 3-D position. Every other pixel $p_2'$ whose distance to $p_1'$ is smaller than the highest number of skipped pixels, $||p_1' - p_2'|| < m$, is a candidate also to become labelled as obstacle. The final test to determine whether $p_2'$ is labelled as obstacle is done by checking whether its corresponding 3-D point, $p_2$, is at a distance from $p_1$ shorter than an approximation of the maximum allowed distance between 3-D points to be considered compatible, $||p_1 - p_2|| < H_{max}$.

As mentioned, in order to save computation, the compatibility test is conditionally applied in the coarse analysis phase. Concretely, the compatibility test is applied only if the *local saliency* increases between scanned pixels or an additional constraint is applicable. Local saliency is preferred over a pixel-wise one to reduce the effects of poor lighting conditions, which in some situations make objects’ upper part appear more salient than their lower part. Local saliency is computed by taking the maximum saliency from the set of pixels that share the column of the pixel being analysed and that are contained in its truncated triangle. Owing to the fact that the compatibility test is skipped over non-interesting regions and therefore with low likelihood of containing obstacles, certain regions of the environment are more coarsely analysed and so computational cost is saved.

Let us now describe the additional constraint that forces the execution of the compatibility test during the coarse analysis. The constraint aims at performing a fine analysis whenever an obstacle is detected and a *progressive* fine-to-coarse analysis of the obstacle’s boundaries. The latter effect is particularly useful in handling noisy data in and around obstacles. For this
Figure 6: Typical example of pixels analysed with saliency-based SVR. (c) Pixels analysed based on the saliency map depicted in (b). White pixels in (c) correspond to points that have been skipped by the detector due to lack of saliency or computed range, whereas grey pixels correspond to points that have been analysed. ROI with $r_{\text{max}} = 10\, \text{m}$, $H_{\text{min}} = 0.10\, \text{m}$, $H_{\text{max}} = 0.4\, \text{m}$, and $\theta = 40^\circ$. Space-variant resolution with $n = 3$, $m = 6$, $n_{\text{max}} = 30$, and $w_{\text{max}} = 20$. Note the focus on obstacle regions.

Along the same line of reasoning, instead of analysing every row of the input image, $n + w$ rows are skipped, where $w$ is incremented every time an analysed row does not contain any pixel with a positive compatibility test. To avoid large jumps $w$ is upper bounded by $w_{\text{max}}$. Whenever a compatibility test succeeds, $w$ is zeroed. This procedure intends to reduce the computational load in environments with few obstacles or when these are located mostly in the far-field. Because the truncated triangles associated with points in the near-field are quite large, skipping rows from the image’s bottom when no obstacle is found there greatly reduces the computational load. Fig. 6 illustrates the operation of the SVR on a typical input image.

Finally, every time a pixel is labelled obstacle, the saliency of all pixels encompassed by its upper truncated triangle is increased by a given percentage, $\lambda$. This reinforcement of the detected obstacle’s presence raises the chances of selecting other obstacle’s pixels for compatibility testing. The use of saliency to guide a task-specific detector is rather typical (Itti et al., 1998; Frintrop, 2006). However, the proposed model exhibits a novel characteristic to saliency-based systems by allowing the results of the detector to modulate the saliency map.

By not relying on 3-D features, visual saliency is able to guide the detector without being misguided by potential 3-D artefacts. Fig. 7 shows that this property helps in the reduction of false positive rate.
Figure 7: Typical results of the SVR with and without saliency modulation. No additional filters have been turned on. (c)-(d) Black, false positives; dark grey, true positives; bright grey, false negatives. ROI with $r_{\text{max}} = 10\, \text{m}$, $H_{\text{min}} = 0.10\, \text{m}$, $H_{\text{max}} = 0.4\, \text{m}$, and $\theta = 40^\circ$. SVR with $n = 3$, $m = 6$, $n_{\text{max}} = 30$, and $w_{\text{max}} = 20$. Without the use of saliency, a set of false positives hamper the robot from finding the passage on the right-hand side of the image.

### 6.3 Voting Filter

Stereo-based 3-D reconstruction is a rather noisy process. This characteristic is particularly problematic when small distant obstacles must be detected. In this case, the challenge is to devise a set of filters to remove the noise without hampering performance and accuracy. The direct filtering of the 3-D point cloud is a computationally expensive task. The cost comes mainly from the fact that it is not possible to know beforehand which regions of the image are important to be handled. An alternative is to perform post-filtering on the segmented obstacles. This focuses the filtering process on those 3-D patches that are known to be relevant for the overall system. However, the higher the noise level the more obstacle segments exist and consequently the more expensive their treatment is. Here, we propose a complementary mechanism: the use of a voting mechanism fully embedded in the obstacle detector. Being embedded in the detection process, the previously mentioned limitations are circumvented. On the one hand, only 3-D points that are relevant to the detector are analysed. On the other hand, removing noise during the detection process reduces the number of obstacle points to be considered by the subsequent segmentation phase.

Let $S_p$ be the set of 3-D points encompassed by the upper truncated cone emanating from the 3-D point $p$. These points are said to be voted by $p$. Conversely, let $R_p$ be the set of 3-D points whose upper truncated cones encompass $p$. These points are said to vote on $p$. See Fig. 8 for an illustration of these concepts. A direct voting mechanism would be to reject $p$ as an obstacle if the cardinality of both sets $S_p$ and $R_p$ did not reach a given threshold. However, the effects of projection make the theoretical maximum size of both sets depend on the distance $p$ is from the camera. In other words, farther obstacles are represented by fewer pixels than closer obstacles. Hence, this approach would result in a scale-variant filtering mechanism and so is inaccurate.
Figure 8: Diagrams illustrating the voting mechanism associated with a given 3-D point \( p \) and its corresponding projection on the image plane, \( p' \). (a) Illustration depicting voting relationships between a given 3-D point \( p \) and its neighbours, according to the compatibility test. (b) Projection onto the image plane of the situation depicted in (a). \(|A_p'|\) and \(|B_p'|\) correspond to the area in pixels of the left and right filled truncated triangles, respectively. If the number of pixels without computed disparity is subtracted from each of these quantities, one obtains the theoretical maximum number of times that \( p \) is able to vote and being voted, respectively.

To make it invariant to scale, the voting mechanism applies a threshold to the cardinality of the distance-normalised versions of both \( R_p \) and \( S_p \). As before, let \( p' \) be the projection of \( p \) onto the image plane. Let \( A_p' \) be the area in pixels of the upper truncated triangle emanating from pixel \( p' \) (see Fig. 8). Let \( A_p'' \) be the sub-set of \( A_p' \) with associated 3-D information. \( A_p'' \) is said to be the theoretical maximum number of points that may be voted by \( p' \) and consequently by \( p \). Similarly, let \( B_p' \) be the area in pixels of the lower truncated triangle emanating from pixel \( p' \) (see Fig. 8). Let \( B_p'' \) be the sub-set of \( B_p' \) with associated 3-D information. \( B_p'' \) is said to be an approximation of the theoretical maximum number of points that may vote on \( p \). This approximation builds on the assumption that the truncated triangles associated to the projections of all 3-D points composing the same obstacle have equal area. Because those points are close to each other, the taken assumption renders a good approximation.

The following test on the number of votes normalised by their theoretical maximum can now be used to determine whether point \( p \) is accepted as an obstacle:

\[
\left( \frac{|S_p|}{|A_p''|} > v \right) \lor \left( \frac{|R_p|}{|B_p''|} > v \right),
\]

where \( v \) is an empirically defined threshold. With the voting mechanism, compatibility between two 3-D points is no longer a sufficient condition to consider them as obstacles. Now, a higher level of robustness is attained by defining obstacles with a many-to-many relationship. With this method the detector becomes considerably resilient to the presence of 3-D artefacts (see Fig. 9) and even to the type of noise generated by a partial damage of one of the lenses composing the stereoscopic vision sensor (Santana et al., 2008).
Figure 9: Typical results of the small obstacle detector without voting mechanism (b), with voting mechanism (c), and with voting mechanism plus SVR (d). In (b) and (c) the SVR has been turned off. (b)-(d) Black, false positives; dark grey, true positives; bright grey, false negatives. ROI with $r_{\text{max}} = 10\text{m}$, $H_{\text{min}} = 0.10\text{m}$, $H_{\text{max}} = 0.4\text{m}$, and $\theta = 40^\circ$. Voting mechanisms with $v = 0.2$ and $a = 25$. SVR with $n = 3$, $m = 6$, $n_{\text{max}} = 30$, and $w_{\text{max}} = 20$. Note the considerable reduction of false positives when the voting mechanism is turned on. Note also that with SVR, computation time is saved by one order of magnitude for the same detection rate. The lack of data (white pixels) within texture-deprived obstacles is due to failure in the stereo-based 3-D reconstruction process.

6.4 Area Filter

Although the voting filter is extremely powerful, experimental results will show that its operation can be better exploited when in conjunction with an area filter (see Section 8). The area filter comes into play to remove any residual noise left by the voting filters. For this purpose, the obstacle points are first segmented in the 3-D space (Manduchi et al., 2005). The area filter then eliminates those segments whose volume projected onto the image plane is characterised by having a small area. Formally, an obstacle point $\mathbf{p}$ is re-labelled as non-obstacle if

$$|A_{\mathbf{p}}| < \frac{a \cdot 10^2}{p_z^2},$$

(6)

where $A_{\mathbf{p}}$ is the set of points composing the segment that encompasses $\mathbf{p}$ and $a$ is an empirically defined scalar. This test verifies whether the projected area of the segment is below a pre-specified area $a \cdot 10^2$, properly normalised by the squared distance to the obstacle. This normalisation procedure introduces scale-invariance into the filter.

6.5 Pitch-Roll Compensation

All the above geometrical considerations assume that the camera is not pitched nor rolled with respect to the ground-plane. A possible way of removing this constraint is to compensate for small variations on the camera’s attitude by overestimating the truncated triangles’ size.
Let $Q$ be the set of 3-D points obtained for the current scene. Let $\alpha$ and $\theta$ be the pitch and roll angles of the sensor with respect to the ground-plane, which is estimated as in Section 5. Let $R_{(\theta, \alpha)}$ be a 3-D rotation matrix built upon both pitch and roll angles. The 3-D points composing the set $P$ provided to the obstacle detector now correspond to the elements of $Q$ properly rotated by $R_{(\theta, \alpha)}$:

$$p = R_{(\theta, \alpha)}q, \quad \forall p \in P, q \in Q.$$  

(7)

This rotation aligns the stereoscopic vision sensor’s local frame of reference with the frame of reference of the ground-plane (van der Mark et al., 2007). This accommodates the 3-D point cloud for the application of the canonical compatibility test. From another perspective, the truncated cone of the compatibility test is implicitly rotated according to the attitude of the vision sensor with respect to the ground-plane. However, the projection of this transformation must also be accounted for. That is, the truncated triangle must also be rotated; otherwise some pixels would be erroneously skipped, whereas others would be unnecessarily analysed in the compatibility test.

A solution to this problem would be to rotate the back-projection of the truncated triangle’s vertexes using $R_{(-\theta, -\alpha)}$, which would then be re-projected onto the image plane to become the new vertexes used in the compatibility test. However, the resulting triangle would most probably no longer be isosceles, which would complicate the scanning procedure within it. This plus the additional projective transformations make this solution computationally expensive. For this reason, in practice only the roll angle is taken into account in this operation. This approximation draws from the empirical observation that this rotation is essential for a proper operation of the small obstacle detector, whereas the disregard of the pitch angle only results in missing the top of some obstacles. Bearing this in mind, the vertexes of the rotated triangle associated to $p$, $v''_l$ and $v''_r$, are obtained by rotating the non-rotated triangle’s vertexes, $v'_l$ and $v'_r$, so as to compensate for the ground-plane’s roll angle:

$$v''_i = R_{(-\theta)}(v'_i - p') + p',$$  

(8)

where $i \in \{l, r\}$ and $R_{(-\theta)}$ is a 2-D rotation matrix. Note that the ground-plane’s roll angle is assumed to project directly onto the image plane, which is an approximation that was found to be sufficient in practice (see Fig. 10).

### 7 Parameter Selection

This section provides guidelines regarding the instantiation of every free parameter included in the model. The ROI definition requires the specification of a minimum range, $r_{min}$, and a maximum range, $r_{max}$. Whereas the former is constrained by the sensor’s field of view, the latter is upper bounded by the point cloud’s noise level. For instance, the shorter the baseline
Figure 10: Roll compensation in the image plane. (a) Illustrative diagram, where the truncated triangle is rotated in order to compensate for a roll angle $\theta$. (b) Typical real situation in which the rotation of the truncated triangle is the result of the roll compensation mechanism. The zoomed image depicts the results of the compatibility test associated with the pixel in the truncated cone’s vertex, where red (darker), green (lighter), and black pixels are compatible, incompatible, and without associated 3-D information, respectively.

of the stereoscopic vision sensor, the smaller the $r_{\text{max}}$. The additional parameter $\varsigma$ is an upper bound in the expected proportion of pixels encompassed in a row that are erroneously beyond $r_{\text{max}}$. A value of 15% is usually sufficient.

The number of iterations $m_{\text{refit}}$ and $n_{\text{hypo}}$ in the ground-plane estimation RANSAC procedure could be theoretically defined, provided that the distribution of the sampling space was known. Because this is not the case on all-terrain environments, these parameters must be defined empirically. A large number of iterations is required only if a large number of obstacles are expected to occur in the scene. The related parameter $d_{\text{plane}}$ defines an upper bound for the distance between a point and a plane hypothesis to be accepted as an inlier. On uneven ground, a small value ($<10\text{cm}$) has the negative effect of fitting the ground-plane estimate to small planar patches of the terrain. In practice, larger patches of the terrain can be covered and consequently more robust estimates attained if a larger value is considered.

The number of non-salient pixels is used in the RANSAC procedure to estimate the discriminative power that saliency has in the current scene. In the process, a small threshold $l$ is used to determine which pixels are non-salient. A small value above zero is usually sufficient for a proper instantiation of this parameter. A measure of the local saliency is used in the ground-plane estimation process to determine which points should be taken into account. The larger the width of the window used to compute local saliency, $q$, is, the more difficult it is to accept points near salient regions. Practice suggests a value below 25% of the image’s width to properly handle cluttered environments. The selection pressure is also controlled by the scaling factor $\rho$. Higher values reduce the chances of selecting non-salient pixels at the cost of a higher number of required samples. Practice suggests that a scaling factor of around 4 manages the trade-off well.

The small obstacle detector also requires the specification of a set of parameters. $H_{\text{min}}$ and $\theta$ are obtained directly from the robot’s mechanical properties, whereas $H_{\text{max}}$ equals $h$ in the hybrid framework. That is, $H_{\text{max}}$ is as high as the smallest obstacle to be detected by the large
obstacle detector. SVR parameters $n_{\text{max}}$ and $w_{\text{max}}$ must also be instantiated. Parameter $n_{\text{max}}$ trade-offs speed with the risk of failing to detect a non-salient obstacle in the coarse analysis of the image. Because large obstacles are those that may not be salient in the image, $n_{\text{max}}$ can be safely instantiated to a large value. Not being sensitive to saliency information, the parameter controlling the maximum number of rows that can be skipped in the coarse analysis of the image, $w_{\text{max}}$, must be at least as small as the expected height in pixels of obstacles detectable at a safe distance. The same reasoning applies to the variable controlling the coarse analysis, $m$. In practice these variables are tuned to achieve the required computational performance under a given accuracy constraint. An example of such constraint is the expected projected width onto the image plane, $\Delta x'$, of an object with a given width $\Delta x$ and located at a given distance $z$ from the sensor. This can be expressed as $\Delta x' = \frac{f}{o} \Delta x$, where $f$ is the sensor’s focal length and $o$ is the pixel’s width. For instance, for a camera with focal length of 4.4 mm and pixel size 0.006 mm, a 20 cm obstacle at 5 m from the robot will have a corresponding size in the image plane of roughly 30 pixels. This means that both $w_{\text{max}}$ and $m$ must have smaller values to ensure that the obstacle is sampled. Parameter $\lambda$, used to detect ascending saliency variations, is empirically minimised under the constraint that the scanning procedure must not get trapped by small local variations resulting from noise in the input image. A good indicator for its parameterisation is 10% of the saliency’s range.

Both voting and area filters are also dependent on two parameters, namely $v$ and $a$. Their values depend mostly on the level of noise needed to be handled. As experimental results will show, by varying these thresholds it is possible to smoothly move on the trade-off curve relating false positive rate and true positive rate. This robustness will also be confirmed by the ability of using the same thresholds in different environments. In practice, by making $\nu \approx 0.2$ and $a \approx 25$ most of the noise impinging stereo-vision is filtered without significant signal loss.

8 Experimental Results

This section presents the experimental results obtained with two different data sets composed of $640 \times 480$ image pairs. The first data set is composed of 36 heterogeneous stereo image pairs (see Fig. 18 in Appendix A) that have been acquired with a 9 cm baseline Videre Design STOC sensor under dynamic conditions and at an approximate height of 1.5 m. This set will be used to quantitatively assess the small obstacle detector and the ground-plane estimation technique. For this purpose all images have been hand-labelled (obstacle/non-obstacle pixels) for ground truth. To reduce imprecision in the hand-labeling process, the ROI of this data set has been limited to 10 m and images were selected so that a dominant ground-plane was actually present and vegetation as much absent as possible. This allows an accurate quantitative analysis of the several components of the system.

To test the hybrid detector, a data set with images containing large obstacles and considerable uneven terrain is required. This data set should also allow the test of the method against more distant obstacles. Finally, the data set should be extensive and disparate. All these aspects have been considered in the second data set, which encompasses three long runs with 798, 998
and 600 frames, obtained at 7.5 Hz in off-road, urban, and mixed environments, respectively. With the stereo head hand-held at an approximate height of 1.5 m, the acquisition process took place by walking in the environments at an average speed of approximately 1 m s\(^{-1}\). To allow the detection of obstacles up to 20 m a sensor with a baseline of 30 cm was used. Every 20 images were hand-labelled to enable a quantitative analysis. The higher complexity of this data set when compared to the previous one results in more ambiguous hand-labels. This explains why the components of the system are first tested individually and thoroughly with the previous data set.

Small Vision System (SVS) (Konolige, 1997; Konolige and Beymer, 2007) and Open-CV (Bradski and Kaehler, 2008) were used for stereo computation and other low-level computer vision routines, respectively. Stereo computation uses an area-based L1 norm (absolute difference) correlation method, operating over Laplacian Of Gaussian (LOG) transformed images. The result is interpolated to a precision of 1/4 pixel and the correlation window size is 11 × 11. To increase the amount of information available in the point cloud, the disparity calculation is carried out at the original resolution, and also on images reduced by 1/2. With this multi-scale approach, the extra disparity information is used to fill in dropouts in the original disparity calculation.

SVS also provides a set of standard filters to reject 3-D points that are potentially erroneous at the cost of reducing too much the density of the 3-D point cloud in poorly textured environments. Briefly, according to a threshold \( f_c \), a confidence filter eliminates stereo matches that have a low probability of success due to lack of image texture. A uniqueness filter performs a consistency check to ensure that the minimum correlation value is lower than all other match values by a threshold \( f_u \). Finally, a speckle filter eliminates small disparity regions that are not correct by imposing a threshold \( f_s \) on the minimum region size. The three filters are used in both data sets with \( f_c = 12, f_u = 10, \) and \( f_s = 400 \). The ability of the model to handle noisier point clouds will be demonstrated in a final experiment, in which the strength of both confidence and uniqueness filters will be reduced, \( f_c = 6, f_u = 6 \). With this reduction, the point cloud is denser, but also noisier, and so the overall results of our model are more significant.

The following summarises the parameterisation of the model used in the experiments, which has been defined according to the guidelines presented in Section 7. Robot related parameters, \( H_{\text{min}} = 0.1 \text{ m}, H_{\text{max}} = h = 0.4 \text{ m}, \theta = 40^\circ \), ground-plane estimation parameters, \( m_{\text{refit}} = 10, n_{\text{hypo}} = 40, d_{\text{plane}} = 0.2 \text{ m}, l = 5, q = 150, \rho = 4, \) and large obstacle detection threshold, \( h = 0.4 \text{ m} \), are stable across experiments. To account for the different sensors’ baselines, the ROI is parameterised with \( r_{\text{min}} = 1 \text{ m}, r_{\text{max}} = 10 \text{ m}, \zeta = 15\% \) and \( r_{\text{min}} = 2 \text{ m}, r_{\text{max}} = 20 \text{ m}, \zeta = 15\% \) for the first and second data sets, respectively. Unless otherwise noted, SVR of the small obstacle detector and related filtering mechanisms are set as \( n = 3, m = 6, n_{\text{max}} = 30, w_{\text{max}} = 20 \) and \( v = 0.2, a = 25, \) respectively. The specific values of \( v \) and \( a \) will be justified according to the results obtained with the small obstacle detector in the first data set.
8.1 Ground-Plane Estimation Results

The first set of experiments intends to demonstrate the usefulness of using saliency to modulate the ground-plane hypotheses generation step. Ground-truth is given in terms of obstacle/non-obstacle hand-labels, rather than in terms of ground-plane coefficients. This option results from the fact that the hand-labelling of the obstacles is a much more accurate process than the process of hand-labelling the ground-plane coefficients that better approximate the terrain. This compels us to assess the ground-plane estimation process by indirect means. Concretely, validation is done by comparing the obstacles detected using the plane-based detector with the obstacles present in the ground-truth. The better the estimated ground-plane, the closer the detected obstacles match the ground-truth. This process is repeated with and without using saliency to modulate the ground-plane estimation process.

A large set $M$ of 10000 ground-plane hypotheses per image in the first data set was created with the Saliency-based Ground-Plane Estimation (SalGPE) approach. This large set results in statistics varying $\approx 1\%$ across experiments. A set $U$ with the size of $M$ was created for each image as well, but this time without saliency modulation and thus representing the canonical RANSAC-based Ground-Plane Estimation (GPE). The set of obstacles detected using each ground-plane hypothesis is compared against the ground-truth to obtain the True Positive Rate (TPR), the False Positive Rate (FPR), and the two-class Matthews Correlation Coefficient (MCC). The MCC metric is well known for its ability to handle unbalanced data sets. The closer MCC is to 1 the better the hypothesis matches the ground-truth. Obstacles are those points whose orthogonal distance to the plane is above 0.2m, which is a reasonable upper bound for most wheeled robots. A lower value would be inappropriate for a plane-based detection approach. Then, the mean ($\mu$) and standard deviation ($\sigma$) of the above variables over all hypotheses in both $M$ and $U$ are computed.

According to the MCC results (see Table 2.1(a)), two main image sub-sets emerge. One (grey shaded) aggregates images in which the RANSAC saliency-based hypotheses generation step outperforms the canonical one (refer to the two last columns in Table 2.1(a)). The MCC differences are residual (i.e., $< 5\%$) for the remainder images, meaning that saliency is essentially neutral there. Images without obstacles (13 and 18) have MCC values of 0. Images benefiting from saliency share a characteristic: a considerable presence of objects. In these situations (e.g., Fig. 2), saliency easily segments objects from background. Saliency thus operates better in those situations in which it is most required. In the absence of obstacles an uninformed solution suffices.

Fig. 11(a) depicts the Receiver Operating Characteristic (ROC) results of the experiment. For each image $k$, an arrow is drawn to connect each without-saliency point, $(\mu^U_{TPR}, \mu^U_{FPR})_k$, to its corresponding point with-saliency, $(\mu^M_{TPR}, \mu^M_{FPR})_k$. The closer a point is to the upper-left corner of the graph the better the corresponding set of sampled hypotheses matches the ground truth. A clear dominance of arrows heading towards the upper-left corner is observed.

However, the arrow corresponding to image 23 in Fig. 11(a) goes in a clearly different direction. The justification for this fact relates to the particular configuration of the environment, whose effects on the analysis are the following. Without saliency modulation, the estimated
the effects are the opposite. That is, the better approximation of the estimated plane to the actual

plane is raised slightly off the ground, influenced by the large obstacle in the image. As a result,
by not being fully above the estimated plane, the obstacle induces a low TPR. Moreover, by
being below the estimated plane, some false positives do not contribute to the FPR. Conversely,
because the obstacle does not affect the fitting process when saliency modulation is employed,
the effects are the opposite. That is, the better approximation of the estimated plane to the actual
ground-plane results in that surface variations induce a higher FPR, and a bigger portion of the
obstacle contributes to an also higher TPR.

Figure 11: Ground-plane estimation and obstacle detection ROC plots. For a given image, each arrow connects the TPR/FPR trade-off point obtained without saliency modulation to the TPR/FPR trade-off obtained with saliency modulation. (a) Ground-plane estimation with (SalGPE) vs. without (GPE) saliency modulation. (b) Obstacle detection with (SalOD) vs. without (OOD) saliency modulation. (c) Same as (b) but skipping $n_{\text{slide}}$ pixels for the OOD case. Line $x = y$ displayed for reference.

8.2 Small Obstacle Detector Results

To isolate the saliency modulation capabilities, a stripped down version of the small obstacle detector was tested in an initial experiment. For this purpose, the voting filter was turned off, the maximum number of pixels and rows that can be skipped in the coarse analysis was fixed to 30 and 1, respectively. This means that the skipping procedure was not progressive.

Fig. 11(b) shows the results obtained from comparing the OOD (Talukder et al., 2002; Manduchi et al., 2005) with the stripped down, saliency-based detector (SalOD), i.e., with $(n, m, n_{\text{slide}}) = (1, 2, 30)$. When compared to the OOD, the SalOD exhibits a considerably reduced FPR and only a slightly smaller TPR. This combined result shows the benefits of using saliency to selectively discard non-obstacle points. Moreover, it should be noted that the still undesirable level of TPR reduction is owed in part to the absence of progressive skipping in the SalOD. This signal-to-noise ratio improvement is confirmed by the contrast of the MCC with $(MCC_p^M)$ and without $(MCC_p^U)$ saliency modulation (see Table 2.1(b)). Label $p$ means that ground-plane compensation is on. Saliency contributes in the same amount when the ground-plane compensation is off (see Table 2.1(c)), highlighting its resilience.

To reinforce the evidence that the reduction in FPR is due to the saliency’s selective nature instead of the reduced number of pixels being analysed, an additional test was carried out. The OOD was configured to systematically skip $n_{\text{slide}}$ pixels, rather than $n$, when displacing the
truncated triangle. In this situation, the results (see Fig. 11(c)) show that although the OOD now produces a smaller FPR, due to blindly skipping more pixels, a considerable reduction in the TPR is also observed. As for obstacle detection, smaller TPR means higher risk of collision, saliency shows itself to be a useful cue for informed false positive removal.

To test the full-fledged small obstacle detector (ESalOD), all its features were turned on, including both voting and area filters. Fig. 12 plots the ROC curves of this experiment. A first analysis shows that ESalOD with \((n \times m) = (3 \times 6)\) exhibits a TPR vs. FPR trade-off at least as good as the trade-off exhibited by the OOD. This stems from the observation that the ESalOD ROC curve intercepts the OOD ROC point. The ROC curve also shows that for high values of TPR the ESalOD exhibits a better TPR vs. FPR trade-off than the OOD. This demonstrates how advantageous it is to embed into the detector both visual saliency and voting filter. The curve shows, for instance, that it is possible to reduce the FPR of the OOD by \(\approx 70\%\) while diminishing the TPR by only \(\approx 10\%\).

The area under each ROC curve obtained with either the voting or the area filters alone is smaller than the area obtained with both filters operating in conjunction. This shows that the latter configuration exhibits a better overall performance. Note that for low FPR values (i.e., \(< 0.15\)) none of the filters alone is capable of approaching the ROC curve obtained with both filters operating simultaneously. This clearly shows their complementary role. This is further confirmed by the ROC curve of the area filter (see Fig. 13), where the higher the area under the curve, the more active the area filter is.

From the analysis of the area filter ROC curve (see Fig. 13) we conclude that the config-
Figure 13: Impact of the area filter. Each plot represents the average of the ROC curves obtained over all images in the first data set, for a given area filter configuration. For a given image and filter’s parameterisation, $a \in \{0, 5, 25\}$, the ROC curve is built by sliding the threshold of the votes filter over its domain, $v \in \{0, 0.05, \ldots, 1\}$. The absence of the area filter, $a = 0$, results in the lowest area under the curve, which shows the usefulness of the filter. The relative performance of the other two parameterisations, $a \in \{5, 25\}$, swaps at the intersecting point of their ROC curves. Nevertheless, $a = 25$ is shown to be the most interesting configuration as it is the one that performs better for higher TPR values.

uration $a = 25$ (see Equation 6) is the most adequate for the first data set. After a thorough inspection of the model’s behaviour in the first data set we also conclude that the configuration $v = 0.2$ (see Equation 5) for the voting filter shows the best performance. With this configuration, false negatives are mostly absent and the false positives are mostly concentrated around obstacles, resulting only in their enlargement (see Fig. 9).

8.3 Hybrid Obstacle Detection Results

Having confirmed the performance of the model’s individual components, this section tests the hybrid solution as a whole. For this purpose, the more extensive second data set is used. The terrain in this data set is much more uneven, and large obstacles at different ranges are much more frequent, which will help highlighting the advantages of focussing each detection technique on a specific type of obstacle. As will be shown, this specialisation promotes a better TPR vs. FPR trade-off as well as a reduced computational cost. The ROI in these experiments has been changed to $r_{\text{min}} = 2 \text{ m, } r_{\text{max}} = 20 \text{ m}$.

Table 2 summarises the quantitative results obtained for the three runs composing the data set. The table’s first row shows the ability of the model to maintain a high TPR with the parameterisation obtained from the first data set. This is an exhibition of the model’s robustness to different sensors and environments. The second row shows that with a less filtered disparity
Table 2: Results obtained with the hybrid detector (mean ± standard deviation). Each row corresponds to a given configuration. The small obstacle detector in the “base” configuration is parameterised as for the first data set. The “+ disp.” configuration takes the same parameterisation but with a much denser/noisier disparity map. The “+ filters” configuration adds to the previous configuration a stronger filtering mechanism, $v = 0.6$, $a = 70$. The last configuration adds to the previous configuration a larger upper bound to the saliency-based pixel skipping procedure, $n_{max} = 100$.

map, and consequently with a denser and noisier point cloud, the number of false positives grow for both off-road and mixed environments. But interestingly the TPR grows as well. This shows the robustness of the model to changes in the stereo-based 3-D reconstruction process. If the growth in FPR is nevertheless undesirable, which depends on the model’s client, one can reduce it by empirically pushing further the voting and area filters to $v = 0.6$, $a = 70$. The third row of the table shows exactly this. See for instance that the FPR is reduced on average by 55% with only a TPR average reduction of 7%.

To assess the sensibility of the model to the mechanism that most strongly constrains the saliency-based SVR, $n_{max}$ was extended from 30 to 100 pixels. The results in the last row of the table show that no significant difference can be observed with this new parameterisation. This means that the saliency map is accurate enough to guide the detector. This last configuration of the hybrid detector will be used in the remaining experiments. The videos with the results overlaid are available in the authors’ website\(^1\). The videos permit the qualitative verification of the model’s ability to detect the vast majority of the obstacles present in the several tested environments. Moreover, the videos also show that most of the false positives are not stable across frames, and consequently they should be easily filtered out under a probabilistic mapping framework.

Table 3 summarises the results obtained with both small and large obstacle detectors in isolation. The goal of these experiments is to study the contribution of each technique to the overall hybrid model. The first row of the table shows that the small obstacle detector alone is not able to obtain the TPR level of the hybrid solution (cf. bottom row of Table 2). This is mostly because it fails to detect some large obstacles whose 3-D point cloud is sparser or noisier (e.g., Fig. 14(a)). This is a limitation of methods that identify obstacles based on geometrical relationships between neighbour 3-D points. Sometimes, the small obstacle detector alone also fails to label as obstacle those gentle slopes that due to their height are not navigable (e.g.,

---

\(^{1}\)Videos with results overlaid at [http://www.uninova.pt/~pfs/index/ODVideos.html](http://www.uninova.pt/~pfs/index/ODVideos.html)
Table 3: Results obtained with isolated detectors (mean ± standard deviation). Each row corresponds to a given configuration. The “small” configuration considers only the output of the small obstacle detector as considered in configuration “+ n_{max}” (see Table 2). The “h = 0.4 m” configuration considers only the large obstacle detector. The last configuration considers only the large obstacle detector, but this time with h = 0.1 m and also encompassing negative obstacles.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Urban Run TPR</th>
<th>Urban Run FPR</th>
<th>Off-Road Run TPR</th>
<th>Off-Road Run FPR</th>
<th>Mixed Run TPR</th>
<th>Mixed Run FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>0.63 ± 0.23</td>
<td>0.04 ± 0.03</td>
<td>0.62 ± 0.24</td>
<td>0.01 ± 0.01</td>
<td>0.63 ± 0.28</td>
<td>0.05 ± 0.05</td>
</tr>
<tr>
<td>h = 0.4 m</td>
<td>0.81 ± 0.17</td>
<td>0.04 ± 0.03</td>
<td>0.80 ± 0.21</td>
<td>0.02 ± 0.02</td>
<td>0.79 ± 0.25</td>
<td>0.04 ± 0.06</td>
</tr>
<tr>
<td>h = 0.1 m</td>
<td>0.93 ± 0.06</td>
<td>0.13 ± 0.13</td>
<td>0.95 ± 0.07</td>
<td>0.19 ± 0.16</td>
<td>0.92 ± 0.13</td>
<td>0.17 ± 0.14</td>
</tr>
</tbody>
</table>

In conclusion, the two techniques are shown to perform in a complementary way, thus justifying the benefits of a hybrid solution. By making h = 0.1 m, the ability of the plane-based solution to detect both small and large obstacles was tested in a final experiment. As this configuration intends to also substitute the use of the small obstacle detector, it must also be able to detect negative obstacles. Therefore, in order to allow that points below the ground-plane would also be considered obstacles in this test, the distance of 3-D points to the ground-plane is considered in absolute terms. The third row of Table 3 shows that with this configuration the FPR grows approximately 44 times the growth of TPR (e.g., Fig. 14(d)). See, for instance, the case of the off-road run, in which an 850% increment of FPR is followed by only a 19% increment of TPR. Therefore, the hybrid solution exhibits the best TPR vs. FPR trade-off.

Typical outputs of the hybrid obstacle detector in the off-road, urban, and mixed environments can be depicted in Fig. 15, 16, and 17, respectively.

8.4 Computational Performance

Table 4 summarises the computation time spent at each step of the proposed model. It also compares the computation time of the hybrid detector with the computation time taken by the OOD, proposed by Talukder et al. (2002) and Manduchi et al. (2005). These results were obtained with a Linux-based 2.8GHz Intel Core 2 Duo Laptop equipped with 4GB of RAM. The model runs on a single core.

The hybrid obstacle detection takes on average 183ms on the second data set, as opposed to the 294ms obtained with the first data set. The 61% increment of computation time relates to the different ROI used in each data set, which is in turn a consequence of using sensors with different baselines. When covering larger areas of the near-field, the truncated triangles
Figure 14: Comparison between hybrid (middle row) and isolated (bottom row) detectors over a set of typical input images (top row). Results are organised in a columnwise manner, each column being associated to a given configuration (see Table 3). Detected obstacles are overlaid in dark grey over the corresponding input images. White pixels are those out of the ROI or without computed 3-D information and are thus discarded by the detection process.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Stereo</th>
<th>Saliency</th>
<th>Plane</th>
<th>Detection</th>
<th>Total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (9 cm)</td>
<td>44 ± 1</td>
<td>51 ± 4</td>
<td>31 ± 7</td>
<td>168 ± 106</td>
<td>294 ± 107</td>
<td>6464 ± 4074</td>
</tr>
<tr>
<td>2 (30 cm)</td>
<td>59 ± 1</td>
<td>53 ± 7</td>
<td>25 ± 5</td>
<td>46 ± 25</td>
<td>183 ± 26</td>
<td>3198 ± 909</td>
</tr>
</tbody>
</table>
Figure 15: Typical results obtained with the proposed hybrid model in the off-road long run.

Figure 16: Typical results obtained with the proposed hybrid model in the urban long run.
9 Conclusions

A model for stereo-based all-terrain obstacle detection was presented. By hybridising two complementary obstacle detection techniques, the model innovates at the architectural level. A common characteristic of the two techniques is that they perform in the image space, rather than in a digital terrain map.

The presented hybrid system is capable of searching for obstacles with more than 10cm height up to a range of 20m on uneven terrain. It performs at 5Hz on 640 × 480 images. This has been attained by focussing each detection technique on a particular type of obstacle (i.e., large vs. small) depending on each technique’s characteristics. Also novel at the architectural level is the extensive use of visual saliency to guide the detection process. This mechanism was shown to improve robustness, accuracy, and computational efficiency of obstacle detection.

The technique employed for small obstacle detection is known for its ability to perform well on uneven terrain. However, in its classic form, it is also known to be computationally...
expensive and brittle in the face of noise. These limitations have been overcome in the proposed model with the novel use of saliency-based space-variant resolution, with a mechanism for the camera’s pitch-roll compensation, and with a voting mechanism. The latter was shown to be extremely powerful in enabling the operation of the detector in the presence of noisy 3-D point clouds. The fact that the filter is fully embedded in the detector’s operation, almost discards the need for time consuming post-filters.

The revealed success of visual saliency in all-terrain obstacle detection is in part due to the novel way conspicuity maps are blended in this work. Another key aspect is that the detector updates the saliency map whenever an obstacle is detected. This allows the saliency map to guide the detector while being opportunistically corrected by it.

With the exception of the compatibility test, all other aspects of the proposed model are easily parallelisable. This opens the door for future Graphics Processing Unit (GPU)-based implementations to further reduce processing cycle. As a further improvement, the number of independent parameters shall also be reduced. We also expect in the future to assess the benefits for obstacle detection of modulating the saliency map with top-down knowledge, such as the expected appearance of most frequent obstacles.

Acknowledgements

This work was partially supported by FCT/MCTES grant No. SFRH/BD/27305/2006. We thank Paulo Santos for his participation in the early developments of the obstacle detector. We also acknowledge the useful comments provided by Nelson Alves and by the four anonymous reviewers of this paper.
Figure 18: Stereo data set (left images only) obtained with a 9cm baseline configuration.
References


R. Labayrade, D. Aubert, and J. P. Tarel. Real time obstacle detection in stereovision on non flat road geometry through v-disparity representation. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 646–651. IEEE, June 2002.


