Robust ground plane detection and tracking in stereo sequences using camera orientation

Paul Herghelegiu, Adrian Burlacu and Simona Caraiman
Faculty of Automatic Control and Computer Engineering
Technical University “Gheorghe Asachi”
Iasi, Romania
{pherghelegiu | aburlacu | sarustei}@tuiasi.ro

Abstract— **The research on assisting visually impaired people to navigate in unknown environments commonly employs image processing or computer vision techniques. One key element that needs to be identified in the surrounding environment of a blind user is the ground plane. Its accurate detection is highly important because the user is able to move freely in that area. An assistive device should primarily assist the user in avoiding the obstacles that lie on the ground surface. The images that will be further processed are usually acquired using depth acquisition devices worn by the user. Therefore, in contrast with automotive or robotics applications, which also heavily rely on ground plane extraction for obstacle detection, the camera orientation has more degrees of freedom. This leads to the requirement of designing more complex solutions for the ground plane detection in the case of assistive systems for the visually impaired. In this paper we introduce an algorithm to detect the ground plane taking into consideration the orientation of the camera, namely its pitch and roll. The proposed algorithm is based on an efficient processing of the v-disparity map associated to each frame. A two-step decision making approach to determine the most suited area that corresponds to the ground plane is described. Experimental results with synthetic datasets are provided to prove the efficiency and robustness of the proposed approach.**

Keywords—visually impaired people; assistive technologies; ground plane detection.

I. Introduction

Nowadays, a great importance is given to developing assistive technologies to help the visually impaired people deal with their every-day life. Such technologies focus on providing visual environmental data to the users by means of other senses. These include the haptic sense when data is passed to the user using various body parts like the tongue, the torso, the hand or others. Another way of providing data to the users is by using the auditory sense. This means that visual information is transformed into different sounds.

No matter the way information is provided to the visually impaired person, most sensory substitution approaches employ image processing techniques. Information about the surroundings of the users is acquired using different cameras. This information is processed and transformed into data that the visually impaired person can understand. As the surrounding environment typically contains a considerable amount of data, a very important step of the image processing algorithms is to filter the data, i.e., decide which information should be provided to the user and which should be ignored.

A visually impaired person uses a white cane to navigate. The cane is about 2 meters long and it is used to acquire information about the elements that are in the front of the user, at the ground level. This information is of significant importance as the main objective of such a user is to navigate without hitting an obstacle. Knowing that there is no obstacle in front allows the user to navigate freely in that area. Therefore, the ground plane located in front of the user is of significant importance. Most algorithms that aim to transform the visual data into information that can be understood by the visually impaired persons employ a ground plane detection step. Accurate ground plane detection is also important in higher-level image processing techniques like object tracking algorithms.

In this paper we present an algorithm for ground plane detection in stereo sequences. The algorithm is tailored for assistive applications for the visually impaired, which target a 3D reconstruction and segmentation of the environment.

Typically, the cameras that acquire visual information about the surrounding environment are worn by the users. There are approaches where the cameras are worn on the head but also on other body parts like the torso. Regardless of the camera positioning with respect to the user, the user movement and orientation while navigating in the environment introduces a significant variation of the ground pose in the acquired images. This is an important distinction from other applications where ground plane detection is attempted, like autonomous driving cars or autonomous robots. The proposed algorithm incorporates information about the roll and pitch rotations of the camera in order to infer valuable information about the ground plane orientation.

Real-time computation is an essential requirement for the targeted application. Thus, the proposed algorithm processes the images acquired from the environment in a 2D space. While processing the 3D point cloud built for each frame usually provides more reliable information, it is also more computationally demanding. In order to cope with the misdetections associated to 2D-based methods, we introduce a temporal coherence approach. This means that detecting the ground plane in one frame is correlated with the detection results from the previous one.

In this paper we aim to adapt the ground plane detection approaches successfully applied for automotive and robotics environments to perform fast and accurate ground plane extraction for the development of assistive devices for the visually impaired. We define a robust method that can be applied in complex outdoor environments, where more heterogeneous structures and objects are present.

For testing purposes, we used a series of 3D virtual environments specially designed to generate benchmark stereo sequences for the development of human assistive applications. Since virtual scenes can provide ground truth information, testing using such synthetic data can offer valuable information about the efficiency of the algorithm and acknowledge worst-case scenarios. Moreover, different environment scenarios can be tested without the need to physically find these locations or recreate some special situations in real life environments.

The paper is structured as follows: Section 2 describes the related work. In Section 3 the details of our approach are presented. Evaluation results are discussed in Section 4. The paper is concluded in Section 5.

II. Related Work

Ground plane detection in stereo sequences has been the subject of many research works performed in various application fields, ranging from robotics, to automotive and assistive technologies, in the context of environment sensing and understanding. In most of these applications, the identification of the ground plane represents an essential step towards achieving robust obstacle detection methods.

In indoor environments, a navigable floor map can be constructed by detecting planar surfaces, and ground plane in particular, in the data acquired with RGBD sensors (e.g., Kinect) [1-4]. Most of the approaches exploit a form of normal estimation and decomposition for the acquired point cloud, followed by clustering to detect the planar surfaces. The ground plane can then be estimated as being the largest planar surface under some orientation and positioning constraints.

The reliability of RGBD data allows for more accurate normal computation for the acquired point cloud than in the case of stereo sensors. These sensors produce depth maps containing more noisy records. Thus, plane segmentation in stereo depth maps requires distinct techniques that are able to cope with this level of noise in the measurements. Model fitting techniques, such as RANSAC or Hough transform, are usually employed to fit the (ground) plane model in noisy data.

In the context of assistive systems, in most approaches, the ground plane information is either explicitly extracted from the 3D data stream [5-7]. It also can be implicitly exploited like in Saez's approach [8] where the 3D model of the environment is permanently aligned with the horizontal axis. There are two main approaches to achieve the explicit detection of the ground plane: processing the point-cloud associated to the disparity/depth map [5,6] or detecting the ground plane directly in the disparity domain [7]. While the point-cloud processing solution is completely invariant to camera tilting, the disparity domain approach is less computationally demanding.

Rodriguez et al. [6] designed an obstacle detection system based on 3D ground detection. Ground plane detection is achieved using a model fitting technique, i.e., RANSAC. This is a global approach, in contrast with Bujacz's local approach, which exploits the similarity between neighboring patches computed for the point cloud, to iteratively group them into larger surfaces [5]. Mattoccia et al. also employed a RANSAC approach for ground plane detection in the v-disparity domain [7].

V-disparity image processing has recently become a very popular approach for ground plane estimation, particularly for road extraction in automotive applications [9-13]. The v-disparity image can be understood as the disparity histogram of each row. The major planar surfaces in the scene have corresponding line representations in the v-disparity images. Vertical surfaces are mapped into vertical line segments in the v-disparity image, while the ground plane corresponds to a slanted line segment. A Hough transform parameterization of the v-disparity image can then be employed to detect the most significant line segment. The approach relies on the assumption that the ground plane represents the dominant planar surface in the scene. Several other assumptions are exploited for ground plane estimation in automotive applications: distinct road features are present on the ground [9], the ground is well represented in the acquired image and clear of obstacles near the vehicle [12].

Also, a very important distinction of assistive applications, which usually exploit a head-worn camera acquisition system, is the higher number of the degrees of freedom for the camera motion. The slope and position of the ground plane line in the v-disparity image are directly connected to the camera orientation. Under stable conditions these parameters can be inferred during an on off-line calibration stage. Then under driving conditions the ground plane line must be one of those parallel to the line defined under stable conditions and within a certain disparity range.

The constraints considered for the ground plane detection using v-disparity map approaches for automotive applications do not usually comply with assistive devices designed for visually impaired people. The algorithm proposed in this paper defines a ground plane tracking procedure to account for the situations in which the ground plane is poorly represented in the image or even not visible at all. These situations correspond to the cases when the ground is occluded, e.g. the user is facing a wall, or not present in the field of view, e.g. the user is facing up. In such situations, the traditional v-disparity processing approaches incorrectly detect the ground plane by associating it with the largest visible planar surface, i.e., the wall, or even are unable to detect it at all. Also, the proposed algorithm differs from the existing approach [12] by using new information, such as the stereo camera roll, to provide a more accurate ground plane.

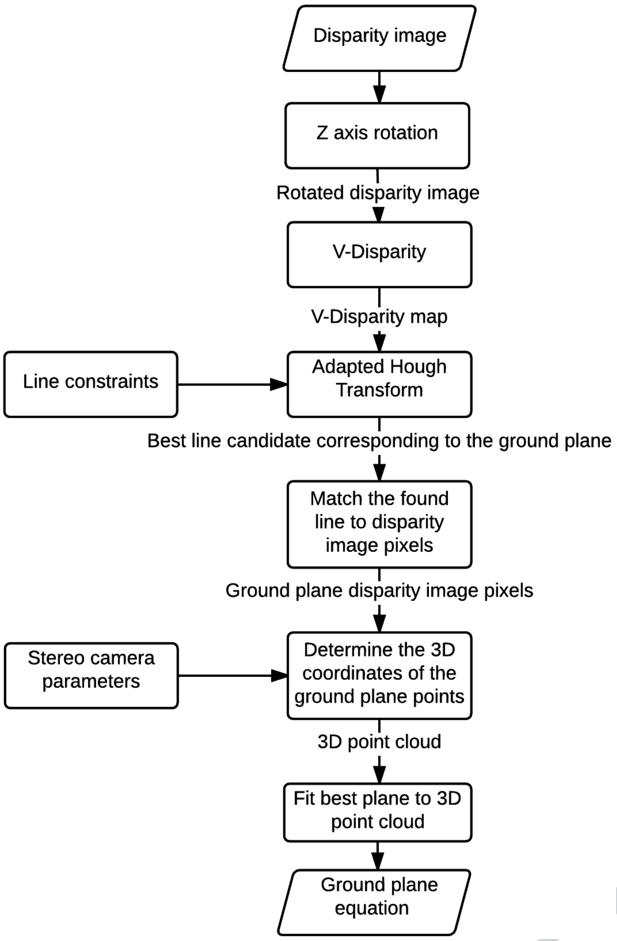
Disparity image
Z axis rotation
Rotated disparity image
V-Disparity
V-Disparity map
Line constraints
Adapted Hough Transform
Best line candidate corresponding to the ground plane
Match the found line to disparity image pixels
Ground plane disparity image pixels
Stereo camera parameters
Determine the 3D coordinates of the ground plane points
3D point cloud
Fit best plane to 3D point cloud
Ground plane equation

Figure 1: Overview of the proposed ground plane detection algorithm.

III. Method Description

The main steps of the algorithm introduced in this paper are presented in Figure 1 and are detailed in the followings. The algorithm takes as input a disparity map that has to be previously computed.

Our ground plane detection technique uses a v-disparity representation associated to the disparity image [14]. The v-disparity map is an image with the same number of rows as the disparity image. The number of columns is represented by the maximum disparity value in the image. Each row of the v-disparity map represents a histogram of the corresponding row in the disparity image. The number of pixels with a specific grayscale value is encoded by the intensity of the pixel of the histogram.

The main idea of using a v-disparity map is that the ground plane is represented by a line or a line segment in the v-disparity map. However, this is true only for the images acquired with no camera roll. If camera roll is present, the ground plane detection using the v-disparity map is not straightforward anymore. This is because the ground plane would not be represented as a straight line in the v-disparity map. Instead, it generates a region of points in the v-disparity map (Figure 2, top). This will cause the detection of the ground plane to fail. Therefore, an image acquired with a camera roll has to be rotated if to be used in the ground plane detection algorithm. This represents the first step of the algorithm we propose, i.e. rotating the disparity image along the z-axis with the angle given by the camera roll, but in the opposite direction. Computing the v-disparity on the rotated image along the z axis (roll angle of the stereo camera system) makes the ground plane to appear as a straight line on the v-disparity map (Figure 2, bottom).

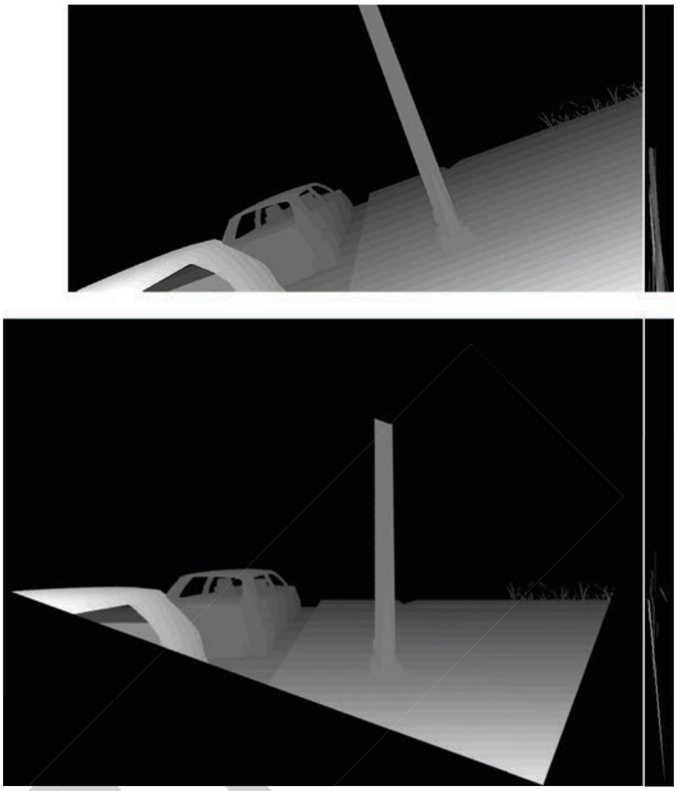

Figure 2. Top: disparity map and its corresponding v-disparity of the original image. Bottom: disparity map and v-disparity after image is rotated along the z axis, around its centre (Note: all images have adjusted brightness for presentation purposes; the maximum disparity in the actual disparity images is 28).

Rotating the disparity image will result in a larger image, but the memory extra-load is not significant. If the extra memory is not available, the disparity image could be rotated around one bottom corner of the image, depending on the roll. If the rotated image must be the same size as the original one, extra care has to be taken not to crop away the regions containing the actual ground plane.

In our approach, we only compute the v-disparity map for the lower half of the image. Besides speeding up the whole computational process, this also accounts for environment structuring particularities. The situations in which ground plane is present in the image but it is not located in its lower half are associated with the presence of obstacles in front of the user at very low distance. Therefore, we account for such situations as otherwise the system could lead the user to a non-accessible area.

The next step of the algorithm is to apply an adapted Hough transform on the v-disparity image in order to detect the ground plane line. The Hough transform requires as input a binary image and it detects the lines in that image. An extension of this algorithm that finds line segments is the progressive probabilistic Hough transform [15]. As the input of the Hough transform is a binary image, every pixel in the v-disparity image that is not zero is considered to be one. This can lead to heavy computational loads due to the high number of pixels in the image. Therefore it is a common practice to apply a threshold segmentation step on the v-disparity image. This intermediate step is performed to remove the pixels that have a very dark color in the v-disparity map. These pixels correspond to low sized objects or regions that cannot be associated with the ground plane. Upon completion, the Hough transform is applied on the new v-disparity map.

The output of the algorithm consists in a set of line segments. As the number of line segments can be quite high (some thousands in some cases), we implemented a selection mechanism to best fit the ground representation in the v-disparity map with the camera orientation and the ground plane of the previous frame. The algorithm takes into consideration various line constraints, tracks the detected ground plane in consecutive frames and consists in two steps described in the followings:

Step 1: determines what line segments correspond from the slope point of view. They will be further considered in the algorithm.

Along with the camera roll, the pitch angle of the camera affects the v-disparity representation of the ground plane [12] (see Figure 3). However, its influence is only related to the position of the line in the v-disparity map, i.e., the abscissa intercept.

The slope g of the line that corresponds to the ground plane in the v-disparity map can be computed using the following equation [12]

$$g = h / (b \cdot \cos\theta), \tag{1}$$

where h is the camera height above the ground plane level, b is the stereo baseline and θ is the tilt angle of the camera (or the pitch).

Step 2: selects the most suited line segment considering the detected ground plane in the last processed frame. The second step is required as multiple line segments that fulfill the condition of step one can be obtained. Also, it ensures the tracking of the ground plane in consecutive frames.

The selection algorithm is described by the following pseudo-code:

```
Step 1:
Compute slope $g$ from the camera orientation and its parameters (Eq. 1)
for each line segment ($LS$) in the line segments set returned by the Hough transform
   Compute $d = |slope(LS) - g|$
   if $d < \varepsilon$
      Consider $LS$ in Step 2
```

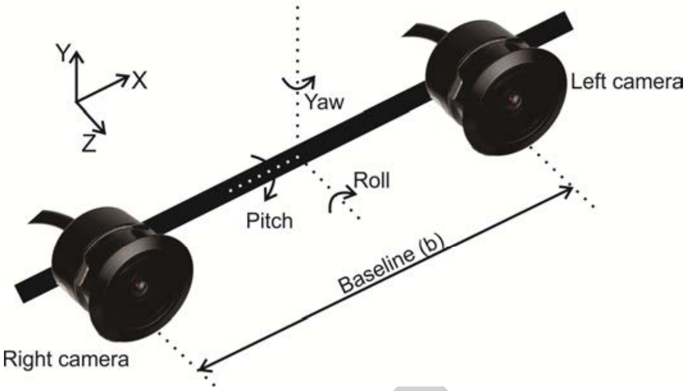

Figure 3. Stereo camera rig. Roll, pitch and yaw according to the stereo camera coordinate system.

```
Step 2:
if no ground plane was detected in the previous frame (suitable $LS$ was found)
   Select the $LS$ with the lowest $d$
else
   Select the $LS$ with the closest intersection point with the Ox axis to the intersection point with the Ox axis of the $LS$ of the previous frame
```

The value of ε was empirically determined to be 3.5. Too low value can lead to no ground plane line segment to be selected, even if it can be observed in the v-disparity image. Too high value of it and its purpose fades away as too many lines are considered, even vertical ones in some cases. A value between 3 and 4 ensures that a fair number of line segments can be considered as good candidates for the ground plane detection.

The line segment selection algorithm takes into consideration the case when no suitable line segment was detected in the previously processed frame. A special case that includes this situation is when processing the first frame of the acquisition process. This frame cannot be compared to any previous frame from the point of view of the line in the v-disparity map. This case also applies when in a frame the ground plane is not detected properly or is not present at all.

After selecting the best suited line segment, the next steps are performed to determine the ground plane equation that corresponds to the selected line segment. In order to compute the ground plane equation, first we determine the points in the disparity image that correspond to the detected line segment in the v-disparity map.

The next step is to determine the 3D positions of the pixels selected from the disparity image. This is done using the camera parameters, i.e. the focal length and the baseline. This step outputs a set of 3D points that correspond to the line selected in the v-disparity map.

Having a set of 3D points, we compute next the equation of the plane that approximates the points best. We do this by using a single value decomposition algorithm on an over defined system. After performing all these steps, a plane equation that corresponds to the ground plane is computed.

IV. Evaluation

As stated in the previous sections, the algorithm we propose in this paper was tailored for assistive systems for the visually impaired people. This proved that identifying the ground plane is more complex than finding it in automotive and robotics applications. Starting with a pre-computed disparity map, the algorithm was built on new constraints generated by the target user of our application.

To determine the accuracy of the proposed algorithm as well as for ethical reasons, synthetic test data was used. These data were generated in a virtual environment framework. The framework provides the disparity map. The map is accurately generated and it is not affected by any type of errors. When using real world data, the disparity map is computed based on left and right images acquired using a stereo-camera. Among the methods that can be used to compute a disparity map there are two major categories: methods based on correlation (block-matching, semi-global block matching [16]) or methods based on statistical optimization (efficient large-scale stereo [17]). Another method of acquiring a depth map is to use a depth camera.

The development of the 3D scenes in the virtual environment datasets was done using the Unity game engine. The scenes were designed to mimic common outdoor locations. Ground truth segmentation information was obtained by assigning a label and unique ID to each object in the scene. Thus, for the virtual environment testing scenarios the ground plane as well as camera orientation are straightforward to compute.

To allow the visual inspection of the results, a software application was developed. The points in the original left image for which the distance to the determined ground plane is lower than a threshold were determined. A 5 cm threshold for the presented results was chosen. These points were overlaid in a different color over the original image. Thus, a visual inspection can be performed to analyze the test cases that correspond to misdetections.

The images were acquired with a virtual stereo camera with a 15cm baseline, positioned at a height of 165 cm above the ground. The camera is manipulated by simulating the movement of an avatar.

In the scenario used in this paper, the roll of the camera is directly provided by the virtual environment framework. When using real-world data, a common solution is to use and inertial measurement unit (IMU) attached to the camera. The IMU device provides the camera orientation.

We implemented the proposed algorithm in C++ using the OpenCV image processing suite [18]. We used the OpenCV implementations of the probabilistic Hough transform and of the singular value decomposition algorithm.

Regarding the selection of the line segment on the output results of the Hough transform, several mechanisms were investigated. As presented in the related work section, some approaches take into consideration the line segment length. Others consider the number of points in the disparity image that correspond to the line segment [19]. Both these approaches rely on the assumptions that the ground plane occupies a large region of the image or is spread across the entire height of the image. These approaches fail when these assumptions are not met and provide un-satisfactory results. Taking into account these disadvantages, our algorithm considers a two step procedure that allows a more accurate line segment selection. This leads to an increased robustness of the ground plane estimation.

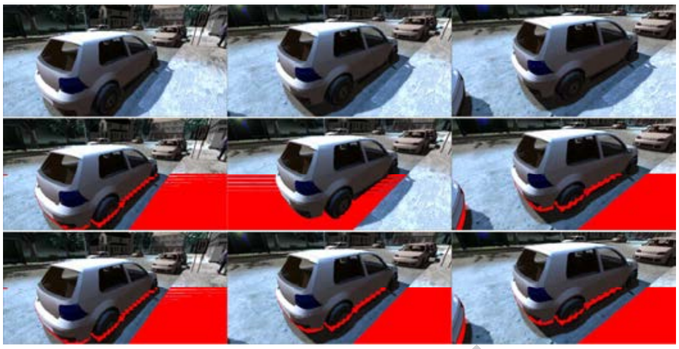

Figure 4. Top row: three consecutive frames with the sidewalk and the road visible. Middle row: detected ground planes using just the slope equation. Bottom: ground plane algorithm results using our presented algorithm.

A frequently encountered situation when the ground plane detection algorithm tends to fail is when an image contains two regions that both can qualify as ground plane. This is common in the images that capture a sidewalk near a road. Both surfaces fulfill the requirement from the point of view of the slope (see Eq. 1). Some algorithms might switch from the sidewalk to the road and back during consecutive frames. This case might lead to potentially dangerous situations when if the detected ground plane is rendered via different senses to visually impaired users. We present such a case in Figure 4. The top row depicts three consecutive frames captured by the left camera. The middle row shows the results of an algorithm that uses just the slope of the line segment to select the ground plane in the v-disparity image. As the line segments of the road and sidewalk in v-disparity image fulfill the slope requirements, the detected ground plane switched from one ground plane to another. The bottom row presents the results of our algorithm. Even if the ground plane corresponding to the road is larger, the algorithm selects the sidewalk in all three frames.

By taking into consideration the camera roll, the proposed algorithm accurately detects the ground plane in images acquired with a camera roll. Figure 5 presents such a case when the stereo camera is rotated along the z-axis. Common detection algorithms fail in such a case as the ground plane is not represented by a line in the v-disparity map but by a cluster of points. The detected surface in Figure 5 is not continuous due to the rounding errors caused by the use integer disparity values. The performance of the proposed ground plane detection algorithm was evaluated using a dataset consisting in 240 frames recorded in the virtual

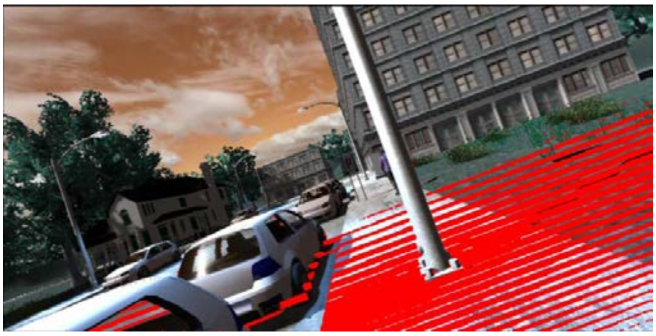

Figure 5. Detected ground plane in a z-axis rotated image.

environment. The ground plane was accurately detected in 225 frames. The ground plane was miss-detected in the situations when several line candidates were detected in the v-disparity map. In the second step of the algorithm these candidates had the same distance to the previous x-intercept. In this case, the algorithm randomly selects one.

V. Conclusions

In this paper we have introduced a ground plane detection algorithm based on v-disparity maps. Our approach is intended to be used in developing assistive technologies designed for the visually impaired people. In such applications, the stereo camera is worn by the user. This leads to the orientation of the camera to have more degrees of freedom than the ones considered in developing automotive applications. For this reason, the algorithm introduced in this paper takes into consideration the roll and pitch of the camera. The pitch is used to compute the slope of the ground plane line segment in the v-disparity map. To accurately detect the ground line using the v-disparity map the disparity image was rotated using the roll information of the camera. The proposed algorithm tracks the detected ground plane between consecutive frames based on the similarity of its representation in the corresponding v-disparity maps. The tracking also helps in distinguishing between the cases when an image presents multiple surfaces that can be considered as ground plane (sidewalks, roads).

The proposed algorithm was tested using synthetic data, providing promising results. Based on these results, we plan to adapt the algorithm to real world scenarios, using images acquired by cameras worn by visually impaired people. The time performance of the algorithm suggests that interactive frame rates can be achieved even when the processing is performed on portable devices.

One limitation of the presented algorithm that we have to mention is encountered when the user goes uphill or downhill and the camera is aligned to the horizontal axis. This situation can occur if the images are acquired with a camera worn by a visually impaired person. We plan to address this limitation in future work.

Acknowledgment

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 643636 “Sound of Vision”.

References

[1] F. Ribeiro, D. Florencio, P. Chou, and Z. Zhang, Auditory augmented reality: Object sonification for the visually impaired, in 14th International Workshop on Multimedia Signal Processing (MMSP), pp.~319—324, IEEE, 2012

[2] R.Hulik, V. Beran, M. Spanel, P. Krsek, P.Smrz, Fast and Accurate Plane Segmentation in Depth Maps for Indoor Scenes, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, 2012, pp. 1665-1670.

[3] H.J. Hemmat, A. Pourtaherian, E. Bondarev, P. H. N. de With, Fast planar segmentation of depth images. Proc. SPIE 9399, Image Processing: Algorithms and Systems XIII, 93990I (March 16, 2015)

[4] D. Kırcalı, F.B. Tek, Ground Plane Detection Using an RGB-D Sensor, Information Sciences and Systems 2014: Proc. of the 29th Intl. Symp. on Computer and Information Sciences, pp. 69-77, Springer, 2014

[5] M. Bujacz, Representing 3D scenes through spatial audio in an electronic travel aid for the blind, PhD Thesis, Technical University of Lodz, 2010.

[6] A. Rodriguez, J.J. Yebes, P.F. Alcantarilla, Bergasa, M. Luis, J. Almazan, A. Cela, Assisting the Visually Impaired: Obstacle Detection and Warning System by Acoustic Feedback, Sensors 12(12), pp. 17476-17496, 2012.

[7] S. Mattoccia, P. Macri, 3D glasses as mobility aid for visually impaired people, Proc. of the ECCV2014 Workshop, 2014.

[8] J.M. Saez Martinez, F. Escolano Ruiz, Stereo-based Aerial Obstacle Detection for the Visually Impaired, Workshop on Computer Vision Applications for the Visually Impaired, 2008.

[9] N. Soquet, D. Aubert and N. Hautiere, "Road Segmentation Supervised by an Extended V-Disparity Algorithm for Autonomous Navigation," 2007 IEEE Intelligent Vehicles Symposium, Istanbul, 2007, pp. 160-165

[10] J. Zhao, M. Whitty and J. Katupitiya, "Detection of non-flat ground surfaces using V-Disparity images," 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, 2009, pp. 4584-4589

[11] D. Yiruo, W. Wenjia and K. Yukihiro, "Complex ground plane detection based on V-disparity map in off-road environment," Intelligent Vehicles Symposium (IV), 2013 IEEE, Gold Coast, QLD, 2013, pp. 1137-1142.

[12] J. Zhao, J. Katupitiya and J. Ward, "Global Correlation Based Ground Plane Estimation Using V-Disparity Image," Proceedings 2007 IEEE International Conference on Robotics and Automation, Roma, 2007, pp. 529-534

[13] N. Fakhfakh, D. Gruyer, D. Aubert. Weighted V-disparity Approach for Obstacles Localization in Highway Environments. IEEE Intelligent Vehicles Symposium, Jun 2013, Australia. 8p, 2013

[14] R. Labayrade, D. Aubert, and J. P. Tarel, “Real time obstacle detection in stereovision on non flat road geometry through ”v-disparity” representation,” in Intelligent Vehicle Symposium, 2002. IEEE, vol. 2, June 2002, pp. 646–651 vol.2.

[15] J. Matas, C. Galambos, J.V. Kittler, Robust Detection of Lines Using the Progressive Probabilistic Hough Transform. CVIU 78 1, pp 119-137 (2000)

[16] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 328–341, Feb 2008.

[17] A. Geiger, R. Martin and U. Raquel. "Efficient large-scale stereo matching." In Asian conference on computer vision, pp. 25-38. Springer Berlin Heidelberg, 2010.

[18] Open Source Computer Vision Library, Itseez, 2015.

[19] A. Broggi, C. Caraffi, R. I. Fedriga and P. Grisleri, "Obstacle Detection with Stereo Vision for Off-Road Vehicle Navigation," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, San Diego, CA, USA, 2005, pp. 65-65. doi: 10.1109/CVPR.2005.503