Künstl Intell
DOI 10.1007/s13218-015-0407-7

TECHNICAL CONTRIBUTION

A Camera-Based Mobility Aid for Visually Impaired People

Tobias Schwarze[1] · Martin Lauer[1] · Manuel Schwaab[2] · Michailas Romanovas[3] · Sandra Böhm[4] · Thomas Jürgensohn[4]

Received: 12 August 2015 / Accepted: 5 October 2015

Abstract We present a wearable assistance system for visually impaired persons that perceives the environment with a stereo camera system and communicates obstacles and other objects to the user in form of intuitive acoustic feedback. The system is designed to complement traditional assistance aids. We describe the core techniques of scene understanding, head tracking, and sonification and show in an experimental study that it enables users to walk in unknown urban terrain and to avoid obstacles safely.

Keywords Wearable navigation aid · Visual impairment · Binocular vision · Sonification

✉ Martin Lauer
martin.lauer@kit.edu

Tobias Schwarze
tobias.schwarze@kit.edu

Manuel Schwaab
manuel.schwaab@hsg-imit.de

Michailas Romanovas
michailas.romanovas@dlr.de

Sandra Böhm
boehm@human-factors-consult.de

Thomas Jürgensohn
juergensohn@human-factors-consult.de

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[2] Hahn-Schickard (HSG), Villingen-Schwenningen, Germany

[3] German Aerospace Centre (DLR), Neustrelitz, Germany

[4] Human-Factors-Consult GmbH (HFC), Berlin, Germany

1 Introduction

Moving through unknown urban areas is highly challenging for visually impaired persons. The white cane as traditional navigation aid allows to sense the space directly in front of the person, however, it does not provide any information about objects which are farther away than one meter. Moreover, the white cane does not allow to detect objects like overhanging branches of trees and open windows, which pose great danger to visually impaired persons. A guide dog as more intelligent assistive technology is unaffordable for most blind persons. Therefore, the development of intelligent technical aids would be an important contribution to increase the autonomous mobility of these persons.

Early approaches towards assistance systems for the visually impaired trace back to the 1960s, when experiments with wearable ultrasonic range sensors were carried out (e.g. [8, 13]). Several approaches have been developed in recent years [3, 7]. Most of these systems notify the user about non-traversable directions in the scene [4, 6], or they guide the user into walkable free space [10, 16]. Both options do not require a deep technical level of scene understanding, but the interpretation of the situation is left to the user. However, the correct interpretation of the acoustic or haptic feedback is hard. None of these systems did make it to a product so that the white cane and the guide dog are still the only mobility aids for visually impaired persons.

Meanwhile, the progress in the field of environment perception and scene understanding has reached a level that allows to build useful applications in many domains like robotics, driver assistance systems, and surveillance. This progress has motivated us to design a new kind of assistance aid for visually impaired persons. In contrast to

Published online: 15 October 2015

previous systems, it interprets its environment so that it can provide information on a high level of abstraction which facilitates its usage. However, the requirements for such a system are high since it must be wearable, lightweight, and unobtrusive. The environment must be recognized robustly and the user must be notified in an intuitive way about its surrounding without affecting its natural way of sensing.

In this paper we want to describe the basic design of this assistance system, introduce the main algorithms for scene understanding, and describe the way in which the user is informed acoustically about its environment. We will show with an experimental study how blind persons can benefit from it.

2 Requirements and System Design

A mobile navigation aid must provide reliable information in an intuitive form to the user. Since the visual feedback cannot be used in our context, our system is based on the idea to provide acoustic feedback, i.e. to encode relevant information in terms of specific sounds which are provided to the user. Based on this idea we can augment the natural acoustic environment by an artificial acoustic world, which describes the local environment of the user. Figure 1 exemplifies this idea.

The perception of the environment is based on a head-worn binocular camera system. The binocular camera provides texture information as well as distance measurements. Both kinds of information are necessary to be able to determine the geometrical layout of the environment, to detect obstacles, to track objects over time, and to distinguish different kinds of objects. Moreover, the head-mounted camera allows us to perceive the environment from a natural point of view, and it allows the user to control in which direction the camera points. This might be useful to sense information from a specific direction.

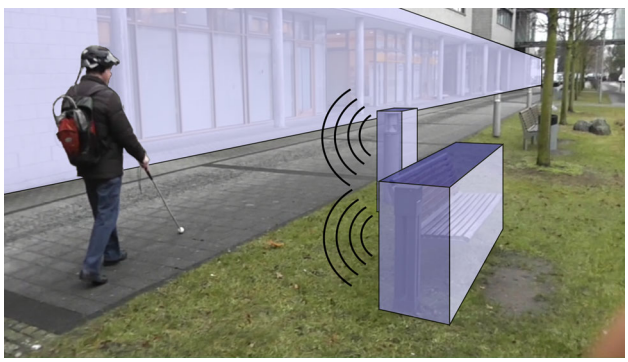

Fig. 1 Illustration of the basic idea of a navigation aid for blind persons. The invisible world is translated to a world of spatially arranged sound sources, which describe objects next to the user. The sounds are generated by earphones

The latencies between the perception of an object with the camera and the acoustic notification of the user sums up to around 100 ms. A direct coupling between perception and sonification would confuse the user since head movements during the processing time would lead to erroneous localization of objects. Therefore, we augment the system with an inertial measurement unit (IMU) which tracks the head position with minimum latency. This allows to convert the environment as perceived by the camera into the coordinate system which reflects the head pose at the time of sonification. Moreover, the knowledge of the head movement over time allows to accumulate knowledge that was perceived in the whole sequence of camera images.

The implementation of this system design was done on a prototype level using a bicycle helmet that was equipped with a binocular camera and an IMU (see Fig. 2). Earphones are integrated into the helmet with a small distance to the user's ear to avoid blocking the natural surrounding sound. Interviews revealed how crucial this fact for visually impaired people is, since otherwise important information about the environment would be lost.

All the computations were done on a notebook computer, which was worn in a backpack.

3 Binocular Environment Perception

One of the technical core challenges in the development of the assistance aid was to develop algorithms for the camera-based environment perception which are reliable and efficient enough to be operated in real time on a wearable system with limited computation power.

In comparison to traditional travel aids like the laser cane [1] it is not sufficient to detect the walkable free space, we rather need to understand what is limiting the free space. In urban environments a large variety of objects might occur at the same time in the vicinity of the user, e.g. buildings, parking cars, bicycles, trees and bushes, shop displays, chairs and tables, stairs leading up or down, and moving pedestrians. Only a small part of this information

Fig. 2 Prototype system built out of a bicycle helmet, a binocular camera, earphones, and an IMU (IMU not visible)

can be communicated to the user. Hence, it is necessary to aggregate the information and generate an abstract representation of the environment in which we ignore the irrelevant details.

A large part of inner-urban scenes is covered by walls, facades, fences, or bushes. These structures can be understood as the background of the scene. Scene background is always static. It provides valuable high-level context knowledge that can be used for obstacle detection and scene understanding. Furthermore, the alignment of facades is a valuable orientation hint for visually impaired users. Other objects like obstacles and other traffic participants are placed in front of the background and define the foreground. Foreground objects are typically smaller than background objects and they might move themselves. These observations motivated us to model the geometric scene background structure independently of movable foreground objects.

The underlying environment model can best be described as a blocks-world composed of planar surfaces that represent the scene background and independently moving boxes, which represent foreground objects and obstacles. This provides information on a level beyond any traditional means of mobility assistance for the visually impaired.

The task of the vision system is to build and maintain this environment model while the user is moving through the scene. A dense disparity estimator provides the basis for extracting the geometric scene background structure (Sect. 3.1) within which we detect and track generic obstacles (Sect. 3.2). The disparity estimator is accurate enough for our purpose within a range of approximately 10–20 m. We represent all measurements in a global reference frame to be able to deal with objects that are temporarily leaving the field of view. We estimate our position within that frame using a new combination of visual odometry and IMU (Sect. 4).

3.1 Scene Geometry as Background

Our scene geometry model consists of a composition of planar surfaces in global 3d-space. Determining these planes is a multi-model fitting problem that we treat with a combination of multi-model RANSAC plane fitting and least-squares optimization. Planes are estimated in the non-Euclidean disparity space before they are transformed into Euclidean 3d-space [14].

We apply a set of geometric constraints to regularize the plane estimation process. This includes a constraint that all planes are orthogonal to a common ground plane, which we estimate in a first step. Furthermore, in structured urban environments we make use of vanishing directions, which are aligned with buildings and the ground and which correspond to normal vectors of facades. This enables us to track planes even under heavy occlusion or when a plane temporarily becomes unobservable due to the camera orientation [15].

3.2 Foreground Object Detection

We are not interested in distinguishing specific foreground objects like cars, bicycles, pedestrians, etc., but we want to represent all kinds of foreground objects in a unified representation. Therefore, we define a generic foreground object as an assembly of spatially neighboring points above the ground plane which do not belong to parts of the scene background. After removing all points from the disparity data which belong to the background, we partition the remaining points into small segments in which all points are located close to each other. We apply single linkage agglomerative clustering and use as distance measure the difference of the disparity of two points. Although quite simple, this method yields an over-segmentation of the scene with small computational expense.

In the next step we group the segments into objects. We apply a reasoning process based on the previously instantiated objects in the environment model. We assign each segment to the closest object based on two features, (a) the overlap in image space of a segment with the projected outline of an object, and (b) the Mahalanobis distance in 3d-space between segment and object. The group of segments that was assigned to the same object forms an observation for this object.

Finally, we update a Kalman filter with the observed objects. The filter estimates the bounding box dimensions, the position, and the velocity for each object (see Fig. 3). Segments which could not be assigned to any existing object are used to initialize new objects in the environment model. Objects are deleted when they are in the field of view, however, have not been observed for a number of consecutive frames.

4 Egopose Estimation

As mentioned in Sect. 2 the knowledge of the exact head pose and movement is very important to provide consistent acoustic feedback to the user. Moreover, it allows to track the environment model over time and to estimate the movement of observed objects consistently. The task of head tracking can be solved in two ways, (a) by using the IMU, or (b) by camera-based visual odometry. The IMU has the big advantage that it operates almost delay free with a high update rate of 400 Hz. It is based on 3-axis accelerometers, gyroscopes and magnetic field sensors. It provides the orientation angles and the rotation rates of the head reliably even in the case of strong head movements.

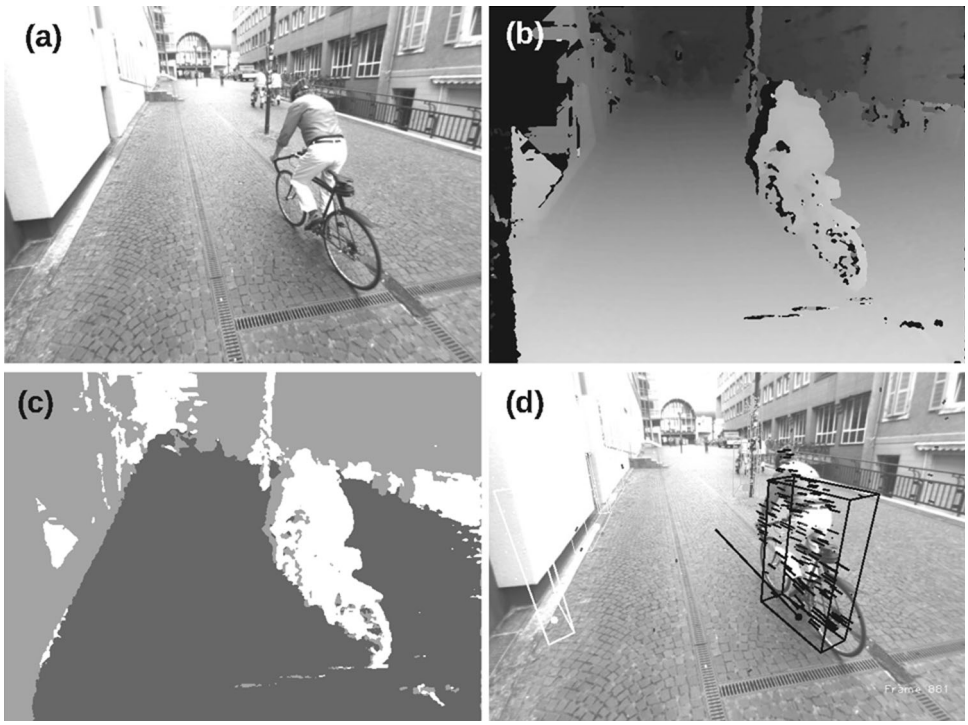

Fig. 3 Illustration of the camera image analysis. **a** Camera image. **b** Dense disparity map. *Gray*-values encode the distance from *light* (large disparity) to *dark* (small disparity). **c** Scene background classified into ground plane (*dark gray*) and facades (*light gray*). **d** Objects with aligned bounding boxes. Sparse feature flow vectors of the cyclist are *highlighted*. The *line on the ground* depicts the estimated motion

However, it does not provide reliable estimates of the head position and translation. In contrast, visual odometry provides both, but it suffers from larger delays of 50–100 ms and a small update rate of 10 Hz. It is based on the idea to track the movement of salient points in the images over time. Assuming that the majority of those points is located on static background objects, we can derive the ego motion of the camera [5].

To overcome the shortcomings of both methods we combined them in an integrated approach [11].

The different update rates of the two sensors and the need to achieve delay free head pose estimates for sonification required a complex modeling of the estimation task. The core of our approach is a common orientation-filter based on inertial and magnetic field measurements [9]. The filter state contains the orientation represented by a quaternion and the bias of the gyroscope.

Similar to a gyroscope, visual odometry also measures the rotation. Up to statistical errors, integrating the gyroscope over the time interval between the capturing times of subsequent video frames yields the same measurement as the rotation calculated by visual odometry. Thus, while visual odometry does not provide any new information concerning the rotation, it can statistically improve the orientation estimates and can also help to detect and correct irregular measurements, e.g. in the case of magnetic distortion. This fusion scenario with the involved measurements is sketched in Fig. 4.

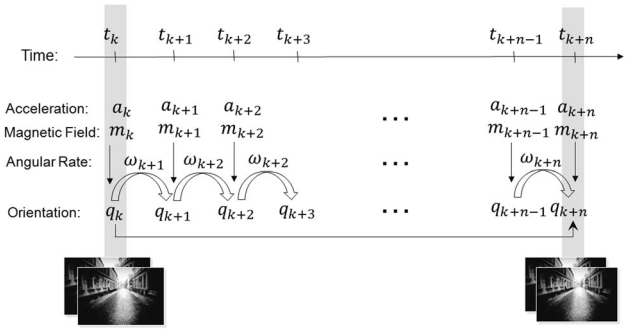

Fig. 4 Measurements of IMU and visual odometry in their temporal order. At points in time t_k and t_{k+n} we obtain rotation estimates from visual odometry while in between we only obtain updates from the IMU

The *stochastic cloning approach* proposed in [12] allows us to consider incremental measurements which relate states of non-consecutive points in time. In our setup, we obtain IMU measurements for all points in time and visual odometry measurements only for $t_k, t_{k+n}, t_{k+2n}, \ldots$ At time t_k the state is augmented with a clone of it, which will remain an estimate for time t_k. At time t_{k+n} the rotation calculated by visual odometry becomes available and can be related to the corresponding incremental inertial measurement, which can be calculated as the difference between the present orientation estimate and the cloned orientation. This provides the innovation step of the filter.

The position of the head is determined outside of the filter solely based on visual odometry.

The filter delivers the head orientation with a frequency of 400 Hz and the head position with 10 Hz, which allows almost delay free prediction of objects. Figure 5 provides the estimated trajectory for a walk of approximately 500 m.

5 Acoustic Feedback

Based on the environment model determined from the camera images and the egopose estimate, the position of objects and obstacles in the vicinity of the user can be calculated relative to the current head pose. Each object is represented by a sound source which encodes the spatial location through binaural rendering. Binaural rendering refers to the technique of creating sounds that can be localized in direction and distance using headphones [7]. It requires the head-related transfer function (HRTF) of both ears to render a sound depending on the direction and distance as if it was naturally distorted through the outer ear and delayed according to the ear distance [2]. An acoustic image of the environment arises which the user can freely interact with by turning or moving the head. In order to be accepted by the users, it is important that the cognitive load of interpreting the acoustic feedback remains small. Here, it was important to carefully select sounds for sonification which are intuitive to understand.

To avoid confusion, the artificial sounds need to be clearly distinguishable from natural environmental sounds.

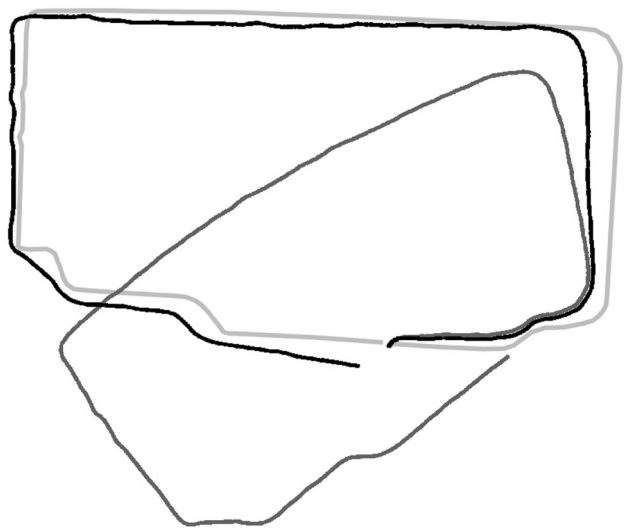

Fig. 5 *Top view* of the estimated trajectory for a walk of 500 m length using pure visual odometry (*dark gray*) and the stochastic cloning Kalman filter (*black*). The ground truth is depicted in *light gray*

Additionally, they need to be pleasant to listen to and they need to encode semantic information about the object, e.g. its kind, its motion or its potential danger in an intuitive way. Furthermore, we decided to keep the number of different sounds small so that the required time to learn the semantics of each sound remains small. Therefore, we had to group objects into a small number of categories for which we assigned a specific sound. Objects within the same category should exhibit some form of similarity. The criteria on which to base this similarity are largely subjective. Furthermore, we expected them to differ strongly between visually impaired and non-visually impaired persons.

To find a good solution we conducted a study with 26 visually impaired persons. They were given the task to group a set of 40 named obstacles (like poles, trees, traffic signs, flowerbeds, drop-offs and holes, stairs, bushes, benches) into arbitrary categories which needed to be defined by the participants themselves. This categorization was used to extract a measure of similarity between all pairs of objects, which we used to cluster the 40 obstacles into categories that best resemble the participants' individual categorization. These categories are (a) wide objects (e.g. ticket machines, cars, benches), (b) pole-like objects, (c) elevated objects (awnings, barriers), (d) non-traversable areas (flowerbeds), (e) approaching vehicles, (f) approaching persons, (g) drop-offs and holes, (h) high curbs and stairs, and (i) crosswalks.

Besides being semantically meaningful, it is important that the sounds can easily be located by the users. Localizing sounds in terms of their direction and distance requires the sound to be composed of a wide frequency spectrum. The human outer ear distorts the frequency spectrum depending on the sound direction to allow sound source localization. This learned relationship between sound direction, distance and perceived sound appearance has been shown to improve with a priori known sounds. However, sounds exhibiting wide frequency spectra often conflict with the requirement of comfort. Especially high frequencies can only be used carefully. Moreover, the sounds selected should be in some kind of harmony with each other since usually multiple sounds will be rendered simultaneously.

We carried out experiments in a sound simulator to find appropriate sounds. 18 real and synthesized sounds were played to 30 persons, 15 of them visually impaired. The participants were equipped with headphones in which each sound was played from 20 different directions in a 140° field of view in front of the head. The participants were asked to point a marker towards the virtual sound source which we used to measure the localization error. In a second step we evaluated the comfort of each sound by asking a grade between 1 and 6. Furthermore, we asked the

participants to assign the sounds to the previously defined categories.

In urban environments there are typically many more objects in the vicinity of a person than the number of sounds that are distinguishable simultaneously. Therefore, we developed a strategy which selects the three most relevant objects in terms of distance and relative position to the walking direction of the user. Based on this selection, sound sources are virtually placed to their position. The sounds are convoluted with the HRTF of the users and their amplitudes are adapted to the distance of the objects.

Fig. 6 The second scenario used in the experiments

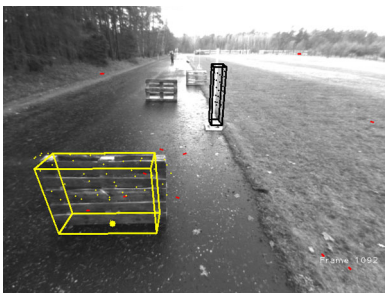

Fig. 7 Camera view with obstacles detected in the obstacle course. Two sounds are emitted in this situation

6 Experiments

In order to check whether the overall system design is useful for potential users we performed experiments with a set of visually impaired persons under realistic conditions. Since the output of the system influences the behavior of the user, it was important to test the whole control loop, which includes the perception of the environment with the camera system, the sonification of the sensed objects, and the behavior of the human user.

The experiment was carried out with 8 visually impaired persons at the age of 20–50 years. Five of the persons are independently mobile (four with a white cane, one with a guide dog). The remaining three persons are more restricted in their mobility and rely on the support of others.

As a first part of the experiment we set up a training scenario consisting of two big obstacles on an open field. The test persons were asked to navigate towards the obstacles and pass between them to validate the sound localization concept. This phase took on average 10 min.

The second part of the experiment consisted of a more complex scenario with the purpose to find out whether the persons could use the acoustic information to redirect their path of travel in order to avoid collisions. The task was to navigate along the turf between a pavement and lawn as orientation guideline using the white cane. We placed different obstacles (low boxes, high poles, and one overhanging obstacle) with a few meters distance along this path, some directly on the path, some to the left and right (see Figs. 6, 7). The system classified these obstacles into flat, pole-like, dynamic and overhanging obstacles, each with a specific sound. The overall experience with the system took around 30–45 min per participant.

The evaluation of the experiment was done in the form of interviews with the test persons.

We observed that the feedback principle was immediately clear to all participants. The first training scenario suited well to get familiar with the concept of spatial sounds. In the beginning of the second scenario, the users tended to stop walking whenever a new obstacle was sonified and turn their head in order to confirm the sound direction. After some minutes of experience, new obstacles caused them to decrease their walking speed until the white cane touched the obstacle. It turned out that it was difficult to assess the distance to objects based on the amplitude of sound alone. The training phase was too short to develop a proper sense for the relationship between sound volume and distance. This effect was intensified by the use of four different sounds which were perceived differently loud by the individual participants.

During a second and third walk through the obstacle course some test persons had developed a sense that allowed them to avoid obstacles before they could touch them with the white cane. However, they reported that it was very difficult to assess the object extension since it was not reflected in the feedback.

Two participants who felt instantly very comfortable with the system were using it afterwards to walk freely on the sidewalk of a nearby road. We asked the users to verbally communicate the recognized obstacles. The test persons were able to follow the sound that was virtually emitted from a guiding person which was walking in front of them.

Most test persons were skeptical about artificial spatial sounds after participating in our simulator studies. However, the experiments under realistic conditions turned out

to be a positive experience for these users. The acoustic overlay did not cause the users to feel limited in their natural sense of hearing. The concept of informing the persons about the environment rather than generating navigation clues was received positive. Devolving the decision making to the assistance system is not desired by most visually impaired persons, they want to stay in control. All participants could imagine to apply such an assistance system in future.

7 Conclusions

Improving individual mobility for visually impaired persons was the overall goal of our work. With the development of a system that combines environment perception with a binocular camera, head tracking with visual odometry and an IMU, and sonification of relevant objects in the vicinity of the user, we were able to overcome limitations of present assistive aids. A core challenge in this development was to bring together the limited technical possibilities of a wearable platform with the demands of the users.

The perception of the environment was based on the modeling of a geometric scene background, which was estimated from the binocular camera images using robust estimation techniques. The foreground was modeled as a set of generic, potentially movable objects. Those objects are tracked over time and contribute to a compact representation of the local environment of the user, which serves as the base for sonification. A second core technique was the development of a robust head tracking algorithm that combines an inertial measurement unit and visual odometry to achieve high frequent, robust head pose estimates with small latency. Unlike solutions that are solely based on inertial sensors, our solution is able to estimate the head orientation as well as its position.

The technical solution for sonification was based on a set of simulator studies that were designed to identify which way of sonification suits best to the needs and wishes of the visually impaired users and which sounds allow for an intuitive interpretation of the virtual acoustic world. Based on these results we developed a sonification concept that notifies the user about potentially dangerous objects. However, the decision into which direction to walk is still left to the user which increases acceptance of such a system.

Our final experiments under realistic conditions gave evidence that the assistance system is useful for visually impaired persons and that it can be used in an intuitive way. It extends the sensing range from approximately 1 m (white cane) to 10–20 m and, thus, allows the user to avoid obstacles and dangerous situations earlier. Moreover, it allows to detect obstacles like tree branches or barriers, which cannot be recognized with the white cane.

We plan to miniaturize the system and integrate the sensors into an eyeglass frame so that the helmet and the backpack can be removed. This will allow visually impaired persons to use it in everyday life. Moreover, once a user is wearing such a system, it can be used to provide additional information like the position of stairways, bus stops, and zebra crossings or the text on signs and guideposts.

Acknowledgements This work was supported by the German Federal Ministry of Education and Research within the research project *OIWOB*, Grant no. 13EZ1126.

References

1. Benjamin J, Malvern J (1973) The new C-5 laser cane for the blind. In: Carnahan conference on electronic prosthetics, pp 77–82
2. Blauert J (1997) Spatial hearing: the psychophysics of human sound localization. MIT Press, Cambridge
3. Dakopoulos D, Bourbakis N (2010) Wearable obstacle avoidance electronic travel aids for blind: a survey. IEEE Trans Syst Man Cybern Part C Appl Rev 40(1):25–35
4. Fajarnes GP, Dunai L, Praderas VS, Dunai I (2010) CASBLiP—a new cognitive object detection and orientation system for impaired people. Trials 1(2):3
5. Geiger A, Ziegler J, Stiller C (2011) Stereoscan: dense 3d reconstruction in real-time. In: IEEE intelligent vehicles symposium, pp 963–968
6. Gonzalez-Mora J, Rodriguez-Hernandez A, Burunat E, Martin F, Castellano M (2006) Seeing the world by hearing: virtual acoustic space (VAS) a new space perception system for blind people. In: Information and communication technologies (ICTTA '06), vol 1, pp 837–842
7. Hermann T, Hunt A, Neuhoff JG (eds) (2011) The sonification handbook. Logos Publishing House, London
8. Kay L (1964) An ultrasonic sensing probe as a mobility aid for the blind. Ultrasonics 2(2):53–59
9. Kraft E (2003) A quaternion-based unscented Kalman filter for orientation tracking. In: 6th international conference of information fusion, vol 1, pp 47–54
10. Pradeep V, Medioni G, Weiland J (2010) Robot vision for the visually impaired. In: 2010 IEEE computer society conference on computer vision and pattern recognition workshops, pp 15–22
11. Romanovas M, Schwarze T, Schwaab M, Traechtler M, Manoli Y (2013) Stochastic cloning Kalman filter for visual odometry and inertial/magnetic data fusion. In: 16th international conference on information fusion, pp 1434–1441
12. Roumeliotis S, Burdick J (2002) Stochastic cloning: a generalized framework for processing relative state measurements. In: IEEE international conference on robotics and automation (ICRA '02), vol 2, pp 1788–1795
13. Russel L (1965) Travel path sounder. In: Rotterdam mobility research conference
14. Schwarze T, Lauer M (2015) Geometry estimation of urban street canyons using stereo vision from egocentric view. In: Informatics in control, automation and robotics. Lecture notes in electrical engineering, vol 325, pp 279–292. Springer International Publishing, Berlin

15. Schwarze T, Lauer M (2015) Robust ground plane tracking in cluttered environments from egocentric stereo vision. In: 2015 IEEE international conference on robotics and automation, pp 2442–2447
16. Shoval S, Ulrich I, Borenstein J (2003) Navbelt and the guide-cane. IEEE Robot Autom Mag 10(1):9–20

Tobias Schwarze studied computer engineering at Berlin Technical University and is working as researcher at Karlsruhe Institute of Technology in the area of camera-based environment perception for wearable systems.

Martin Lauer studied computer science at Karlsruhe University and obtained his Ph.D. from Osnabrück University. He is working as research group leader at Karlsruhe Institute of Technology in the areas of computer vision, machine learning, and autonomous driving.

Manuel Schwaab studied computer science at Hochschule Furtwangen University. Until 2012 he was working with Fraunhofer IIS and is now a researcher at Hahn-Schickard in the Sensors and Systems group. His research interest include image and video processing, digital filtering and optimal estimation and control.

Michailas Romanovas studied electronics engineering at Vilnius Technical University (Lithuania) and Microsystems Engineering at Hochschule Furtwangen University. Until 2014 he was working at Hahn-Schickard and is now with the German Aerospace Center. His research interests include inertial sensors and systems, statistical signal processing, localization systems, estimation algorithms and sensor fusion.

Sandra Böhm studied psychology and education at Erfurt University and Human Factors at Berlin Technical University. She is working as expert for design, evaluation, and accessibility of human-machine-interfaces at Human-Factors-Consult GmbH in Berlin.

Thomas Jürgensohn studied electrical engineering and obtained his Ph.D. from Berlin Technical University. He is founder and manager of the Human-Factors-Consult GmbH in Berlin and in parallel Professor for human factors in vehicular technology.