Can We Unify Perception and Localization in Assisted Navigation? An Indoor Semantic Visual Positioning System for Visually Impaired People

Haoye Chen, Yingzhi Zhang, Kailun Yang, Manuel Martinez, Karin Müller and Rainer Stiefelhagen

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology
Correspondence: kailun.yang@kit.edu

Abstract. Navigation assistance has made significant progress in the last years with the emergence of different approaches, allowing them to perceive their surroundings and localize themselves accurately, which greatly improves the mobility of visually impaired people. However, most of the existing systems address each of the tasks individually, which increases the response time that is clearly not beneficial for a safety-critical application. In this paper, we aim to cover scene perception and visual localization needed by navigation assistance in a unified way. We present a semantic visual localization system to help visually impaired people to be aware of their locations and surroundings in indoor environments. Our method relies on 3D reconstruction and semantic segmentation of RGB-D images captured from a pair of wearable smart glasses. We can inform the user of an upcoming object via audio feedback so that the user can be prepared to avoid obstacles or interact with the object, which means that visually impaired people can be more active in an unfamiliar environment.

Keywords: Visual Localization · 3D Reconstruction · Semantic Segmentation · Navigation Assistance for the Visually Impaired.

1 Introduction

With the help of mobility aids such as a global navigation satellite system (GNSS) device, it is possible for visually impaired people to travel more independently. Although such mobility aids can navigate visually impaired people to the entrance of the right target building, the unknown indoor environment remains a labyrinth for them [3]. The situation indoors is a more demanding challenge than outdoors, as each room can have a different layout and indoor navigation systems are not on the market. For visually impaired people it is difficult to find their own way to the desired destination without the company of a personal guide. In addition, the arrangement of movables can change when returning to a familiar place, which can be dangerous for people with visual impairments if they rely on their memory for navigation.

On the other hand, vision-based navigation aids have made remarkable progress in recent years [5, 8, 20], making it possible to perceive the environment and localize oneself effectively, which significantly improves the mobility of visually impaired people. However, most of these tools work outdoors and address each task separately, which increases the response time that is clearly not advantageous for safety-critical assisted navigation. In these scenarios, a system, which can capture and convey both positional and cognitive messages, offers significant support to visually impaired people.

In this paper, we aim to cover scene perception and localization desired by navigation assistance in a unified manner. We present a semantic visual localization system for the visually impaired under indoor circumstances, in order to help them to acquire the overall information about their surroundings and relative position of objects nearby. The system reconstructs a 3D copy of the user's surroundings in real time from a stereo camera. Meanwhile, it associates semantic concepts of nearby objects with corresponding entities in the 3D reconstruction by using pixel-wise semantic segmentation. As the system maps the real world into a digital one, we can estimate the user's location according to the camera position in the 3D map. Finally, through audio feedback, semantic concepts combined with their position can give the user intuitive awareness and understanding of their surroundings (e.g., the system can tell the user what kind of obstacles are in front of him or tell him where a door is).

2 Related Work

In recent years, the robotics community has well explored Simultaneous Localization and Mapping (SLAM) [2] problems. Visual-SLAM research still has a substantial potential thanks to astonishing achievements by computer vision and computer graphic techniques. ORBSLAM [9] exhibits a system for monocular, stereo, and RGB-D cameras, including loop closing, relocalization, and map reuse. ElasticFusion [18] is capable of estimating a dense 3D map of an indoor environment. Kimera [14] enables mesh reconstruction and semantic labeling in 3D. However, there is still a huge gap between localization and assistance, as visually impaired people always rely on the surrounding semantic information to localize themselves, which is not necessarily mapped to the corresponding positioning results from visual SLAM algorithms.

Deep neural networks have achieved excellent results in semantic segmentation. SegNet [1] was presented as an encoder-decoder architecture for pixel-wise semantic segmentation. ENet [10], Fast-SCNN [11] and ERFNet [12, 13] were proposed as efficient architectures for fast inference. ACNet [4] introduced an attention complementary module to exploit cross-modal features, which is also used in RFNet [17] that facilitates real-time RGB-D segmentation. In this work, we use RFNet due to its real-time performance and fusion capability. Despite these progress, in previous wearable systems, semantic segmentation has only been used for unified scene perception [20], leaving rich opportunities open to assist localization.

In the field of assisted navigation with computer vision methods for visually impaired people, Lin et al. [5] proposed an outdoor localization system for visually impaired pedestrians. Hu et al. [3] presented an indoor positioning framework based on panoramic visual odometry, which attained robust localization performance due to the large field of view. Lin et al. [6] put forward a data-driven approach to predict safe and reliable navigable instructions by using RGB-D data and the established semantic map. Liu et al. [7] built a solution for indoor topological localization with semantic information based on object detection, which is the closest to our work. Our work differs from these works as we aim to use the dense semantic maps produced by an RGB-D segmentation network to improve localization, since the pixel-wise results, which are extremely informative during orientation and navigation, not only allow the user to recognize nearby objects, but also facilitate the detection of walkable areas.

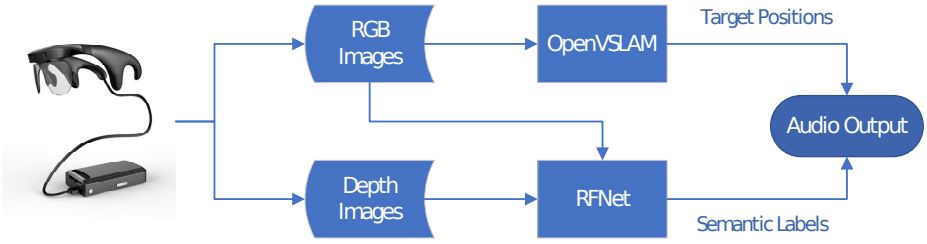

Fig. 1. An overview of the proposed system. The smart glasses provide RGB images to OpenVSLAM to establish localization and mapping. Meanwhile, the RFNet generates semantic labels from the RGB and depth images. We select the target positions with their semantic labels to produce audio prompts for the user.

3 System Description

In this section, we describe the hardware and software components as well as the interaction of the components of the entire system. Figure 1 gives an overview of the proposed system.

3.1 Hardware Components

The system consists of a RealSense camera R200, a pair of bone-conduction earphones, as well as an NVIDIA Jetson AGX Xavier processor. The camera and earphones are integrated into a pair of wearable smart glasses, as it is shown in Fig. 2. We perform the semantic segmentation and localization on the embedded processor Xavier in real time.

Fig. 2. Devices and the real-time results of our system. The left image shows the user wearing the devices. The blue box in left image indicates the smart glasses while the orange box indicates the Xavier processor. The windows on the screen exhibit (1) the depth image, (2) the input frame for the SLAM system, (3) the semantic segmentation result and (4) the 3D map.

3.2 Software Components

Our approach is based on the OpenVSLAM [16] framework, which provides our system with robust mapping and localization in real time. We feed the OpenVSLAM with color images captured by the RGB camera. The tracking module estimates the camera pose of the current frame. We assume that the area covered by the camera reveals an interesting direction for a visually impaired user. We utilize the 3D landmarks generated by the mapping module in the current frame to calculate the distance between the user and the objects. In this scenario, the camera center is considered as the location of the user. We choose the area within a distance of half a meter to one meter as the target area. In parallel, our system takes the color images and the depth images to the image segmentation component of our system. Subsequently, we acquire the semantic labels of the target area from the segmentation results. Normally, our target area covers several semantic labels. It is trivial to determine the final label by choosing the most frequent label of this area. The semantic segmentation approach is derived from the RFNet [17], a real-time fusion network. It provides the system with fast inference and high accuracy of semantic segmentation by fusing RGB-D information from the camera, as shown in Fig. 3.

Training of the Computer Vision Model. We trained the RFNet with the SUN RGB-D indoor scene understanding benchmark suit [15]. SUN RGB-D contains 10355 RGB-D images with dense indoor semantic labels of 37 classes. We resized all images to 480×640 and applied data augmentation during the training. The pixel classification accuracy is 15.5% on 2000 test images. Fig. 3 shows some results of RFNet on the SUN RGB-D dataset. We use Intel RealSense R200 as the input device for both the SLAM part and segmentation part. The

stream resolution is 480×640 with a frame rate of 60 fps. As shown in Table 1, We achieve approximately 69.3ms/frame inference speed of the semantic segmentation and 59.9ms/frame tracking speed of the SLAM system, which is fast for navigation assistance on the portable embedded processor. Fig. 4 shows the mapping results in small rooms. When the system detects objects near the user, the system generates a audio feedback with semantic information every 1 second.

3.3 Interaction of the components

Figure 1 shows the general interaction of the components of our proposed system. In order to support reliable obstacle avoidance, we keep searching for the nearest landmarks to the camera center in the map. When the distance is reaching a certain interval (i.e., between 0.5m and 1m), the user is informed of the semantic label of the target area. We embed this information in a sentence (e.g., "A table is in front of you") and send it to a text-to-speech module to generate audio feedback for the user.

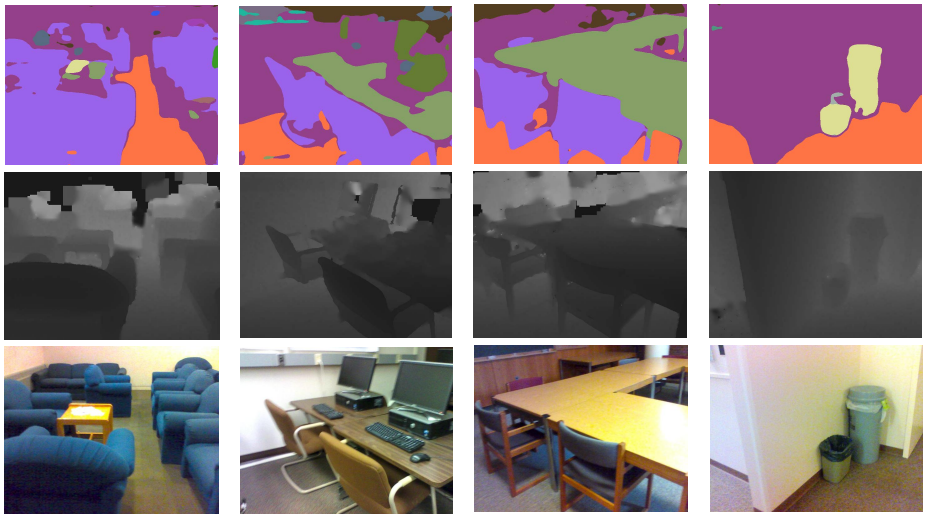

Fig. 3. Semantic segmentation results on SUN RGB-D dataset. From top to bottom: semantic maps, depth maps and RGB images.

4 Pilot Study

Our system aims to enable navigation in unstructured indoor environments. Thus, the user must be made aware of impassable areas in their path, as well as possible obstacles. Hence, we focused our evaluation on the ability to detect

Table 1. System specifications and speed analysis.

Camera resolution	Camera fps	Inference speed	Mean tracking time
480×640	60fps	69.3ms/frame	59.9ms/frame

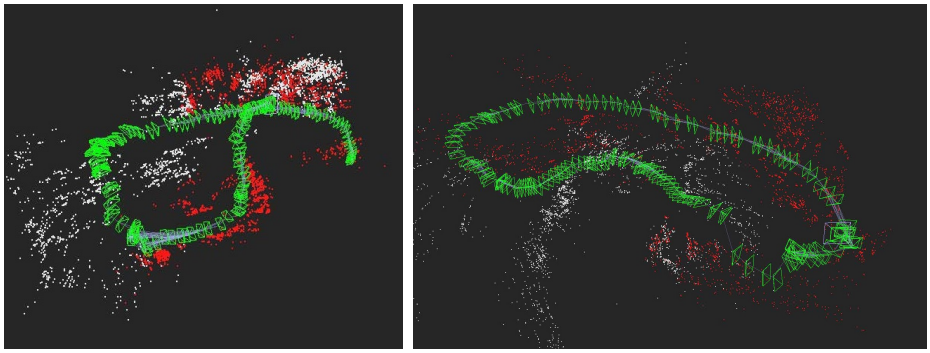

Fig. 4. Mapping results using the wearable glasses. The square trajectory on the left image indicates the result of walking around a table. The right image shows the walking trajectory along a corridor.

obstacles and objects blocking the path. However, we also evaluated our prototype through a user study where a blindfolded participant walked around a table with chairs and other small obstacles along the path. The goal of the task was to see if the user was capable of completing a circuit around the table while evading all possible objects. In this regard, the system test was successful. The user received timely audio feedback that warned him of all obstacles, and the test was completed without any collision with an obstacle.

On the other hand, the test was useful to identify some of the limitations of our device that impacted the user experience. On one side, the field of view of the camera did not cover all areas in front of the user. To prevent hazards from outside the camera frame, the user had to scan the environment by slightly moving their head. This task, however, was quite intuitive and posed no problem during the test, but it was noted that a device with a larger field of view would be advantageous for this application. A second limitation found was that due to the generalization of our model (see Fig. 5), the results of semantic segmentation differ under various environments. This resulted in a handful of times where obstacles that were not present on the scene were nonetheless notified to the user. While the user had no problem dealing with them, those diminish the confidence of users in the system, and thus further improvements in the semantic segmentation would benefit the user experience.

5 Conclusions

We presented an approach for visually impaired people to gain more mobility and orientation capacity in an indoor environment. The system makes it possible

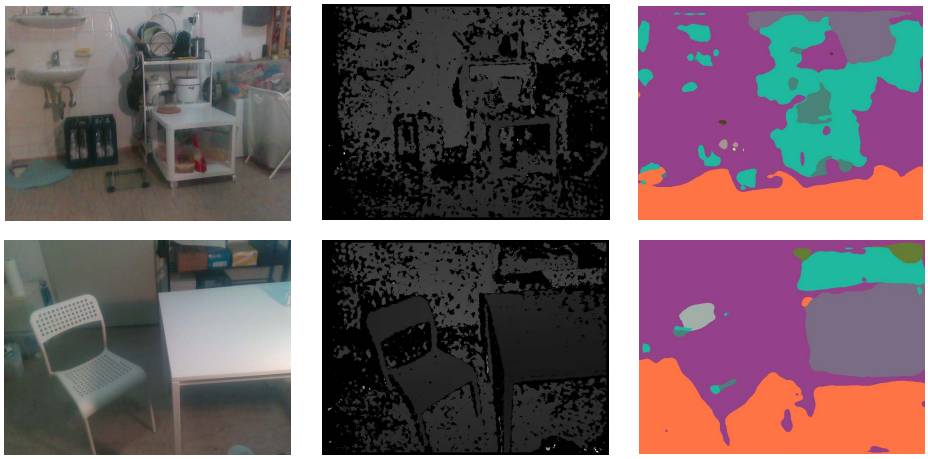

Fig. 5. Semantic segmentation results in our indoor environments.

to provide additional information that it is not easy to obtain with traditional mobility aids such as the white cane. Combined with semantic contents of the environment, the system can provide visually impaired people with different options of their actions (i.e., not only avoidance but also interaction).

In future work, we plan to use the semantic information to improve the localization further (i.e., to estimate what kind of room the user is currently located in). We also will test our system with persons with visual impairments to adapt the system to their special needs. Furthermore, we will robustify semantic perception in real-world domains [19] and improve the computational efficiency of the visual positioning system.

6 Acknowledgement

The work is partially funded by the German Federal Ministry of Labour and Social Affairs (BMAS) under the grant number 01KM151112. This work is also supported in part by Hangzhou SurImage Technology Company Ltd. and in part by Hangzhou KrVision Technology Company Ltd. (krvision.cn).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
2. Cadena, C., Carlone, L., Carrillo, H., et al.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on Robotics (2016)

3. Hu, W., Wang, K., Chen, H., et al.: An indoor positioning framework based on panoramic visual odometry for visually impaired people. Measurement Science and Technology (2019)
4. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: International Conference on Image Processing (2019)
5. Lin, S., Cheng, R., Wang, K., Yang, K.: Visual Localizer: Outdoor Localization Based on ConvNet Descriptor and Global Optimization for Visually Impaired Pedestrians. Sensors (2018)
6. Lin, Y., Wang, K., Yi, W., Lian, S.: Deep learning based wearable assistive system for visually impaired people. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
7. Liu, Q., Li, R., Hu, H., Gu, D.: Indoor topological localization based on a novel deep learning technique. Cognitive Computation (2020)
8. Martinez, M., Roitberg, A., Koester, D., et al.: Using technology developed for autonomous cars to help navigate blind people. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2017)
9. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics (2017)
10. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147 (2016)
11. Poudel, R.P., Liwicki, S., Cipolla, R.: Fast-scnn: fast semantic segmentation network. arXiv:1902.04502 (2019)
12. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems (2018)
13. Romera, E., Bergasa, L.M., Yang, K., et al.: Bridging the day and night domain gap for semantic segmentation. In: Intelligent Vehicles Symposium (2019)
14. Rosinol, A., Abate, M., Chang, Y., Carlone, L.: Kimera: an open-source library for real-time metric-semantic localization and mapping. In: International Conference on Robotics and Automation (2019)
15. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
16. Sumikura, S., Shibuya, M., Sakurada, K.: OpenVSLAM: A Versatile Visual SLAM Framework. In: Proceedings of the 27th ACM International Conference on Multimedia (2019)
17. Sun, L., Yang, K., Hu, X., et al.: Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images. arXiv:2002.10570 (2020)
18. Whelan, T., Salas-Moreno, R.F., Glocker, B., et al.: Elasticfusion: Real-time dense slam and light source estimation. The International Journal of Robotics Research (2016)
19. Yang, K., Bergasa, L.M., Romera, E., Wang, K.: Robustifying semantic cognition of traversability across wearable rgb-depth cameras. Applied optics (2019)
20. Yang, K., Wang, K., Bergasa, L.M., et al.: Unifying terrain awareness for the visually impaired through real-time semantic segmentation. Sensors (2018)