An Intuitive Mobility Aid for Visually Impaired People based on Stereo Vision

Tobias Schwarze and Martin Lauer
Karlsruhe Institute of Technology (KIT)
{tobias.schwarze,martin.lauer}@kit.edu

Manuel Schwaab
Hahn-Schickard (HSG)
manuel.schwaab@hsg-imit.de

Michailas Romanovas
German Aerospace Centre (DLR)
michailas.romanovas@dlr.de

Sandra Böhm and Thomas Jürgensohn
Human-Factors-Consult GmbH (HFC)
{boehm,juergensohn}@human-factors.de

Abstract

We present a wearable assistance system for visually impaired persons that perceives the environment with a stereo camera and communicates obstacles and other objects to the user. We develop our idea of combining perception on an increased level of scene understanding with acoustic feedback to obtain an intuitive mobility aid. We describe our core techniques of scene modelling, object tracking, and acoustic feedback and show in an experimental study how our system can help improving the mobility and safety of visually impaired users.

1. Introduction

People with severe visual impairment are faced with huge challenges when moving through unknown environments. For many people independent movement is restricted to well known areas. The traditional white cane allows to sense the space directly in front of the person, but it does not provide any information about objects further away. Overhanging objects like tree branches or open windows, which pose great danger, cannot be sensed. Guide dogs as the most auxiliary assistive aid are unaffordable for most blind persons. The development of intelligent and affordable technical mobility 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. [12, 20]). Several approaches have been developed in the recent years [6]. Most of these systems notify the user about non-traversable directions in the scene [8, 17, 10], or they guide the user into walkable free space [23, 16]. Both options do not require a deep technical level of scene understanding. Either the difficult task of correctly interpreting the haptic or acoustic feedback is left to the user [15], which can cause substantial cognitive load, or the navigation is completely taken over by the system.

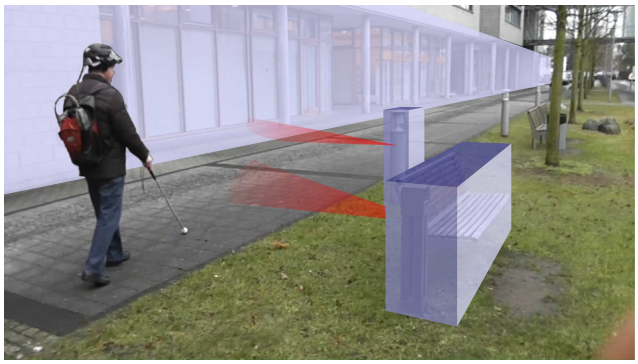


Figure 1: Our assistance system detects obstacles in the surrounding and transmits them to the user through spatial sounds.

Meanwhile, the progress in the field of environment perception and scene understanding becomes visible with the launch of intelligent applications and systems as seen in robotics, driving assistance, or surveillance. These systems are able to understand different aspects of their environment, they detect and track objects, assess risks and act accordingly. This has motivated us to develop an assistance aid which interprets its environment in order to offer feedback on a high level of abstraction to the user. This facilitates the usage significantly, but also entails high requirements since the system must be wearable, lightweight, and unobtrusive. Furthermore, the sensed information must be conveyed in an intuitive manner to the visually impaired user, which does not interfere the natural sensing.

In this paper we describe the basic design of such assistance system. We introduce the methods for scene understanding and show how acoustic feedback is applied to intuitively inform the user about its environment. We report

Figure 2: Prototype setup built out of a bicycle helmet, a binocular camera, headphones and an inertial measurement unit (IMU).

on a experimental study which shows how blind persons can benefit from such system.

2. Requirements and System Design

A mobile navigation aid must provide reliable information in an intuitive form to the user. The purpose of our mobility aid is to inform the user about objects in the local environment. This information shall enable the person to plan further ahead and avoid obstacles more prospectively. We convey the location and relevant semantic information of objects through binaural acoustic feedback. Using a headphone, the natural acoustic environment is augmented with sounds which can be localized in terms of direction and distance. This idea is exemplified in Fig. 1.

The environment perception (Section 3) of our system builds upon a head-worn binocular camera. This allows to perceive the environment from a natural point of view, and offers to easily direct the viewing range towards points of interest or objects of interaction [14]. These opportunities come along with the challenging task of operating under almost unconstrained and unpredictable camera motion. Based on the binocular camera images we detect and track generic static and moving objects and classify them into predefined groups of obstacles.

Special care needs to be given to the aspect of communicating the information to the user. The generated feedback needs to transport as much information as possible of the sensed environment, while it needs to be intuitive enough to be used without extensive training. Acoustic feedback has been shown to offer this potential [11], but is critical to apply in our context since visually impaired persons strongly rely on the hearing sense. Acceptance can only be expected if natural sounds are not blocked but carefully augmented with artificial sounds. Physically, we ensure this by placing the headphones slightly in front rather directly on the outer ear. Bone conduction speakers can be considered as an alternative, but require more careful setup and positioning. Furthermore, the selection and filtering of these sounds is critical for intuitive and pleasant usage. We summarize a range of experiments regarding this in Section 5.

The generated sounds are perceived relative to the head orientation. To create a realistic acoustic impression of a world-fixed sound source, the sound position relative to the moving head needs to be updated frequently and with minimal delay. These requirements are hard to achieve with a sequential computer vision process chain since computing times are by orders of magnitudes higher than the required acceptable delay between head and sound motion. For this reason an important building block in our system design is a module that estimates the head orientation with minimal delay fusing data from an inertial measurement unit (IMU) and camera (Section 4).

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.

A forward directed camera with limited aperture angle perceives only a small part of the environment surrounding the user. This might be sufficient to warn of imminent collisions but it is not sufficient to inform about objects next or even behind the user. To inform about such vanished objects we need to keep track of everything that was once seen.

In comparison with traditional travel aids like the laser cane [3] it is not sufficient to detect the walkable free space, we rather need to understand what is limiting the free space. In urban environments with buildings, parking cars, cycles, trees and bushes, shop displays, chairs and tables, stairs leading up and down, or moving pedestrians this is a large amount of information. Only a small part of this information can be communicated to the user. Hence, it is required to condense the information into an abstract representation of the environment, in which we ignore irrelevant details. This representation has to be flexible and expressive enough to depict the variety of different objects and their motion relative to the user, while it needs to be compact enough to keep the computational processing load small.

A large part of inner-urban scenes is covered by high walls, building facades, fences, or bushes. These kind of natural and man-made structures can be understood as a scene background in front of which small, independently positioned objects define a foreground. Foreground and background differ strongly in their extension and the fact that the scene background is always static. The scene background can provide high-level context knowledge that can be applied for obstacle detection. Furthermore, the alignments of building facades are valuable orientation hints for visually impaired users. Objects of interest are usually part of the foreground, which motivates us to model the geometric scene background structure independently of movable foreground objects in our environment representation.

The underlying environment model can best be described as a blocks-world composed of planar surfaces representing the scene background geometry and independently moving aligned boxes which represent foreground objects or obstacles. This provides information on a level beyond any traditional mobility aid 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 (Section 3.1) within which we detect and track generic obstacles (Section 3.2). To handle objects moving out of camera view we represent all measurements in a global reference frame and estimate our position within that frame using a combination of visual odometry and the inertial measurement unit (Section 4).

3.1. Scene Geometry

Our scene geometry model consists of a composition of planar surfaces in global 3d-space. Specifically, we keep track of a common ground plane, and structures like building facades, fences or bushes which constitute planes orthogonal to the ground plane.

3.1.1 Plane estimation

Measuring such planes is a multi-model fitting problem that we treat with a combination of multi-model RANSAC plane fitting and least-squares optimization. To avoid the non-linear stereo reconstruction error in Euclidean XYZ-space [21] we determine planes directly in disparity ($uv\delta$-)space as

$$\alpha u + \beta v + \gamma + \delta(u,v) = 0 \tag{1}$$

with (u,v) being the image coordinates and $\delta(u,v)$ the according disparity measurement. Given the camera calibration, the $uv\delta$ plane can be expressed in XYZ-space through a normal vector and camera distance as $\mathbf{p} = (\mathbf{n}, d)$

$$\mathbf{n}_x X + \mathbf{n}_y Y + \mathbf{n}_z Z + d = 0 \tag{2}$$

$$(n_x, n_y, n_z, d) \propto (\alpha f, \beta f, \alpha c_u + \beta c_v + \gamma, bf) \tag{3}$$

with focal length f, principal point (c_u, c_v) and stereo baseline b.

We apply the RANSAC scheme and generate plane hypotheses by repeatedly sampling planes through 3 random points. A plane is evaluated by counting the support points with point-to-plane distance $|\alpha u+\beta v+\gamma+\delta(u,v)|$ smaller than a disparity margin ϵ around the plane to find the best hypotheses.

Having obtained an initial solution we optimize the parameters using robust iterative least-squares estimation. The set of $uv\delta$ plane support points $(u_i, v_i, \delta_i),\ \ i = 1, \ldots, N$ is used to update the plane parameters by solving the linear system

$$\begin{pmatrix} \sum u_i^2 & \sum u_i v_i & \sum u_i \\ \sum u_i v_i & \sum v_i^2 & \sum v_i \\ \sum u_i & \sum v_i & N \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = - \begin{pmatrix} \sum u_i \delta_i \\ \sum v_i \delta_i \\ \sum \delta_i \end{pmatrix} \tag{4}$$

This estimation is applied a few iterations until no considerable update in the parameters remains.

3.1.2 Plane tracking

An estimated plane is transformed into an XYZ-plane $\mathbf{p}_{local}$ and added to the global environment model as $\mathbf{p}_{global} = \left(T_k^{-1}\right)^{\mathrm{T}} \mathbf{p}_{local}$ with T_k being the current ego-pose (cf. Sec. 4). In the next video frame we estimate T_{k+1} and use the predicted plane $\mathbf{p}^-_{local} = T^{\mathrm{T}}_{k+1}\mathbf{p}_{global}$ as initialization for the least-squares plane optimization.

Using this plane detection and refitting scheme our system is able to keep track of the ground plane. Special care needs to be taken in order to handle cases of heavy occlusion and situations in which the camera is temporarily pointed away from the surface. To handle theses situations we extract the vertical scene vanishing direction from the input images. It is represented by a vector $\mathbf{n}_V$ which coincides with the plane normal vector $\mathbf{n}$ estimated in the disparity data. Both information are fused in a Kalman filter and allow robust tracking of the ground in cluttered environments [22].

3.1.3 Vertical structures

To estimate planes which represent vertical scene structures we want to constrain the plane orientation to be orthogonal to a given plane (here the ground plane, represented by its normal vector $\mathbf{n}$), i.e. enforce the inner product of their Euclidean normal vectors to be zero, while optimizing the plane parameters in $uv\delta$ space. We seek the parameters which minimize

$$\underset{\alpha,\beta,\gamma}{minimize} \sum_{i=1}^{N} (\alpha \cdot u_i + \beta \cdot v_i + \gamma + \delta_i)^2$$
$$subject\ to\ \mathbf{n}_x(\alpha f) + \mathbf{n}_y(\beta f) + \mathbf{n}_z(\alpha c_u + \beta c_v + \gamma) = 0 \tag{5}$$

The constraint can be reformulated to

$$\gamma = \underbrace{-(\frac{f\mathbf{n}_x}{\mathbf{n}_z} + c_u)}_{=:k_1}\alpha + \underbrace{-(\frac{f\mathbf{n}_y}{\mathbf{n}_z} + c_v)}_{=:k_2}\beta \tag{6}$$

and inserted into the cost term. The resulting linear system for α and β is listed in equation (7).

To initialize planes vertical to the ground we apply a RANSAC variant in which vertical plane hypotheses are created from two $uv\delta$-points and the orthogonal ground

$$\begin{pmatrix} \sum u_i^2 - 2k_1 \sum u_i + N k_1^2 & \sum u_i v_i + k_1 \sum v_i + k_2 \sum u_i + N k_1 k_2 \\ \sum u_i v_i + k_1 \sum v_i + k_2 \sum u_i + N k_1 k_2 & \sum v_i^2 - 2k_2 \sum v_i + N k_2^2 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -k_1 \sum \delta_i - \sum u_i \delta_i \\ -k_2 \sum \delta_i - \sum v_i \delta_i \end{pmatrix} \quad (7)$$

plane normal vector. Planes are deleted from the environment model when they could not be remeasured for a few subsequent frames.

3.2. Generic Obstacle Detection and Tracking

In contrast to vision systems trained to detect obstacles of specific categories like pedestrians, cyclists or cars based on their appearance (e.g. [7]), our detection stage needs to be independent of obstacle appearance in order to detect arbitrary obstacles. Furthermore it is not sufficient to detect obstacles based on their motion [1, 5], since most parts of the scene which we are analysing are static.

In our case an obstacle can be defined as an assembly of spatially neighboured points, which do not belong to parts of the scene background geometry. Hence, detecting obstacles leads to a segmentation problem, in which each segment represents an object detection [24], which needs to be associated with known objects to be tracked over time [2]. Segmentation of low-resolution disparity data is a challenging problem and hardened by the facts that the number of objects is usually unknown and hardly any prior knowledge about their shape or size can be applied – foreground objects can be as tall as a truck and as small as a post.

We treat this problem as a combination of clustering and tracking before detection. To avoid merging objects close together into one detection we partition the foreground disparity points into small segments in which all points are clearly located close to each other. We apply single linkage agglomerative clustering and use as distance measure the difference of disparity of two points to yield an over-segmentation of the scene with small computational expense.

After segmenting we group the segments into objects. We apply a reasoning process based on the previously instantiated object tracks in the environment model. Each segment becomes assigned to the closest object based on two features, (a) the overlap ratio in image space of a segment with the projected contour of an object $\frac{A_{Segment} \cap A_{Object}}{A_{Segment}}$, and (b) the Mahalanobis distance in 3d-space between the ground plane projection of segment and the objects' center of gravity. The group of segments that was assigned to the same object forms an observation for this object.

The state of an object consists of its position, its velocity and direction, and a 3d aligned bounding box. An extended Kalman filter with constant velocity model updates the state with the observed objects. Furthermore, each object keeps a history of reconstructed 3d points of the past 20 observation. This allows us to determine the object contour in the current camera image for segment assignment and to measure the object extend in order to update the bounding

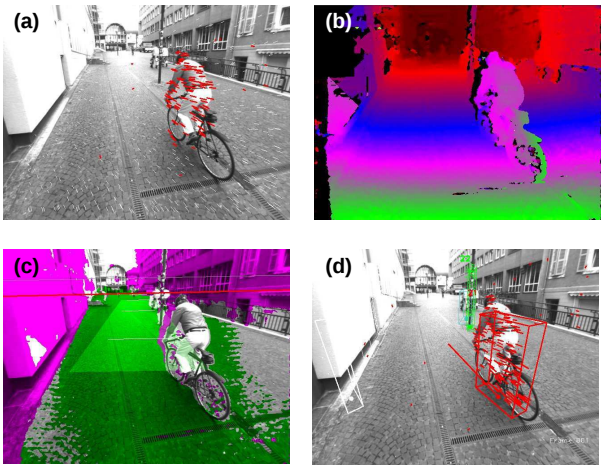


Figure 3: Vision algorithms applied in the system. (a) Feature flow of moving objects (b) dense disparity estimation (c) tracked ground plane (green) and building facades (purple) (d) tracked obstacles with aligned bounding boxes. The line indicates the predicted motion of the cyclist.

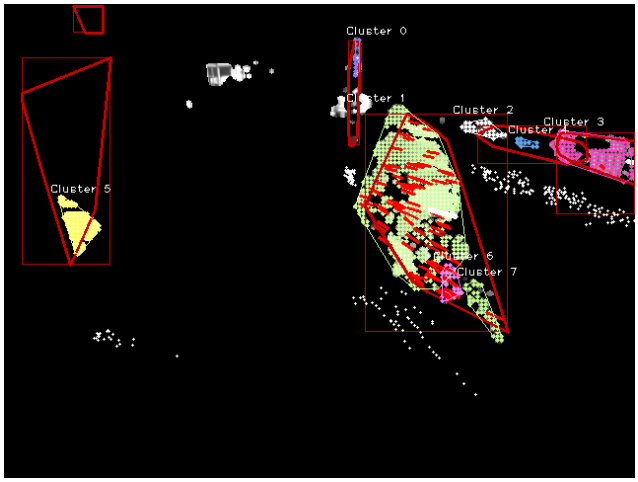
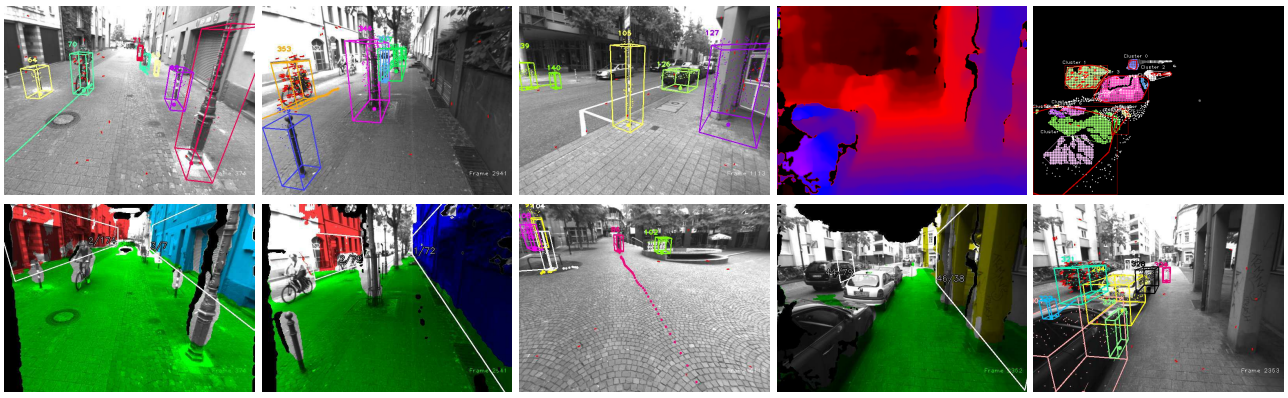


Figure 4: Disparity over-segmentation and object grouping for situation in Figure 3: Disparity segments are shown as colored dots. 3d points of existing obstacles are projected into the current view to find their contour in image space, here depicted as red polygons.



(a) (b) (c) (d)

Figure 5: Results of the vision algorithms. **(a)** and **(b)** Estimated scene geometry with ground surface and building facades shown as colored overlay (bottom row), and estimated obstacle bounding boxes (top row) with their velocity indicated by a line. **(c)** small posts detected in 12m distance (top), passing cyclist tracked up to 20m distance (bottom). **(d)** Cluttered disparity data of parking cars (top left) and its segmentation (top right) after removing the scene geometry (bottom left). Resulting obstacles (bottom right) with two small erroneous instantiated objects in the vicinity of the car (green and blue).

box dimensions. We align the boxes with the main principal component of their 3d points projected onto the ground plane (white dots in Figure 4).

We initialize objects in the environment model with segments which could not be assigned to any existing object. Objects are deleted from the representation when they are in the field of view, but have not been re-detected for a number of consecutive frames.

3.3. Results

The algorithms are embedded into a parallelized software framework in order to ensure a high data throughput. We capture images of 640x480 pixels with 30fps. The disparity is estimated using OpenCV semi-global matching at half-resolution while the egomotion is computed parallel by means of visual odometry (libViso2 [9]) with around 20fps. Using the disparity data we update the scene geometry and the foreground objects parallel with around 15fps on an i7 2.4 GHz dual-core notebook.

Figure 5 shows results of obstacle detection and geometry estimation. Depending on the size, objects are initialized into the tracking scheme in a distance between 10 (small posts) and 20 meters (cars) and tracked until they leave the field of view. Possible kinds of errors are close objects merged into a single track, or objects becoming segmented in multiple tracks (Fig. 5d). While the first is normally uncritical in our application, the second can lead to confusing feedback when single obstacles are reported with multiple sound sources. To avoid confusing the user with such ghost objects we apply temporal filtering and delay the initialization until an object was successfully observed 5 times.

As in all object detection methods based on surface models, the proper estimation of scene geometry is important to avoid wrong object initializations. Since we align all building facades relative to the ground, the estimation of the ground plane is the most significant. Our plane tracking scheme shows to be robust also in situations of temporary total occlusion or situations in which the user's head is temporarily pointing too far up. In these cases we predict the plane until it is visible again and new measurements can be made.

4. Egopose Estimation

The information of camera position and camera orientation w.r.t the environment model is used in two ways: First, it is needed to update the global environment model with measurements obtained from processed camera images. Secondly, a delay free localization enables to predict a local view onto the model which we use to generate feedback. The localization needs to be locally accurate enough to update the environment model with current measurements, but we do not require a globally consistent long term estimation.

This task of head tracking can be solved by camera based visual odometry. A few conditions need to be met here: scene illumination has to be sufficient to avoid motion blur and the captured scene needs to contain textured and apparent static parts. Rapid camera motion must be limited, which can not be guaranteed with uncontrolled head-worn cameras. A principal drawback in our application is the larger latency of up to 50ms, which can cause confusion when perceiving the artificial, environment-fixed sound sources.

As an alternative, an inertial measurement unit can be

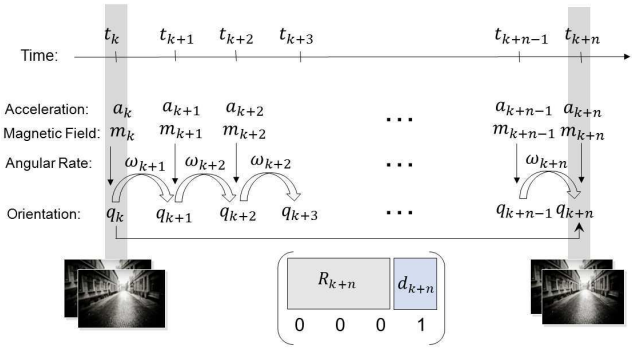


Figure 6: Measurements of IMU and visual odometry related to the orientation measurement. At points in time t_k and t_{k+n} we obtain rotation estimates from visual odometry while in between we only obtain updates from the IMU.

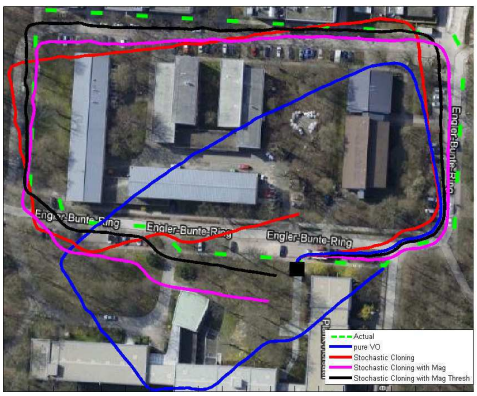


Figure 7: Estimated trajectory for a walk using pure visual odometry (blue) and the stochastic cloning Kalman filter (red). Further improvements are achieved by including measurements of the earths' magnetic field (pink and black). The ground truth is depicted with a green dashed line.

applied which offers incremental absolute orientation estimation with heading and roll compensation based on a combination of 3-axis accelerometers, gyroscopes and magnetic field sensors. Because of high processing rates this estimation is also precise under strong motion and can outperform camera motion estimation here. However, translational motion can not be directly measured and would require a global reference like GNSS for a drift-free estimate.

To overcome the shortcomings of both methods we combined both in an integrated approach. We can obtain delay-free orientation measurements even under strong head accelerations and benefit from accurate orientation and translation estimation through visual odometry in low-velocity situations. To fuse both measurements one has to deal with the different update rates and latencies of the two sensors. Our implementation builds upon libViso2 [9] to estimate the camera motion and fuses inertial measurements into the estimation using a Stochastic Cloning Kalman filter [18].

The core of the filter is a common orientation filter based on inertial and magnetic field measurements (compare e.g. [13]). The filter state contains the orientation represented by a quaternion q_k and the bias of the gyroscope. Similar to a gyroscope, visual odometry also measures the rotation. Integrating the gyroscope over the interval between the capturing of subsequent video frames yields the same measurement as the rotation calculated by the visual odometry using these frames. Thus, while the visual odometry does not provide any new information, it can statistically improve the orientation estimation and can also help to detect or handle irregular measurements, e.g. in the case of magnetic distortion. This fusion scenario with the involved measurements is sketched in figure 6.

The *stochastic cloning approach* proposed in [19] allows us to consider incremental measurements which relate a state at time t_k with a state at time t_{k+n}. For that purpose, the state is augmented at time t_k with a clone of it. This clone will remain an estimate for the time t_k so that the augmented state at time t_{k+n} contains an estimate for the time t_{k+n} as well as for the time t_k, which can be used for correction with the incremental measurement. Accordingly, stochastic cloning is suitable to fuse the orientation information given by the gyroscope and visual odometry in a statistical consistent way.

Every time, say t_k, when a new pair of video frames is captured the state is augmented by a clone of the orientation quaternion. Then, at time $t_{n+k} = t_k + \Delta t_{Video} = t_k + n\Delta t_{IMU}$ the next pair of video frames is captured and visual odometry determines the rotation based on the pair of subsequent camera images. Obviously, the incremental rotation calculated by visual odometry is a measure of the difference between the orientation at time k and $k+n$ 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 according to the translation provided by visual odometry every n IMU samples. As the translation $\mathbf{d}_{k+n}$ calculated by visual odometry is given in the local frame it has to be rotated into the global frame before it can be incremented:

$$\mathbf{t}_{k+n} = \mathbf{t}_k + (q_{k,k+n}^{-1} \otimes \mathbf{d}_{k+n} \otimes q_{k,k+n}) \qquad (7)$$

where $\otimes$ denotes quaternion multiplication. The current egopose (as used in Sec. 3) is expressed as affine transformation

$$T = \begin{bmatrix} R(q_{k+n}) & \mathbf{t}_{k+n} \\ 0^{\mathrm{T}} & 1 \end{bmatrix} \qquad (8)$$

with $R(q)$ the left-handed rotation matrix equivalent to the rotation quaternion q.

Fig. 7 provides the estimated trajectory for a walk of approximately 500 m.

5. Acoustic Feedback

The acoustical feedback generated by the system shall offer an enhanced perception of the surrounding environment. This can have a warning as well as an informing function. On the basis of the environment model detailed in Section 3 and the ego-pose estimation in Section 4, the distances and directions of objects and obstacles around the user can constantly 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 a headphone. 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 [4]. An acoustic image of the environment arises which the user can freely interact with by turning or moving the head. The acoustic representation of the environment can carry lots of information. Several aspects need to be considered in order to keep the cognitive load of interpreting the feedback small and the system intuitive to use. The most crucial is the selection of appropriate sounds, which we treated in a row of surveys and simulator studies with visually impaired as well as sighted persons.

To avoid confusion, the system sounds need to be clearly distinguishable from natural environmental sounds. Additionally, they need to be pleasant to listen to and transport semantic information about the obstacle, e.g. its kind, its motion or its potential danger in an intuitive way. To keep the cognitive load small and minimize required training efforts, the number of different sounds has to be limited. We conducted a study with 26 visually impaired persons to find an appropriate categorization based on a set of 40 different obstacles. The best fitting categorization consisted of (a) wide objects (e.g. ticket machines, cars, benches), (b) pole-like objects, (c) elevated objects (awnings, barriers) and (d) approaching dynamic objects. Additional desirable categories were drop-offs and holes, high curbs and stairs, and crosswalks. Objects of these categories are currently not modelled in the vision framework.

The technical aspect of locatability plays an important aspect. Localising 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 localisation. Sounds exhibiting wide frequency spectra often conflict with the requirement of comfort. Especially high frequencies can only be used carefully. The sounds selected should furthermore be in some kind of harmony with each other since usually multiple sounds will be rendered simultaneously. To find appropriate sounds, we carried out an experimental study with 30 persons (15 of which visually impaired) in a sound simulator. Using a headphone we played 18 synthesized sounds from 20 different directions distributed in a 140° field in front of the head. The participants pointed a marker towards the perceived sound source which we used to automatically measure the localization error. Furthermore, we asked to assign the sounds to the previously defined categories and asked a grade to judge the comfort. On the horizontal plane, experimental studies with real sounds reveal angular localization accuracies of around 10°.The localization errors in our virtual sound experiments were about twice as high. However, in reality we are able to turn the head towards a sound source, which strongly increases the localization accuracy but is not reflected in our experiments.

A final important step is the selection of relevant obstacles in the current situation. In urban environments there are typically many more objects in the vicinity of a person than the number of sounds that are distinguishable simultaneously. To keep the acoustic feedback intuitive we select the three most relevant objects in terms of distance and deviation from the current walking direction. Sound sources are virtually placed at their positions. Using the current egopose (Sec. 4) the sound locations are transformed into local head coordinates, convolved with the HRTF of the users and their amplitudes adapted to the distance.

6. Experiments

The experimental setup consists of two Flea2 cameras with a baseline of around 18 cm and wide-angle lenses of 3.5 mm focal length mounted on a helmet (Figure 2). The IMU (Xsens MTi-300) is flush-mounted into the helmet on top of the person's head. The cameras are calibrated to each other and to the IMU. The headphones are Sennheiser PX 100-II and mounted to the helmet sharing the IMU coordinate frame. Thereby we avoid the required extrinsic calibration between headphone and camera frame. The computing platform is an i7 2.4GHz dual-core notebook carried in a backpack.

6.1. Field Test

Proving the system concept and assessing its value and usefulness for visually impaired persons required testing the system under realistic conditions. Since the behaviour of the user is influenced by the system output, it was important to test the whole control loop containing the perception algorithms, the acoustic feedback and the user behaviour.

The developed prototype was put to a field test with 8 visually impaired persons at the age of 20 to 50 years. Five of the participants are independently mobile, the remaining

Figure 8: Parcours used in the fieldtest.

three persons are more restricted in their mobility and rely on the support of others.

As a first part of the test we set up a training scenario consisting of two big obstacles on an open field. The participants were asked to navigate towards the obstacles and pass between them to validate the sound localization concept. The feedback principle was immediately clear to all the participants. The scenario is well suited to experience the concept of spatial sound and allows simple interaction by approaching the object or passing it, which causes the sound to move correspondingly around the user. After around 10 minutes we moved on to a more complex 2nd parcours.

The parcours consisted of a more complex scenario with the purpose to find out whether the participants 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. Along this path we placed different obstacles (low boxes, high poles and one obstacle hanging overhead) with a few meters distance, some directly on the path, some to the left and right (see Figure 8). The system classified these obstacles into flat, pole-like, dynamic and overhead obstacles, each with a distinct sound. In the beginning the participants tended to shortly stop walking whenever a new obstacle was sonified and turn their head in order to confirm the sound direction. Later, new obstacle sounds caused them to decrease their walking speed until the white cane touched the obstacle. It turned out that the distance to objects was difficult to assess based on the sound intensity alone. The training period 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 parcours some probands had developed a sense that allowed them to avoid obstacles before they could touch them with the white cane. The biggest reported difficulty was to assess the object extension, since it was not reflected in the feedback.

Most participants were sceptical about the principle of artificial spatial sounds after participating in our simulator studies. The experience with the system under real conditions turned out more positive than expected for these users. The acoustic overlay did not cause them to feel limited in their natural sense of hearing. The concept of informing the person about the environment rather than generating navigation clues was received positive. Devolving the decision making to the assistance system is a high hurdle for most visually impaired persons, they like to stay in control. All of our participants could imagine to apply such system for assistance.

7. Conclusion

With the aim of improving the individual mobility of visually impaired persons we have developed a wearable, camera based aid. In this work we developed our conceptual idea of combining scene perception on object level with spatial acoustic feedback to overcome the 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 estimation of the geometric scene background and the detection and tracking of generic static and dynamic objects within the scene foreground. This compact abstract representation serves as a base for the acoustic feedback. A robust technique for head-tracking was developed which combines an inertial measurement unit with camera based visual odometry to allow high frequent measurements of head position and orientation with small delay.

In a set of surveys and simulator studies we adapted the feedback concept to the wishes of the visually impaired users. We selected sounds which allow good obstacle localization and intuitive interpretation of the virtual acoustic world. The developed concept notifies and warns about potential dangers, but the user stays in control how to use this information, which increases acceptance of such systems.

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.

Acknowledgement

This work was supported by the German Federal Ministry of Education and Research within the research project OIWOB. The authors would like to thank the ”Karlsruhe School of Optics and Photonics” for supporting this work.

References

[1] H. Badino, U. Franke, C. Rabe, and S. Gehrig. Stereo Vision-Based Detection of Moving Objects under Strong Camera Motion. In *Proceedings of the First International Conference on Computer Vision Theory and Applications*, volume 2, pages 253–260, February 2006. 4

[2] Y. Bar-Shalom. *Tracking and Data Association*. Academic Press Professional, Inc., San Diego, CA, USA, 1987. 4

[3] J. Benjamin and J. Malvern. The new C-5 laser cane for the blind. In *Carnahan Conference on Electronic Prosthetics*, pages 77–82, 1973. 2

[4] J. Blauert. *Spatial hearing : the psychophysics of human sound localization*. Cambridge, Mass. MIT Press, 1997. 7

[5] A. Colombari, A. Fusiello, and V. Murino. Segmentation and tracking of multiple video objects. *Pattern Recognition*, 40(4):1307 – 1317, 2007. 4

[6] D. Dakopoulos and N. Bourbakis. Wearable obstacle avoidance electronic travel aids for blind: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(1):25–35, Jan 2010. 1

[7] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Moving obstacle detection in highly dynamic scenes. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 56–63, May 2009. 4

[8] G. P. Fajarnes, L. Dunai, V. S. Praderas, and I. Dunai. CASBLiP-a new cognitive object detection and orientation system for impaired people. *trials*, 1(2):3, 2010. 1

[9] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, pages 963–968, 2011. 5, 6

[10] J. Gonzalez-Mora, A. Rodriguez-Hernandez, E. Burunat, F. Martin, and M. Castellano. Seeing the world by hearing: Virtual acoustic space (vas) a new space perception system for blind people. In *ICTTA '06. Information and Communication Technologies*, volume 1, pages 837–842, 2006. 1

[11] T. Hermann, A. Hunt, and J. G. Neuhoff, editors. *The Sonification Handbook*. Logos Publishing House, 2011. 2

[12] L. Kay. An ultrasonic sensing probe as a mobility aid for the blind. *Ultrasonics*, 2(2):53–59, 1964. 1

[13] E. Kraft. A quaternion-based unscented kalman filter for orientation tracking. In *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, volume 1, pages 47–54, July 2003. 6

[14] W. Mayol-Cuevas, B. Tordoff, and D. Murray. On the choice and placement of wearable vision sensors. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 39(2):414–425, March 2009. 2

[15] P. B. Meijer. The voice. http://www.seeingwithsound.com, (accessed September 2015). 1

[16] V. Pradeep, G. Medioni, and J. Weiland. Robot vision for the visually impaired. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 15–22, June 2010. 1

[17] A. Rodríguez, J. J. Yebes, P. F. Alcantarilla, L. M. Bergasa, J. Almazán, and A. Cela. Assisting the visually impaired: Obstacle detection and warning system by acoustic feedback. *Sensors*, 12(12):17476, 2012. 1

[18] M. Romanovas, T. Schwarze, M. Schwaab, M. Traechtler, and Y. Manoli. Stochastic cloning Kalman filter for visual odometry and inertial/magnetic data fusion. In *16th International Conference on Information Fusion*, pages 1434–1441, 2013. 6

[19] S. Roumeliotis and J. Burdick. Stochastic cloning: a generalized framework for processing relative state measurements. In *ICRA '02 IEEE International Conference on Robotics and Automation*, volume 2, pages 1788–1795, 2002. 6

[20] L. Russel. Travel path sounder. *Rotterdam Mobility Research Conference*, 1965. 1

[21] T. Schwarze and M. Lauer. Geometry estimation of urban street canyons using stereo vision from egocentric view. In *Informatics in Control, Automation and Robotics*, volume 325 of *Lecture Notes in Electrical Engineering*, pages 279–292. Springer International Publishing, 2015. 3

[22] T. Schwarze and M. Lauer. Robust ground plane tracking in cluttered environments from egocentric stereo vision. In *2015 IEEE International Conference on Robotics and Automation*, pages 2442–2447, May 2015. 3

[23] S. Shoval, I. Ulrich, and J. Borenstein. Navbelt and the guide-cane. *IEEE Robotics & Automation Magazine*, 10(1):9–20, 2003. 1

[24] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey, 2006. 4