Way to Go! Detecting Open Areas Ahead of a Walking Person

Boris Schauerte, Daniel Koester, Manuel Martinez and Rainer Stiefelhagen

Karlsruhe Institute of Technology
Institute for Anthropomatics and Robotics
Vincenz-Prießnitz-Str. 3
76131 Karlsruhe, Germany

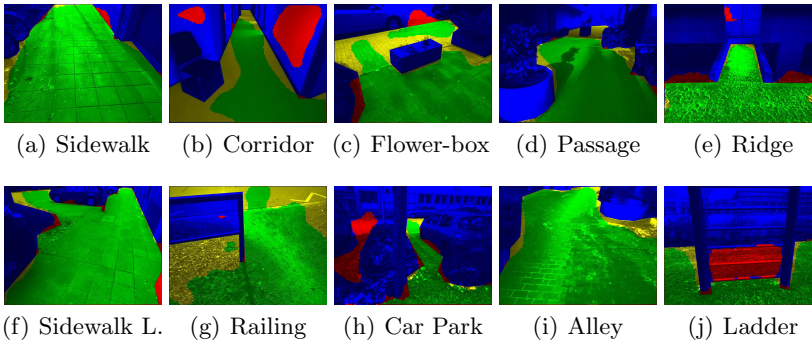
Abstract. We determine the region in front of a walking person that is not blocked by obstacles. This is an important task when trying to assist visually impaired people or navigate autonomous robots in urban environments. We use conditional random fields to learn how to interpret texture and depth information for their accessibility. We demonstrate the effectiveness of the proposed approach on a novel dataset, which consists of urban outdoor and indoor scenes that were recorded with a handheld stereo camera.

1 Introduction

Being able to savely navigate and explore areas in a city is an essential aspect of our everyday lifes. Accordingly, it is also an essential ability that autonomous, humanoid robots will have to master, if we want them to seemlessly operate in our part of the world, outside of controlled factory conditions. Similarly, safe navigation and exploration in urban areas is also an essential task when aiming toward increasing the autonomy, mobility, and overall life quality of visually impaired people. While location and directionality information provided by the global positioning system (GPS) can guide people and robots toward points of interest, GPS is blind with respect to the user's immediate surroundings. Thus, complementary systems are required to recognize hindrances and warn about potential dangers along the desired path.

Many systems have been developed and can be used to detect obstacles, including the classical white cane. Most technical solutions are often targeted towards different applications, imposing specialized constraints. Furthermore, they often rely on specialized, costly hardware such as sonar, radar, or light detection and ranging (LIDAR). This hardware is incapable of perceiving information provided by, e.g., traffic lights, signs, and lane markings. Furthermore, compared to touch sensors for robots or the classical white cane for blind people, an approach based on computer vision makes it possible to smoothly go around obstacles, because obstacles can visually be perceived from a greater distance.

In this paper, we use conditional random fields (CRFs) to determine the obstacle-free area in front of a walking person. Here, in contrast to many existing



(a) Sidewalk (b) Corridor (c) Flower-box (d) Passage (e) Ridge

(f) Sidewalk L. (g) Railing (h) Car Park (i) Alley (j) Ladder

Fig. 1. Exemplary binary classification results (CRF with depth and visual information) overlayed over the original image illustrating the true positives (green), true negatives (blue), false positives (red), and false negatives (yellow). This graphic is best seen in color.

approaches that model and try to detect obstacle classes, we address the dual problem and determine the parts in an image that are not blocked by obstacles. This way, we do not solely rely on the extremely varying characteristics of obstacles. Instead, assuming that the user is constantly guided by our system (see, e.g., [1]) and thus the direct area in front of him is free of obstacles[1], we can leverage the ground texture and depth information directly in front of him or her. To this end, we previously introduced a heuristic that uses depth maps to predict the ground surface normals and used this information as a rough predictor [2]. In this contribution, we investigate three CRF configurations that rely on different features to predict the obstacle-free areas: First, we train a CRF without depth information, which achieves a surprisingly good performance and is suitable for application in, e.g., modern smart phones. Second, we train a CRF that solely relies on specific depth information, which is independent of obstacle texture and also can use different sensors (e.g., Kinect's depth maps). Third, we train a CRF that leverages depth and visual information (see Fig. 1), which achieves the best results in our evaluation. We recorded and annotated a novel, challenging dataset to evaluate our approach. The dataset comprises of 20 videos that were recorded with a handheld stereo camera setup and cover different urban scenes under realistic ego-motion, lighting conditions, and scene complexity.

2 Related Work

The traditional white cane has a long history as a navigational device for visually impaired, especially blind, people. Many attempts have been made to create a digital enhancement, e.g., the *GuideCane* [3]. Martinez and Ruiz [4] complement

[1] Please consider that it is not our intention to replace the white cane, but instead we want to complement it. Thus, the user can recover from failures by relying on the classical walking stick.

the white cane and warn of aerial obstacles, such as low hanging branches. A more radical approach, that tries to replace the walking stick, uses sonar sensors and small vibrotactile units to signal feedback to the user [5]. To provide a navigational context inside buildings, Coughlan and Maguchi [6] use colored markers placed throughout a building. These are then detected and processed by a mobile phone application. The need for specific markers is removed by Chen et al. [7] through an Inertial Measurement Unit (IMU) and an a priori known map of the building. Obstacle detection is usually constrained to a small subset, e.g., matching upper body templates of pedestrians [8] or staircases [9]. These can be based on saliency [10], hough transformation [11] or optical flow [12]. As a dual problem, ground plane detection can be achieved through plane fitting [13] using RANSAC approaches. The authors model a relationship between the ground plane disparity and image pixel coordinates. Stereo camera rigs mounted on wheeled vehicles [14, 15] result in a steady camera movement and support a probabilistic model. The dependency of person detection location and size is used to generate a ground plane estimation.

Segmentation is another technique to detect the ground plane and was proposed by Lombardi [16]. In recent years, conditional random fields have achieved state-of-the-art performance for several segmentation tasks such as, e.g., semantic (scene) segmentation (e.g., [17–19]). Semantic segmentation describes the task of labeling each pixel of an image with a semantic category (e.g., "sky", "car", "street"). Accordingly, we chose conditional random fields as starting point to address our task. However, in contrast to the existing work, we are not interested in semantic object classes or types, but instead are interested in answering the question whether or not the region of an image accessible to a walking person? This naturally is related to road detection (e.g., [20, 21, 15]). However, detecting the walkable area in front of persons differs substantially from road detection for cars: First, we have to deal with a large amount of ego-motion that is characterized by the fact that cameras carried by a person are subject to considerably more degrees-of-freedom compared to cameras mounted on cars. Second, humans do not just follow roads, they sharply change direction, often even want to cross roads (and not just on zebra crossings), and they want to walk indoors as well as outdoors. Third, roads made for cars are much wider, straighter, and smoother than the small pathes between obstacles that are common in urban scenarios, see Fig. 3.

3 Open Area Detection

3.1 Depth-based Surface Angle Heuristic

The depth-based surface angle heuristic builds on work done by Koester et al. [2] that determines the accessible section in front of a walking person. Using epipolar geometry, we calculate the disparity of a point and therefore its distance from the camera. Doing so for every image point, we obtain a depth map $\Delta = \{(x_i, y_i, \delta_i)\}$, which allows us to calculate gradient ∇ for small image regions. After convolution of the image in both horizontal and vertical directions, we compute the local

gradient direction for each region. This results in map Φ, which consists of processed image regions and their corresponding gradient directions.

Within Φ, we calculate the accessible section by processing it in vertical bands. Such a band is a column of Φ, i.e., a vertical grouping of gradient regions. Starting with the band's bottommost block, we collect upwards all blocks that match our criteria of an aligned region. Correctly aligned regions are all blocks whose calculated angles deviate less than a certain threshold from a perfectly upright plane surface normal. Upright for this work is defined as being tilted upwards in the camera image, which prevents the algorithm from working on images where the camera is tilted above the used threshold. For our experiments this was not a problem, as the stereo camera system was mounted on a handheld carrier that was rarely tilted sideways more than 15 degrees, but to address this problem, one could simply use an Inertial Measurement Unit in combination with the cameras or estimate the dominant ground plane. When a block does not fit that criteria, the collection process stops and advances to the next vertical band. We repeat this process until the entire map has been processed, but rely on a geometric constraint in this process. When recording a real world scenario with a camera system from a persons point of view, the accessible section is usually connected to the bottom image border. This constraint allows us to focus on the accessible section that is directly in front of a person and not obstructed by any obstacles.

Due to the simplicity of the gradient calculation, the resulting algorithm works in realtime on a fairly recent computer.

3.2 Conditional Random Field

Structure, Learning, and Prediction In general, a CRF models the conditional probabilities of x (here, is it a walkable area?), given the observation y (i.e., features), i.e.

$$p(x|y) = \frac{1}{Z(y)} \prod_{c \in C} \psi(x_c, y) \prod_{i \in V} \psi(x_i, y) \quad , \tag{1}$$

where C is the set of cliques in the CRF's graph and i represent individual nodes. Here, ψ indicates that the value for a particular configuration x_c depends on the input y.

Naturally, our problem is a binary segmentation task, since the location depicted by a pixel can either be blocked by an obstacle or not, i.e. x_i can either be "blocked" or "non-blocked". We use a pairwise, 4-connected grid CRF structure. We linearly parametrize the CRF parameter vector Θ in unary node $u(y, i)$ (i.e., information at an image location) and edge features $v(y, i, j)$ (e.g., relating neighbored image locations). Here, it is important to consider that the cliques in a 4-connected, grid-structured graph are the sets of connected nodes, which are represented by the edges. Thus, we fit two matrices F and G such that

$$\Theta(x_i) = Fu(y, i) \tag{2}$$

$$\Theta(x_i, x_j) = Gv(y, i, j) \tag{3}$$

Here, y is the observed image and $\Theta(x_i)$ represents the parameter values for all values of x_i. Similarly, $\Theta(x_i, x_j)$ represents the parameter values for all x_i, x_j. Then, we can calculate

$$p(x;\Theta) = \exp\left[\sum_i \Theta(x_i) + \sum_j \Theta(x_i, x_j) - A(\Theta)\right] \quad , \tag{4}$$

where $A(\Theta)$ is the log-partition function that ensures normalization.

We use tree-reweighted belief propagation (TRW) to perform approximate marginal inference, see [22]. TRW addresses the problem that it is computationally intractable to compute the log-partition function $A(\Theta)$ exactly and approximates $A(\Theta)$ with

$$\hat{A}(\Theta) = \max_{\mu \in \mathcal{L}} \Theta \cdot \mu + \hat{H}(\mu) \quad , \tag{5}$$

where $\hat{H}$ is TRW's entropy approximation [22]. Here, $\mathcal{L}$ denotes the valid set of marginal vectors

$$\mathcal{L} = \{\mu : \sum_{x_{c \setminus i}} \mu(x_c) = \mu(x_i) \wedge \sum_{x_i} \mu(x_i) = 1\} \quad , \tag{6}$$

where μ describes a mean vector, which equals a gradient of the log-partition function. Then, the approximate marginals $\hat{\mu}$ are the maximizing vector

$$\hat{\mu} = \arg\max_{\mu \in \mathcal{L}} \Theta \cdot \mu + \hat{H}(\mu) \quad . \tag{7}$$

This can be approached iteratively until convergence or a maximum number of updates [23].

To train the CRF, we rely on the clique loss function, see [22],

$$L(\Theta, x) = -\sum_c \log \hat{\mu}(x_c; \Theta) \quad . \tag{8}$$

Here, $\hat{\mu}$ indicates that the loss is implicitly defined with respect to marginal predictions – again, in our implementation these are determined by tree-reweighted belief propagation – and not the true marginals. This loss can be interpreted as empirical risk minimization of the mean Kullback-Leibler divergence of the true clique marginals to the predicted ones.

Features As unary depth-based features, we use the surface angle map Φ as presented in Sec. 3.1 and additionally the disparity map. As unary image-based features, we include the following information at each CRF grid point: First, we include each pixel's normalized horizontal and vertical image position in the feature vector. Second, we directly use the pixel's intensity value after scaling the image to the CRF's grid size. We expand the position and intensity information using sinusoidal expansion as described by Konidaris et al. [24, 23].

Fig. 2. Exemplary key frames and binary masks of videos (a) Sidewalk, (b) Corridor, (c) Flower-box, (d) Passage, (e) Ridge, and (f) Narrow.

Third, we append the histograms of oriented gradients (HoG) to encode the texture information.

As CRF edge features, we use a simple 1-constant and 10 thresholds to encode the difference of neighboring pixels. Then, we multiply the existing features by an indicator function for each edge type (i.e., vertical and horizontal), effectively doubling the number of features and encoding conjunctions of features and edge type. This way, we parametrize vertical and horizontal edges separately.

4 Experimental Evaluation

4.1 Dataset

We recorded a dataset to evaluate the detection of all image regions that are not blocked by obstacles [2]. This was necessary, because existing related datasets have been recorded for other use cases and mostly focus either on road scenes or people detection inside pedestrian areas (see Sec. 2). Since we target wearable sensor platforms that can assist visually impaired persons, we recorded the dataset on a handheld mobile platform carried by a pedestrian. Consequently, our dataset contains – among other challenges – realistic (camera) ego-motion on all axes. We recorded 20 videos of varying length that show common urban scenes such as, e.g., walkways and sidewalks with static obstacles (e.g. parked cars, bicycles, and street poles) and moving obstacles (e.g., cyclists and pedestrians). Some example images illustrating the dataset are shown in Fig. 2.

The videos were recorded with a stereo setup consisting of two *Point Grey Grasshopper 2* cameras, which were mounted onto a small metal carrier, axes in parallel, at a fixed distance of about 6cm with respect to the lenses' centers and the used lenses provide a field of view of 82 by 67 degrees. The metal carrier was manually held at breast height and the cameras were pointed towards the ground in front of the carrying person. Furthermore, the cameras were configured to synchronize time as well as the adaptation of gain and exposure. All videos were recorded in 8-bit monochrome mode at a resolution of 1024×768 pixels at 15 frames per second.

Overall, the dataset contains 7789 frames, out of which we labeled every fifth frame (i.e., 3 fps). We did not label the first 30 frames of each video in order to

allow for proper gain and exposure synchronization. We labeled the obstacle-free section as a polygon, where we imposed the constraint that a valid region must connect to the bottom frame boundary to be reachable from the current position, otherwise obstacles could obstruct it. Examples of binary masks created from labeled frames can be seen in figure 2.

4.2 Measures

We use the human ground truth annotation of the walkable area in each labeled frame, see Sec. 4.1, to evaluate our approach with respect to two performance measures: First, we use the pixel-wise binary classification error (i.e., $1-$accuracy) to directly evaluate the goodness of the binary classification. In case of the depth-based heuristic, we calculate the best threshold over all training images. Second, we use the area under the receiver-operator characteristic curve (ROC AUC) to investigate the influence of the decision thresholds on the classification performance. This way, we have a measure of how well behaved the non-binary (probabilistic) prediction maps are, i.e., how far away from the truth are the predicted values typically?

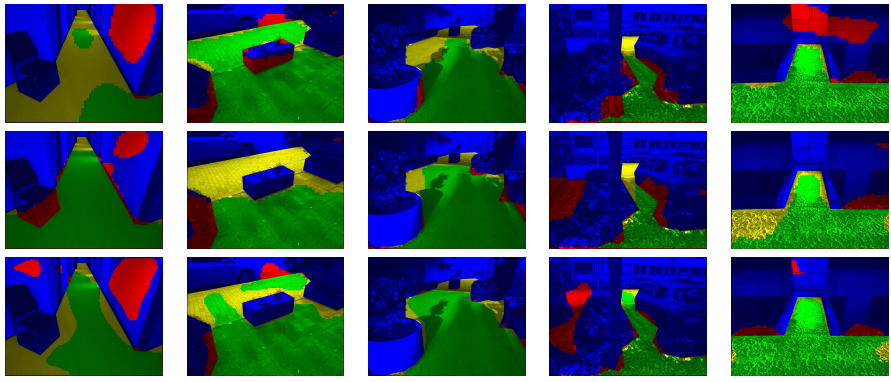
To validate the statistical significance of our results, we perform paired t-tests to ensure that the compared algorithms are in fact better or worse. Here, the results of the two algorithms in question are paired for each video. We reject hypotheses at significance level $\alpha = 0.05$.

4.3 Algorithm Parameters

We use Geiger et al.'s efficient large-scale stereo matching algorithm [25] to calculate the disparity and depth maps from the stereo image pairs. For the surface angle calculation (Sec. 3.1), we use a kernel size of 32×32 pixels, tiling the original 1024×768 pixels image into 32×24 blocks. To heuristically determine the section that is not blocked by obstacles, we consider values that deviate less than $22.5°$ from an optimal perpendicular angle. We train the CRFs using a grid/feature map size of 64×48 pixels.

4.4 Results

As is apparent in Tab. 4, all CRFs outperform the heuristic baseline method on 18 out of the 20 videos in terms of prediction error. Averaged over all videos, the depth-based baseline method achieves a pixel-wise error of 0.209 and a ROC AUC of 0.852. If we train the CRF using the depth and surface angle maps exclusively, then we achieve an error of 0.126 and a ROC AUC of 0.939. Using a pairwise t-test to compare the accuracies achieved on the videos, we can reject the hypotheses of equal mean ($p_{\mathrm{E}} = 0.022$) and that the heuristic approach might be better than the CRF ($p_{\mathrm{I}} = 0.011$). In contrast, if we exclusively rely on image features, then we achieve an error of 0.209 and a ROC AUC of 0.937, which is better than the heuristic approach (we can reject inferiority with $p = 0.030$, but we are unable



(a) Corridor (b) Flower-box (c) Passage (d) Car Park (e) Ridge

Fig. 3. Exemplary binary classifications for different CRF configurations. Top-to-bottom: No depth, depth only, and visual and depth. The binary classification is overlaid over the original image to illustrate the true positives (green), true negatives (blue), false positives (red), and false negatives (yellow). This graphic is best seen in color.

to reject equality) and very similar but slightly inferior in performance to the depth only approach (we are not able to reject any hypothesis). Finally, if we train a CRF with both depth and visual information, it achieves a pixelwise prediction error of 0.111 and a ROC AUC of 0.949. This approach provides the best performance on 16 out of the 20 videos in terms of error minimization and 14 out of 20 in terms of best ROC AUC performance. Confirmed by our statistical tests, we can safely assume that this algorithm in fact provides a better performance than the heuristic baseline (H_0: Inferior? Reject at $p_{\mathrm{I}} = 0.006$; H_0: Equal? Reject at $p_{\mathrm{E}} = 0.012$) and the CRF without depth information (trivially visible, because the results on all videos are better). But, contradicting our expectation given these numbers, we are unable to reject the possibility that the depth only approach is equally good or even better. Why is that? This is caused by two video sequences, namely "Corridor" and "Fence" for which the depth-only results stand out of the other results by being drastically better. If we exclude both video sequences, we can safely reject that the performance of the depth only CRF is equal or better ($p_{\mathrm{E}} = 0$ and $p_{\mathrm{I}} = 0$, respectively) than the performance of the CRF with depth and visual information.

The "Corridor" is an interesting case, because not just the depth-based CRF but even the depth-based heuristic outperform the CRF that uses depth and visual information, see Tab. 4. However, this is most likely explained by the absence of a second indoor video that could provide suitable visual training data in our leave-one-video-out evaluation. Furthermore, it is important to note that the walls and floor in the video are nearly textureless and consequently hardly suited for HoG-like features. The case is slightly different for the "Fence" sequence, for which it is interesting to have a look at the ROC AUC. The ROC AUC of the full – i.e., visual and depth features – CRF is substantially higher ($0.933 > 0.855$)

Fig. 4. Pixel-wise binary classification error and area under the receiver operator characteristic curve achieved in a leave-one-video-out cross-validation procedure. The algorithms are: depth-based surface angle heuristic (D-based), a CRF that only uses depth features (D only), a CRF that only uses visual features (no D), and a CRF that uses depth and visual features (full). The best result for each video is marked bold. Results where D-based outperforms full are underlined.

	↓ Pixel-wise Error (1 − Accuracy)				↑ ROC AUC			
	heuristic	CRF			heuristic	CRF		
Sequence	D-based	D only	no D	full	D-based	D only	no D	full
Alley	0.099	0.088	0.088	**0.066**	0.928	0.971	0.972	**0.979**
Alley L.	0.138	0.124	0.073	**0.054**	0.892	0.958	0.979	**0.985**
Bicycle	0.324	0.141	**0.092**	**0.092**	0.753	0.906	0.946	**0.953**
Car	0.149	0.124	0.090	**0.065**	0.850	0.944	0.935	**0.973**
Corridor	0.204	**0.086**	0.365	0.316	0.819	**0.972**	0.702	0.753
Fence	0.185	**0.126**	0.284	0.260	0.855	**0.936**	0.914	0.933
Flower-box	0.276	0.169	0.160	**0.158**	0.783	0.897	0.964	**0.966**
Hedge	0.186	0.202	0.154	**0.105**	0.836	0.866	0.893	**0.917**
Ladder	0.132	0.176	0.155	**0.112**	0.836	0.920	**0.937**	0.913
Narrow	0.071	0.104	0.106	**0.058**	0.958	0.983	0.982	**0.993**
Pan	0.350	0.127	0.084	**0.063**	0.759	0.940	**0.987**	0.981
Passage	0.195	**0.125**	0.187	0.130	0.850	0.941	0.942	**0.964**
Railing	0.304	**0.189**	0.234	0.203	0.760	**0.851**	0.835	0.835
Ramp	0.269	0.163	0.164	**0.129**	0.803	0.916	**0.970**	**0.970**
Ridge	0.801	0.163	0.258	**0.140**	0.854	0.885	0.910	**0.960**
Sidewalk	0.087	0.056	0.084	**0.046**	0.929	**0.969**	0.945	0.968
Sidewalk 2	0.096	0.073	0.057	**0.043**	0.947	0.978	0.978	**0.982**
Sidewalk L.	0.088	0.110	0.087	**0.070**	0.889	0.979	0.981	**0.986**
Sign	0.146	0.099	0.079	**0.064**	0.890	0.978	0.986	**0.987**
Street	0.083	0.075	0.059	**0.044**	0.940	0.986	0.988	**0.991**
Average	0.209	0.126	0.143	**0.111**	0.852	0.939	0.937	**0.949**

than the ROC AUC achieved by the depth-based heuristic and only marginally worse compared to the depth-only CRF (0.936). This stands in contrast to the considerably higher pixel-wise error ($0.260 > 0.185$ and $0.260 > 0.126$). Accordingly, it is most likely that the actual error is caused by the final decision made by the CRF and, considering from the performance of the CRF without depth-features, might arise from the present visual features.

Overall, it is easy to conclude that depth is an important and reliable information about the accessibility of a ground section ahead. This makes sense, because while the texture of obstacles may vary substantially their main property of physically blocking a certain area is well represented in depth maps. However, if we again closely examine the results, we can see that the visual-only CRF achieves an equal or lower error compared to the depth-based CRF on 12 of the 20 sequences. Thus, it is possible to achieve an accurate prediction even with a single, monocular camera, if we have sufficient and appropriate training data. However, in many cases depth information seems more reliable and especially does not only depend on texture that can vary substantially for obstacles and scenes in general.

Although the CRF that uses depth and visual features provides the overall best performance, all algorithms have their respective use cases: First, CRFs that do not rely on depth features can use monocular cameras, which nowadays can be found in nearly all off-the-shelf mobile phones. Second, the lightweight complexity of the depth-based heuristic[2] stands in contrast to the roughly 2 fps our CRF-based implementations that are not real-time capable yet. Third, the depth-based heuristic and depth-based CRF seem to perform well in the absence of scene specific, targeted training data. Thus, they could serve as fallback in scenes or situations for which the CRFs that include visual information have not been trained.

5 Conclusion

We presented how we determine obstacle-free areas in front of a walking person or (humanoid) robot. In contrast to prior art, we focus on detecting the obstacle-free areas instead of detecting potential obstacles directly, thus addressing the dual problem to classical obstacle detection. Our evaluation dataset consists of 20 videos depicting urban scenes that we recorded using a handheld stereo camera rig. It contains realistic amounts of lighting variations, ego-motion, and scene variety in urban scenarios. Given the dataset, we can train and investigate different conditional random fields for varying sensor configurations, i.e. depth information only, stereo video recordings, and monocular video recordings. To efficiently work with depth information, we use a heuristic that predicts flat ground surfaces in front of the user that typically represent sidewalks, streets, or floors in urban environments. This algorithm also serves as a depth-only, non-CRF baseline algorithm. In summary, we are able to achieve a pixel-wise prediction accuracy

[2] We exclude the time for the depth map calculation, which could be replaced by specialized sensors, e.g., Kinect.

of 0.874, 0.857, and 0.889 for depth-only, monocular images, and stereo images, respectively.

As part of our future work, we plan to investigate how haptic or auditory output modalities can be used to communicate the information to visually impaired users. For this purpose, we also plan to improve the computational efficiency to achieve a high system responsiveness that is essential for auditory or haptic user interfaces. Furthermore, we want to integrate self-localization and tracking to smoothly steer a blind person around obstacles.

Acknowledgments

The work presented in this paper was supported by a Google Research Award for "A Mobility and Navigational Aid for Visually Impaired Persons", the German Research Foundation (DFG) within the Collaborative Research Program SFB 588 "Humanoide Roboter", and the Quaero Programme, funded by OSEO, French State agency for innovation.

References

1. Martinez, M., Constantinescu, A., Schauerte, B., Koester, D., Stiefelhagen, R.: Cognitive evaluation of haptic and audio feedback in short range navigation tasks. In: Proc. 14th Int. Conf. Computers Helping People with Special Needs. (2014)
2. Koester, D., Schauerte, B., Stiefelhagen, R.: Accessible section detection for visual guidance. In: IEEE/NSF Workshop on Multimodal and Alternative Perception for Visually Impaired People. (2013)
3. Shoval, S., Ulrich, I., Borenstein, J.: Navbelt and the guide-cane [obstacle-avoidance systems for the blind and visually impaired]. IEEE Robotics Automation Magazine **10**(1) (2003) 9–20
4. Martinez, J.M.S., Ruiz, F.E., et al.: Stereo-based aerial obstacle detection for the visually impaired. In: Proc. Workshop on Computer Vision Applications for the Visually Impaired. (2008)
5. Cardin, S., Thalmann, D., Vexo, F.: Wearable obstacle detection system for visually impaired people. In: Proc. VR workshop on haptic and tactile perception of deformable objects. (2005)
6. Coughlan, J., Manduchi, R.: A mobile phone wayfinding system for visually impaired users. Assistive technology research series **25**(2009) (2009) 849
7. Chen, D., Feng, W., Zhao, Q., Hu, M., Wang, T.: An infrastructure-free indoor navigation system for blind people. Intelligent Robotics and Applications (2012) 552–561
8. Mitzel, D., Leibe, B.: Close-range human detection and tracking for head-mounted cameras. In: Proc. British Machine Vision Conference. (2012)
9. Hoon, Y., Leung, L.T.s., Medioni, G.: Real-time staircase detection from a wearable stereo system. Proc. Int Conf on Pattern Recognition (2012)
10. Lee, C.H., Su, Y.C., Chen, L.G.: An intelligent depth-based obstacle detection system for visually-impaired aid applications. In: Proc. Int. Workshop Image Analysis for Multimedia Interactive Services. (2012)

11. Labayrade, R., Aubert, D., Tarel, J.P.: Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In: Proc. Intelligent Vehicle Symposium. (2002)
12. Braillon, C., Pradalier, C., Crowley, J., Laugier, C.: Real-time moving obstacle detection using optical flow models. In: Proc. Intelligent Vehicles Symposium. (2006)
13. Se, S., Brady, M.: Ground plane estimation, error analysis and applications. Robotics and Autonomous Systems **39**(2) (2002) 59–71
14. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: Moving obstacle detection in highly dynamic scenes. In: Proc. Int. Conf. Robotics and Automation. (2009)
15. Ess, A., Schindler, K., Leibe, B., Van Gool, L.: Object detection and tracking for autonomous navigation in dynamic environments. The International Journal of Robotics Research **29**(14) (2010)
16. Lombardi, P., Zanin, M., Messelodi, S.: Unified stereovision for ground, road, and obstacle detection. In: Proc. Intelligent Vehicles Symposium. (2005)
17. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. Int. Conf. Machine Learning. (2001)
18. Passino, G., Patras, I., Izquierdo, E.: Latent semantics local distribution for crf-based image semantic segmentation. In: Proc. British Machine Vision Conference. (2009)
19. Verbeek, J., Triggs, B.: Scene segmentation with crfs learned from partially labeled images. In: Advances in Neural Information Processing Systems. Volume 20. (2008) 1553–1560
20. Kong, H., Audibert, J.Y., Ponce, J.: General road detection from a single image. IEEE Trans. Image Processing **19**(8) (2010) 2211–2220
21. Alvarez, J., Lopez, A.: Road detection based on illuminant invariance. IEEE Trans. Intelligent Transportation Systems **12**(1) (2011) 184–193
22. Wainwright, M.J., Jordan, M.I.: Graphical Models, Exponential Families, and Variational Inference. Now Publishers Inc., Hanover, MA, USA (2008)
23. Domke, J.: Learning graphical model parameters with approximate marginal inference. IEEE Trans. Pattern Analysis and Machine Intelligence **35**(10) (2013) 2454–2467
24. Konidaris, G., Osentoski, S., Thomas, P.S.: Value function approximation in reinforcement learning using the fourier basis. In: AAAI Conf. Artificial Intelligence. (2011)
25. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. Proc. Asian Conf. Computer Vision (2011)