# Obstacle Avoidance System for Assisting Visually Impaired People

Alberto Rodríguez, Luis M. Bergasa, Pablo F. Alcantarilla, Andrés Cela*

**Abstract—The aim of the project focuses on the design of an obstacle avoidance system for assisting visually impaired people. A disparity map will be generated thanks to the use of a stereo camera carried by the user. Working on this map will allow to develop an algorithm for the obstacle detection in any kind of scenario. The design will be completed with the introduction of audio signal to assist the blind person to avoid obstacles. To do that, we use the frequency of the signal to codify the obstacle's angle and its intensity to codify proximity. Some experimental results are presented as well as the conclusions and the future works.**

## I. INTRODUCTION

Autonomous navigation is of extreme importance for those who suffer from visual impairment problems. Without a good autonomy, visually impaired people depend on other factors or other people to perform typical daily activities. Within this context, a system that can provide robust and accurate localization of a visually impaired user in urban environments, like city or indoor ones, is much more than desirable.

Nowadays, most of the commercial solutions for visually impaired localization and navigation assistance are based on the Global Positioning System (GPS). However, these solutions are not suitable enough for the visually impaired community mainly for two reasons: the low accuracy in urban environments (errors about the order of several meters) and signal loss due to multi-path effect or line-of-sight restrictions. Moreover, GPS does not work if an insufficient number of satellites are directly visible. Therefore, GPS cannot be used in indoor environments.

Computer vision-based approaches offer substantial advantages with respect to GPS-based systems and constitute a promising alternative to address the problem. By means of visual SLAM techniques [1], [2], it is possible to build an incremental map of the environment, providing at the same time the location and spatial orientation of the user within the environment. In addition, compared to other sensory modalities computer vision can also provide a very rich and valuable perception information of the environment such as for example obstacle detection [3] or 3D scene understanding [4].

In this paper we will show a solution for visually impaired people. This solution will let them get the autonomy they need to move in any kind of scenario without any problem thanks to the obstacle avoidance module. This work is a module of a larger system for assisting visually impaired people based on visual maps. In fact, the objective of this project is to amplify and complete a thesis work called 'Vision Based Localization From Humanoid Robots to Visually Impaired People'. The rest of the paper is organized as follows: in section II there is an introduction to the thesis work 'Vision Based Localization From Humanoid Robots to Visually Impaired People'. In section III, we briefly review the stereo rig calibration and rectification processes in order to obtain accurate localization and mapping results. In section IV, the algorithm for the obstacle detection and warning by an audio signal is described. Section V explains experimental results considering challenging environments with many independent moving objects. Finally, in Section VI, main conclusions are shown.

## II. VISION BASED LOCALIZATION FROM HUMANOID ROBOTS TO VISUALLY IMPAIRED PEOPLE.

In this thesis, several algorithms are proposed in order to obtain an accurate real-time vision-based localization from a prior 3D map. For that purpose, it is necessary to compute a 3D map of the environment beforehand. For computing that 3D map, well-known techniques are employed such as Simultaneous Localization and Mapping (SLAM) or Structure from Motion (SfM). In this thesis, a visual SLAM system is implemented using a stereo camera as the only sensor that allows to obtain accurate 3D reconstructions of the environment. The proposed SLAM system is also capable to detect moving objects especially in a close range to the camera up to approximately 5 meters, thanks to a moving objects detection module. This is possible, thanks to a dense scene flow representation of the en- vironment, that allows to obtain the 3D motion of the world points. This moving objects detection module seems to be very effective in highly crowded and dynamic environments, where there are a huge number of dynamic objects such as pedestrians. By means of the moving objects detection module it is possible to avoid adding erroneous 3D points into the SLAM process, yielding much better and consistent 3D reconstruction results. Up to the best of our knowledge, this is the first time that dense scene flow and derived detection of moving objects has been applied in the context of visual SLAM for challenging crowded and dynamic environments.

In SLAM and vision-based localization approaches, 3D map points are usually described by means of appearance descriptors. By means of these appearance descriptors, the data association between 3D map elements and perceived 2D image features can be done. In this thesis, the author has

*All the authors are with the Department of Electronics, University of Alcalá, Alcalá de Henares, Madrid, Spain. E-mail: rodriguezfdez.alberto@gmail.com, bergasa@depeca.uah.es, pablo.alcantarilla@depeca.uah.es, andres_cela@yahoo.es

investigated a novel family of appearance descriptors known as Gauge-Speeded Up Robust Features (G-SURF). Those descriptors are based on the use of gauge coordinates. By means of these coordinates every pixel in the image is fixed separately in its own local coordinate frame defined by the local structure itself and consisting of the gradient vector and its perpendicular direction. The author has carried out an extensive experimental evaluation on different applications such as image matching, visual object categorization and 3D SfM applications that show the usefulness and improved results of G-SURF descriptors against other state-of-the-art descriptors such as the Scale Invariant Feature Transform (SIFT) or SURF.

In vision-based localization applications, one of the most expensive computational steps is the data association between a large map of 3D points and perceived 2D features in the image. Traditional approaches often rely on purely appearence information for solving the data association step. These algorithms can have a high computational demand and for environments with highly repetitive textures, such as cities, this data association can lead to erroneous results due to the ambiguities introduced by visually similar features. In this thesis, the autor develops an algorithm for predicting the visibility of 3D points by means of a memory based learning approach from a prior 3D reconstruction. Thanks to this learning approach, it is possible to speed-up the data association step by means of the prediction of visible 3D points given a prior camera pose.

The autor has implemented and evaluated visual SLAM and vision-based localization algorithms for two different applications of great interest: humanoid robots and visually impaired people. Regarding humanoid robots, a monocular vision-based localization algorithm with visibility prediction has been evaluated under different scenarios and different types of sequences such as square trajectories, circular, with moving objects, changes in lighting, etc. A comparison of the localization and mapping error has been done with respect to a precise motion capture system, yielding errors about the order of few cm. With respect to the vision-based localization approach for the visually impaired, the author has evaluated the vision-based localization system in indoor and cluttered office-like environments. In addition, it has been evaluated the visual SLAM algorithm with moving objects detection considering test with real visually impaired users in very dynamic environments such as inside the Atocha railway station (Madrid, Spain) and in the city center of Alcalá de Henares (Madrid, Spain). The obtained results highlight the potential benefits of the approach for the localization of the visually impaired in large and cluttered environments.

## III. STEREO RIG CALIBRATION AND RECTIFICATION.

Stereopsis is the impression of depth that is perceived when a scene is viewed with both eyes by someone with normal binocular vision. Binocular viewing of a scene creates two slightly different images of the scene in the two eyes due to the eyes' different positions on the head. These differences, referred to as binocular disparity, provide information that the brain can use to calculate depth in the visual scene, providing a major means of depth perception.

Computer stereo vision is a part of the field of computer vision. It is sometimes used in mobile robotics to detect obstacles. Two cameras take pictures of the same scene, but a distance – exactly like our eyes, separates them. A computer compares the images while shifting the two images together over top of each other to find the parts that match. The shifted amount is called the disparity. The disparity at which objects in the image best match is used by the computer to calculate their distance.

In order to obtain accurate localization and mapping results, a prior stereo rig calibration process is necessary. The stereo rig calibration problem involves the estimation of the intrinsic parameters and distortion parameters of each of the cameras, and the extrinsic parameters (rotation, translation) between cameras. In this Project, both cameras were calibrated independently using the Camera Calibration Toolbox for Matlab [5]. In this way, we can obtain the intrinsics calibration matrix for each of the cameras:

$$K = \begin{pmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$

where fx and fy are the focal lengths and (u0,v0) are the coordinates of the principal point of the camera. Radial (k1,k2) and tangential (p1,p2) distortion parameters are modeled by means of polynomial approximations [6]. After the calibration of each of the cameras, the extrinsics parameters of the stereo rig are estimated. The extrinsics parameters comprise of a rotation matrix RLR and a translation vector TLR between the left and right cameras of the stereo rig.

Once we have obtained the intrinsics and extrinsics of the stereo rig, we can correct the distortion of the images and perform stereo rectification [7]. Stereo rectification simplifies considerably the stereo correspondences problem and allows computing dense disparity or depth maps.

After stereo rectification, we obtain a new camera matrix K, where the left and right camera have the same focal lengths f and principal point (u0,v0). The rotation matrix between cameras RLR is the identity matrix, and the translation vector TLR encodes the baseline B of the rectified stereo rig. Now, considering an ideal stereo system, the depth of one 3D point can be determined by means of the following equation:

$$Z = f \cdot \frac{B}{u_R - u_L} = f \cdot \frac{B}{d_u}$$

where du is the horizontal disparity or the difference in pixels between the horizontal image projections of the same point in the right and left images. Given the depth of the 3D point Z, and the stereo image projections of the same point in both images (uL, uR, v) (notice that in a rectified stereo vL = vR = v) the rest of the coordinates of the 3D point with respect to the camera can be determined as:

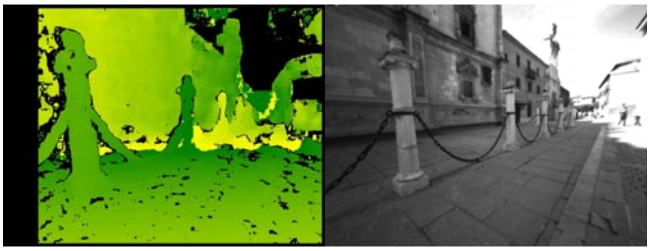

Fig. 1. The image depicts an example of the difficult challenging scenes that we can have in real world crowded environments for the visually impaired. The image shows the resulting disparity map.

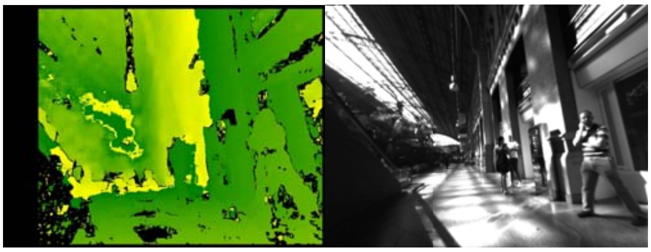

Fig. 2. The images show another example, this time inside a railway station (Atocha, Madrid).

## IV. OBSTACLE DETECTION ALGORITHM.

One of the advantages of stereo vision against monocular one, is that we can exploit the information from two images at once, obtaining dense disparity maps (between the left and right stereo views at each frame). Since for every pixel that has a valid disparity value we know its 3D position (with respect to the camera coordinate frame) we can detect obstacles in any scenario.

The development of an algorithm that detects obstacles is the main task described in this section. This task is possible thanks to the dense disparity map. The most important element of the algorithm is the creation of a cumulative grid. The grid is used represent the presence of obstacles in any image. It is important to say that the dimensions of the grid cover a distance of four and half meters. This is very important because in every image can appear obstacles that are quite far from the camera. As these obstacles do not belong to the grid of interest, they cannot be considered as obstacles. In other words, the obstacle is not going to be considered if it is not closer than five meters to the visually impaired person.

Each frame is analyzed, pixel-by-pixel, and, according to its coordinates, it is determined if it belongs to the ground plane or if it is part of a possible obstacle. The first step is to determine the ground plane. To achieve this, we use the RANSAC algorithm. It basically works as follows:

- A subset of N points is chosen from the disparity image.
- Three points of the subset (randomly selected) are used to estimate the values for A, B, C and D (plane components) in:
- The remaining N-3 points are tested against this model to determine the number of inliers:
- This process is repeated several times and the plane with the higher number of inliers is considered as the ground plane.
- In every frame, the plane components (A, B, C and D) are compared to the components of the previous frame to control that the plane chosen is the correct one.

Once it is determined which pixels belong to the ground, the rest of them are analyzed in order to set if they belong to the grid of interest and if they are part of a possible obstacle. If a pixel is part of an obstacle, it is included in the cumulative grid of interest. As we said before, the grid is used to represent the presence of and obstacle, not its shape.

The grid we are using is quite similar to the one shown in the next figure. Our grid is divided in three zones to represent the distance to obstacle (4.5, 3 and 1.5m). Each zone is divided in portions of 30º (as it is shown in the figure).

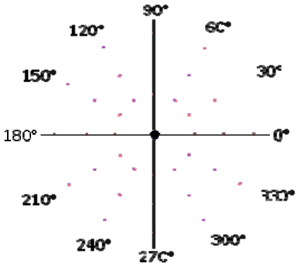

Fig. 3. In this image we can see the polar grid that has been implemented and the equations used to determine the obstacle's position.

As it has been said before, once we have determined which pixels belong to the ground plane, we analyze the rest of them. Thanks to the disparity map, we can obtain the coordinates of every pixel and we can check if they are inside the grid of interest. Before including any pixel in the cumulative grid, we compare the pixel height (y coordinate) with respect to the camera. In this way, we are not considering obstacles which are at a height of, for example, three meters, because they are not a problema for a visually impaired person.

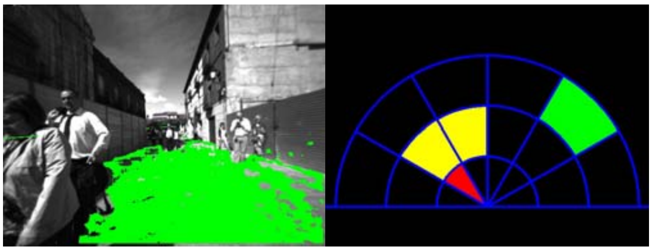

Fig. 4. The image depicts ground plane and obstacle detection in the streets of Alcalá de Henares.

In the images we can see the excellent performance of the algorithm. In the grid we represent the presence of the obstacle and the proximity to the user. Another example is shown in the next figure.

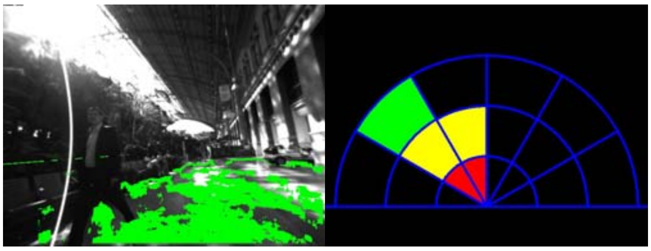

Fig. 5. The image depicts ground plane and obstacle detection inside the railway station of Atocha, Madrid.

In the final part of the algorithm, we warn to the user by audio signals. To implement this, it is used a library called libbeep. Thanks to this library we can configure audio signals to indicate the presence of an obstacle. This library includes a function called beep() that allows the user to control the pc-speaker, allowing different sounds to indicate different events. In this way we can control the frequency, length and repetitions of the acoustic signal.

The grid is analyzed before sending the acoustic signal. We do this to check which obstacle is closer to the user and inform him only about that obstacle. Depending on the obstacle position, the acoustic signal is sent to one earphone or another. If the obstacle appears in the left side, the audio signal will be sent to the left earphone and if it is in the right one, it will be sent to the right earphone. Obviously, if the obstacle is just in front of the user, the acoustic signal is sent to both earphones.

Next figure shows a flowchart that details this part of the algorithm. First step is to check in the grid if there is any obstacle and if it is the closest one to the user. If so, we determine the obstacle position and the corresponding acoustic signal is sent. If it is not the closest obstacle, the grid is analyzed again.

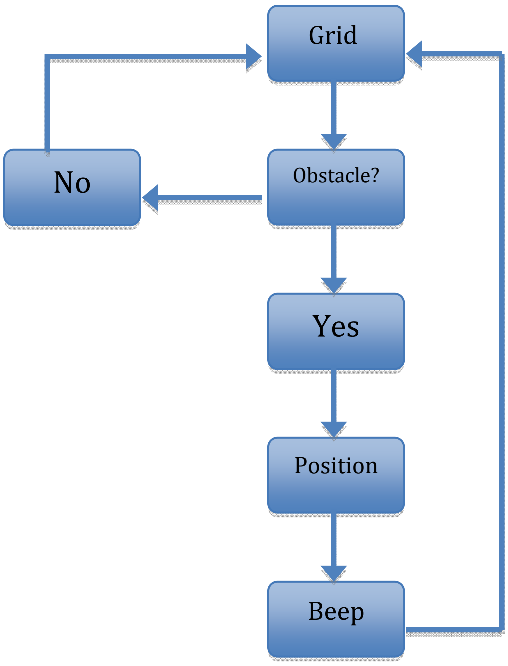

Fig. 6. The flowchart resumes the performance referred to the acoustic warning.

As we have said before, we can control the frequency, length and repetitions of the acoustic signal. Initially we have settled the following configuration but it is quite possible that it will be changed in the future, once we have tried the system with visually impaired people, to satisfy their needs and preferences.

| | 4.5 – 3m | 3 – 1.5m | 1.5 – 0m |
|---|---|---|---|
| **180º - 150º** | f=100, l=50, r=3 | f=1100, l=50, r=3 | f=3100, l=50, r=6 |
| **150º - 120º** | f=200, l=50, r=3 | f=1200, l=50, r=3 | f=3200, l=50, r=6 |
| **120º - 90º** | f=300, l=50, r=3 | f=1300, l=50, r=3 | f=3300, l=50, r=6 |
| **90º - 60º** | f=600, l=50, r=3 | f=1600, l=50, r=3 | f=3600, l=50, r=6 |
| **60º - 30º** | f=700, l=50, r=3 | f=1700, l=50, r=3 | f=3700, l=50, r=6 |
| **30º - 0º** | f=800, l=50, r=3 | f=1800, l=50, r=3 | f=3800, l=50, r=6 |

Fig. 7. The table shows the configuration used for the acoustic warning, where f is frequency (Hz), l is length (ms) and r are the repetitions.

## V. EXPERIMENTAL RESULTS.

Our vision-based system aid for the visually impaired consists of a stereo camera connected through a fireware cable to a small laptop for recording and processing the images. Fig. 8 depicts one image of our vision-based system aid for the visually impaired.

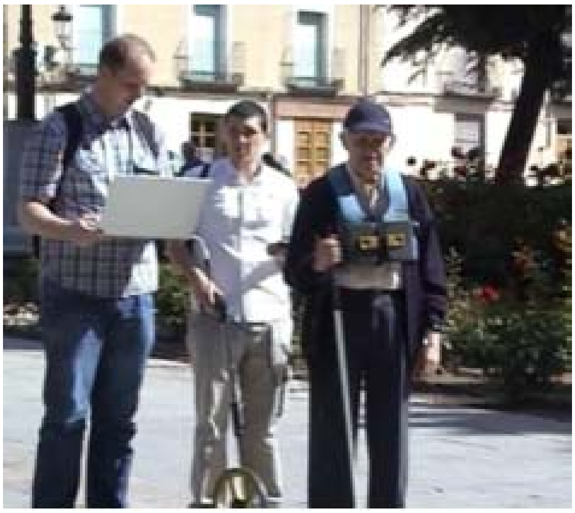

Fig. 8. The stereo camera system is attached to chest of the visually impaired user by means of a non-invasive orthopedic vest. Then the camera is connected to a small laptop by means of a fireware cable.

We conducted large-scale visual SLAM experiments with visually impaired users in highly dynamic environments, with many independently moving objects such as pedestrians or cars. We performed experiments inside the Atocha railway station (Madrid, Spain) and in a crowded area of the city center of Alcalá de Henares (Madrid, Spain). In these experiments, we were mainly interested in evaluating the performance of the obstacle detection algorithm. For this purpose, the visually impaired user received several indications before the start of the sequence about going from one starting point to a final destination.

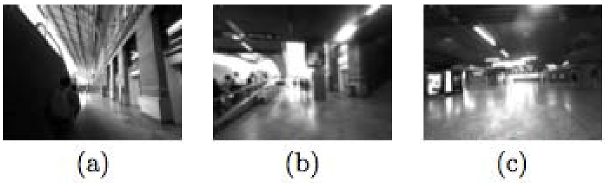

Fig. 9. (a) Start of the route (b) One image sample inside the railway station (c) End of the route: Entrance to the underground station.

For the mentioned experiments, we have used the Bumblebee2 stereo camera. This commercial stereo rig provides highly accurate camera calibration parameters and also stereo rectification and dense depth map generation on-chip. The camera baseline is 12 cm and the horizontal field of view is of 100◦ . The image resolution was 640 × 480 pixels and the acquisition frame rate was about 15 frames per second, considering B&W images.

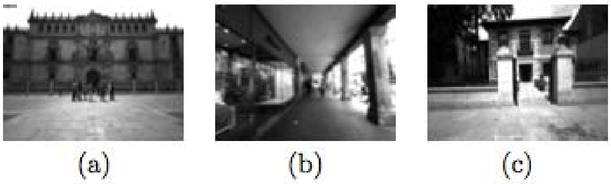

Fig. 10. (a) Start of the route: Façade of the University of Alcalá (b) Mayor street (c) End of the route: Cervantes house.

Apart from this, we have also made an analysis to check how many times the grid detects the presence of an obstacle. We have analyzed the first 1000 frames for the two experiments.

| | **4.5 – 3m** | **3 – 1.5m** | **1.5 – 0m** |
|---|---|---|---|
| **180º - 150º** | 0% | 0% | 0% |
| **150º - 120º** | 35,6% | 26,1% | 3,9% |
| **120º - 90º** | 39,7% | 31,2% | 6.8% |
| **90º - 60º** | 32,1% | 22,6% | 2,1% |
| **60º - 30º** | 19,1% | 11.9% | 5,8% |
| **30º - 0º** | 0% | 0% | 0% |

Fig. 11. The image shows the percentage of obstacles considered in the grid. Alcalá de Henares.

| | **4.5 – 3m** | **3 – 1.5m** | **1.5 – 0m** |
|---|---|---|---|
| **180º - 150º** | 0% | 0% | 0% |
| **150º - 120º** | 42,1% | 33,4% | 6,2% |
| **120º - 90º** | 46,6% | 38,3% | 9,1% |
| **90º - 60º** | 44,1% | 32,5% | 5,4% |
| **60º - 30º** | 25,3% | 18,6% | 2,2% |
| **30º - 0º** | 0% | 0% | 0% |

Fig. 12. The image shows the percentage of the obstacles considered in the grid. Atocha railway station (Madrid).

After obtaining these results we can see that the system can be very helpful for a visually impaired person. With the system they can increase their autonomy specially in indoor environments where there are a lot of people.

## VI. MAIN CONCLUSIONS AND FUTURE WORK.

In this paper, we have shown that it is possible to obtain accurate visual SLAM results in extremely challenging large-scale environments with many independently moving objects. This is possible, due to the detection of moving objects in the image by means of a dense scene flow representation and from derived residual motion likelihoods. When this object detection module is added to the visual SLAM pipeline, we can improve considerably visual odometry estimates and consequently, we obtain more accurate localization and mapping results in highly dynamic environments. We think that our results can be improved considerably in the next future from better dense scene flow representations [8], [9].

The visual SLAM module explained in this paper is an important part of a mobility system towards the autonomous navigation of the visually impaired. Once a persistent map of the environment is created by means of visual SLAM, this map can be used for localization [10] or topological navigation [11] purposes. Given a prior map of the environment and an estimate of the localization of the user within the environment, navigation commands can be computed and transmitted by audio devices to the visually impaired users. We are doing experiments with audio bone conducting, which is a non-invasive technology that allows visually impaired users to listen to other important sound sources in the environment (e.g. vehicles) while receiving navigation commands.

## VI. REFERENCES.

[1] J. M. Sáez, F. Escolano, and A. Peñalver, “First steps towards stereo-based 6DOF SLAM for the visually impaired,” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, 2005.

[2] V. Pradeep, G. Medioni, and J.Weiland, “Robot vision for the visually impaired,” in CVAVI10, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 2010.

[3] J. M. Sáez and F. Escolano, “Stereo-based aerial obstacle detection for the visually impaired,” in European Conference on Computer Vision (ECCV) / Workshop on Computer Vision Applications for the Visually Impaired (CVAVI), Marselle, France, 2008.

[4] A. Geiger, M. Lauer, and R. Urtasun, “A generative model for 3D urban scene understanding from movable platforms,” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, June 2011.

[5] Bouguet, J. (2008b). Documentation: Camera Calibration Toolbox for Matlab.

[6] Heikkila, J. and Silven, O. (1997). A four-step camera calibration procedure with implicit image correction. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pages 1106–1112.

[7] Hartley, R. (1999). Theory and practice of projective rectification. International Journal of Computer Vision, 35:115–127.

[8] C. Rabe, T. Müller, A. Wedel, and U. Frank, “Dense, robust, and accurate motion field estimation from stereo image sequences in real-time,” in Eur. Conf. on Computer Vision (ECCV), 2010, pp. 582–595.

[9] T. Müller, J. Rannacher, C. Rabe, and U. Franke, “Feature and depth-supported modified total variation optical flow for 3D motion field estimation in real scenes,” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011.

[10] P. Alcantarilla, S. Oh, G. Mariottini, L. Bergasa, and F. Dellaert, “Learning visibility of landmarks for vision-based localization,” in IEEE Intl. Conf. on Robotics and Automation (ICRA), Anchorage, AK, USA, 2010, pp. 4881–4888.

[11] A. Ranganathan and F. Dellaert, “Online probabilistic topological mapping,” Intl. J. of Robotics Research, 2010.