Journal on Multimodal User Interfaces (2019) 13:219–230
https://doi.org/10.1007/s12193-018-0281-3

ORIGINAL PAPER

Interactive sonification of U-depth images in a navigation aid for the visually impaired

Piotr Skulimowski[1] · Mateusz Owczarek[1] · Andrzej Radecki[2] · Michal Bujacz[1] · Dariusz Rzeszotarski[1] · Pawel Strumillo[1]

Received: 10 January 2018 / Accepted: 30 October 2018 / Published online: 8 November 2018

Abstract
In this paper we propose an electronic travel aid system for the visually impaired that utilizes interactive sonification of U-depth maps of the environment. The system is comprised of a depth sensor connected to a mobile device and a dedicated application for segmenting depth images and converting them into sounds in real time. An important feature of the system is that the user can interactively select the 3D scene region for sonification by simple touch gestures on the mobile device screen. The sonification scheme is using stereo panning for azimuth angle localization of scene objects, loudness for their size and frequency for distance encoding. Such a sonic representation of 3D scenes allows the user to identify the geometric structure of the environment and determine the distances to potential obstacles. The prototype application was tested by three visually impaired users who managed to successfully perform indoor mobility tasks. The system's usefulness was evaluated quantitatively by means of system usability and task-related questionnaires.

Keywords Interactive sonification · Image sonification · Visually impaired · Electronic travel aid · Depth maps · U-depth · U-disparity

1 Introduction

The visually impaired indicate limited mobility as one of the main problems affecting almost all activities of daily living. The research efforts aimed at building Electronic Travel Aids (ETA) date back to the nineteenth century, when in 1897 Polish ophthalmologist Kazimierz Noiszewski constructed Elektroftalm: a device named "electronic eye" that converted light into sounds or vibrations by using the photoelectric properties of Selenium cells. Although too heavy for practical application, it is considered to be the first electronic sonification interface for the visually impaired [1]. Further attempts were pioneered by Bach-y-Rita [2], who built a number of ETA prototypes that used tactile modality as a channel of communication with the blind. Dynamic development of Information and Communications Technologies (ICT) at the turn of centuries (100 years after seminal efforts by Noiszewski) marked a new chapter in the efforts to design personal aids helping blind people in mobility (laser and ultrasound detectors) and navigation (GPS).

With regard of the non-visual methods used for presentation of information these devices can be subdivided into haptic interfaces and auditory interfaces. An excellent

This work was partially supported by the National Science Centre of Poland under Grant No 2015/17/B/ST7/03884 in years 2016–2018.

✉ Piotr Skulimowski
piotr.skulimowski@p.lodz.pl

Mateusz Owczarek
mateusz.owczarek@p.lodz.pl

Andrzej Radecki
andrzej.radecki@p.lodz.pl

Michal Bujacz
michal.bujacz@p.lodz.pl

Dariusz Rzeszotarski
dariusz.rzeszotarski@p.lodz.pl

Pawel Strumillo
pawel.strumillo@p.lodz.pl

[1] Institute of Electronics, Lodz University of Technology, 211/215 Wolczanska Str., 90-924 Lodz, Poland

[2] Institute of Automatic Control, Lodz University of Technology, 18/22 B. Stefanowskiego Str., 90-924 Lodz, Poland

review of wearable obstacle avoidance ETAs is given in [3]. Due to size factor and cost, in the majority of ETAs, auditory displays are favoured over haptic interfaces that would require complex circuitry to control mechanical stimulations [4]. There are many possible auditory representations of information which can be employed in human–machine interfaces (HMI were widely reviewed in [5]). However, it is sonification, i.e., non-speech audio, which is the method predominantly used for "displaying" the environment to the visually impaired. Quite a comprehensive review of the sonification methods devised for aiding the blind in mobility and travel is given in [6]. Worth mentioning here is the vOICe [7], a widely popularized method for sonifying monochrome images. The employed sonification method, however, is simplistic, not intuitive and requires many weeks of training. In that approach, the vertical coordinate of every pixel corresponds to a specific pure-tone frequency in the range of 500 Hz (bottom image pixels) to 5 kHz (top image pixels), whereas, loudness of the frequency is reflecting the local brightness of the image. This sonification code is used in a looped, one second long, auditory representation of the image that is scanned from left to right. Such a sonification scheme is non-interactive and difficult for the user to control.

In the past decade an important subfield of sonification has emerged, namely: interactive sonification [8]. In such an approach to human–computer auditory interface, the user has been enabled to interact with the sonification process, e.g. the user can define an image region to be sonified or tune the sonification parameters to individual requirements in real time. This is a very important feature of the interface for the blind users, since they can control the speed and amount of auditory information generated by the interface. Thus, the problem of information capacity mismatch between the visual channel and the auditory channel can be alleviated. Large volumes of multidimensional visual data (2D or 3D) that are normally not accessible to the blind user can be converted into one dimensional acoustic signal and encoded using such attributes as loudness, timbre, fundamental frequency (pitch) and the signal envelope. The advantages of interactive sonification techniques in assistive devices for the visually impaired were reported in a number of studies. In [9] the user needed to use a mouse and a keyboard to interactively sonify image edges. Another approach to interactive sonification was adopted in [10]. In this study haptic line graphs made of rubber-bands were explored by touch by the blind user and sonified. Finally, in [11], only simple image primitive shapes (line segments, curved edges, colour) were sonified while intensive image preprocessing methods were applied to recognise scene objects and verbally describe them to the user. Interactive sonification has currently evolved from auditory display methodologies into a mature research study field. From 2004 an Interactive Sonification Workshop (http://interactive-sonification.org) has been organised biannually and attracted researches applying interactive sonification techniques in various disciplines ranging from science, industry and sports to medicine and assistive technologies for the blind.

In this paper we demonstrate how the technique of interactive sonification, which is is capable of representing spatial 3D geometry of the environment, can be applied in a simple travel aid for the blind. We use the so-called depth images and their histograms termed "U-depth" which are simpler for auditory representation to the blind user.

2 3D scene reconstruction basics

2.1 Structured-light 3D scanner

The 3D geometry of the environment can be reconstructed either by using passive techniques like a stereovision camera or active techniques in which light or radio signals are emitted into the environment and the reflected signals are recorded by an appropriate sensor. Active depth sensors are usually based on infrared structured light or Time-of-Flight (ToF) technology [12,13]. An example of ToF camera is Kinect 2 [14], unlike the earlier version of the Kinect [14] and the Structure Sensor [15] which uses infrared structured light pattern projector and a low range infrared CMOS camera. The main constraint of active infrared depth sensors is that they operate reliably only in indoor environments. In the other contexts, e.g. in the direct sunlight the CMOS camera can be "blinded" and the device will not be able to recognize infrared pattern reflected from the surrounding objects. However, for indoor environments, the active depth reconstruction techniques outperform stereovision in terms of reconstruction reliability, depth accuracy and generated image frame rates [16].

For the ETA system presented in this article, the Structure Sensor device was used. The Structure Sensor is a lightweight (95 g), compact (119 × 28 × 29 mm) and features the depth reconstruction range of 40–500 cm. The 3D geometry of the environment is given in the form of the-so called depth map being a 2D array in which each element is a value representing the distance of a 3D scene point to the sensor. The depth map is calculated with a rate of 30 frames per second (fps) with a spatial resolution of 640 × 480 points with depth accuracy between 0.12 and 1% depending on the distance (the larger the distance the lower the accuracy).

The Structure Sensor device is mounted on a user's head with the help of a dedicated headgear made of elastic straps (see Fig. 2). The field of view (FoV) of the sensor is 58° on the horizontal plane and 45° on the vertical plane. This FoV is rather narrow if compared to the human sight, however, the blind participants commented that such a selective probing of the environment helps them to limit the amount of the sonified information and that by using head turns they can

scan the environment with a sufficient spatial width. Henceforth, the Structure Sensor device will be reffered to as the depth camera.

2.2 "U-depth" representation

Binocular disparity occurs in binocular vision systems and is defined as the difference in horizontal coordinates (termed also a parallax) of any point in 3D environment projected onto two images of the stereovision system. The closer the point of the environment to the stereo camera, the larger is its disparity, by the same token, in the case of remote objects the disparity converges to zero. By computing disparities for all scene points within the field of view of the stereovision system one can obtain the so called dense disparity map. From such a map, the depth map representing the 3D geometrical structure of the observed scene can be directly calculated (see Fig. 1) [16].

Literature shows that the U-disparity representation of the environment can prove very effective in obstacle detection for automotive and autonomous robots applications [17–19]. The U-disparity representation is built by computing histograms of consecutive columns of the disparity map. Let us assume the disparity map pixel resolution to be $W \times H$ (i.e., width × height). The size of the U-disparity map is $W \times d_{max}$, where d_{max} is the maximum allowed disparity value. Thus, the value of each point $u(x, d)$ in the U-disparity map is the number of scene points at x-coordinate assuming disparity d. The U-disparity and complementary to it the V-disparity maps were proposed and applied in scene depth analysis tasks in [19–21]. The U-disparity map appears to be a very efficient representation for localizing scene objects (provided the stereovision camera base is parallel to the ground plane) [22]. An object positioned at a well localized distance features a region of the same value in the disparity map, which results in a unimodal histogram with a strong maximum in the U-disparity map (see Fig. 1 showing a 3D scene example and the corresponding depth maps).

The depth camera automatically generates depth maps of the scene instead of the disparity map which comes from the stereovision camera. Thus, in our system we directly use the U-depth maps for detecting scene objects. For the considered application of depth imaging in an electronic travel aid for the blind we assumed the maximum depth value to be $z = 5$ m, i.e. distant objects ($z > 5$ m) are discarded and not sonified. The size of the resulting map is thus $W \times N$, where $W = 320$ is the width of the depth map and $N = 10$ is the number of depth ranges which is equal to the number of sound frequencies used for interactive sonification of the the U-depth map. We have decided that the maximum depth range should be a user-defined parameter, i.e. the user can limit this range to any distance smaller than 5 m.

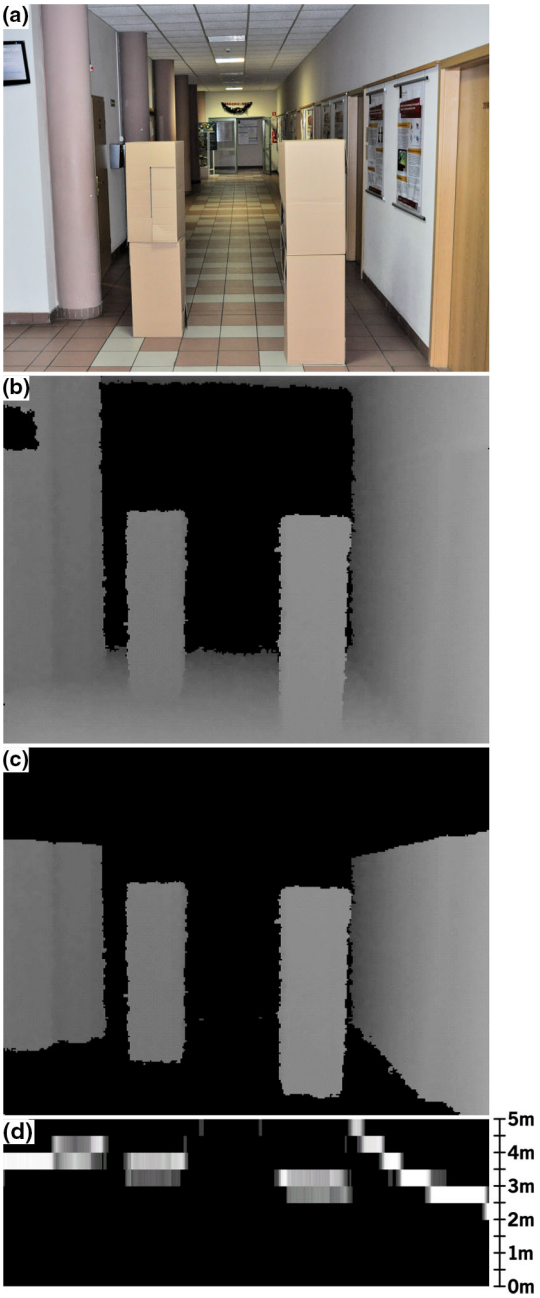

Fig. 1 An example 3D indoor scene (**a**), its depth map (**b**), the depth map with removed ground surface (**c**), the U-depth map computed from the depth map (**d**)

The main application of the system is to aid the visually impaired in navigating in the environment. Thus we hypothesize that recognition of depth position changes of distant objects is considerably important. This hypothesis has been observed during first trials of the system by noting a frequent use of the verbal mode distance detector by the testers. Thus, a linear mapping of depth to frequency scale has been adopted to favour better recognition of variations of distances for distant objects (which are coded by lower frequencies).

For signalling close obstacles the proximity mode by default is automatically activated (see Sect. 3.3). In this mode an alarming sound is generated to warn the user about close obstacles.

2.3 Ground plane estimation

There is a major challenge inherent in reconstructing the geometry of the 3D environment. The ground is visualised by the depth camera as a surface-like object of gradually changing depth. For the purpose of the ETA device such an object should not be sonified as this is the scene region that does not contain any obstacles. Hence in our approach we have developed an algorithm for estimating the location and orientation of the ground surface which is then removed from the further processing pipeline leading to sonification of the U-depth map.

Ground plane estimation consists of the following processing steps. First, the distance between the ground plane and the origin of the depth camera coordinate system is calculated. This value can be identified as the user's height. Moreover, the orientation of the depth camera relative to the ground is estimated as this depends on how the depth camera has been placed on the person's head. In the initialization mode the user is asked to stand in a natural upright position, for which the ground plane region occupies a significant part of the image, e.g. an empty hall or a corridor. For each node (x_i, y_i) of a square grid of size 16pix within the depth map and for two additional points $(x_i - 3, y_i)$ and $(x_i, y_i - 3)$ of the depth map, the corresponding 3D coordinates are calculated. Based on the values of these coordinates the ground plane equation is calculated. If the distance between the ground plane and the origin of the depth camera coordinate system is within a predefined range (150 ± 50 cm), and the angle between the normal vector to the ground plane and the expected normal vector [0.0, 1.0, 0.4] is within a predefined range (0–25°), these points are further used to estimate the ground plane equation using the least square method. Once the ground plane equation $Ax + By + Cz + D = 0$ is calculated for the initial position of the depth camera, it is possible to calculate the user's height h_u and the relative orientation of the depth camera to the ground plane. Similar method is used to track the ground plane equation for consecutive images of the the 3D scene. In this mode, the camera distance boundaries assumptions are limited to a range of $h_u \pm 30$ cm. If the ground plane does not fit the angle and camera distance range, the ground plane equation from the previous frame is used. Such a solution prevents false ground plane detections, e.g. in cases when a user is close to the wall or there are numerous obstacles in the scene occluding the ground.

2.4 Depth map preprocessing

Once the point coordinates in the depth camera coordinate system have been defined, it is possible to remove those whose distance from the identified ground plane is larger than the established threshold. Subsequently, further processing steps of the depth map are made and the following components are removed:

- regions of the ground plane,
- regions in the background ($z > 5$ m),
- regions for which the distance from the ground plane is larger than the pre-selected value (i.e. higher than user's height).

An example of the U-depth map for the indoor test scene is shown in the Fig. 1d. Note that key obstacles, the cardboard boxes and walls are clearly highlighted in the U-depth map. The U-depth map can be interpreted as a top view at the scene.

3 The system and its multimodal interface

The system hardware consists of the Structure Sensor depth camera, a smartphone with Android OS and a pair of open in-ear headphones. The system set up (mounted on a mannequin head) is shown in Fig. 2. The depth map preprocessing procedures (ground plane estimation and removal) and the U-depth map are calculated on the Android-based platform. The depth camera delivers depth images at 320×240 resolution. Images are processed at a rate of approx. 25 frames per sec-

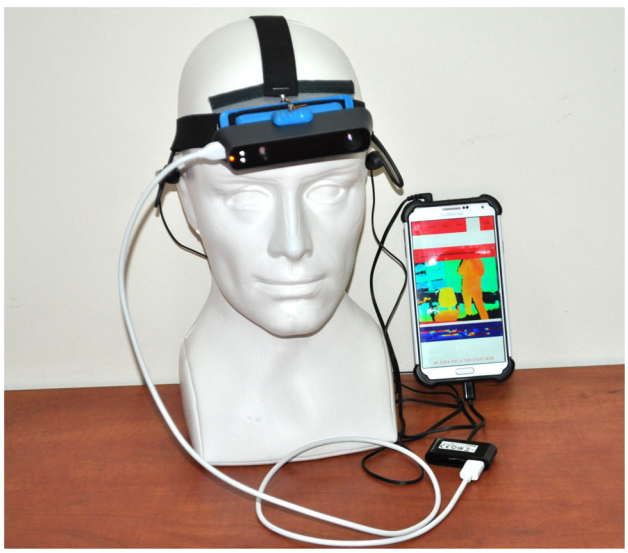

Fig. 2 Hardware components of the electronic system for interactive sonification of 3D scenes

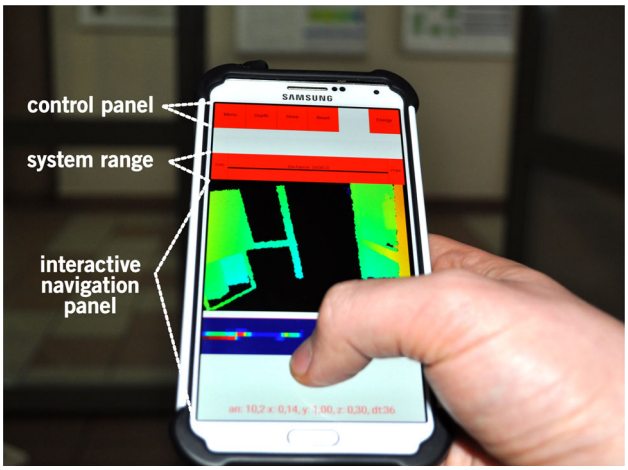

Fig. 3 Touch screen display of the smartphone with the running application for interactive sonification of 3D scenes: the top panel is the control panel, directly below is the panel for setting system range sensitivity, then the depth map is displayed in pseudo-colours and the corresponding U-depth map is displayed below it

ond. The mobile phone can be hidden in the user's pocket. The sonification output stream comes from the mobile phone through the stereo headphones.

3.1 Multimodal user interface

The screen of the Android mobile phone is the touch user interface of the system assisting the visually impaired in navigation. It is divided into three panels (see Fig. 3). The control panel displayed on the top of the screen is not dedicated to the user. It contains control buttons for recording image sequences, disabling screen display or calculating and storing data about the user's height and camera orientation versus the ground plane. The second panel (below the control panel) enables the user to set the maximum depth range value ($z < 5$ m) within which the obstacles will be presented to the user. Phone vibrations inform the user that this panel is selected and depth range modified. When the user sets the distance, the selected value is announced verbally using a text to speech module. The third and the largest panel displays the depth map and its U-depth representation. The user can explore the maps by touch and select 3D scene regions for sonification. We call this panel the interactive navigation panel.

The system works in three operating modes: the interactive sonification mode, the proximity sonification mode and the verbal mode. The interactive sonification mode is activated by touching and holding a finger in the interactive navigation panel. The proximity mode is the default operating mode which is always active while the user in not touching the screen. Finally, the verbal mode is activated on the phone's touch screen by a vertical "fling" gesture (i.e., an upward swipe of the user's finger) at a point of interest on the U-depth map. A more detailed description of these operating modes is provided in the subsequent sections.

3.2 Interactive sonification mode

In the sonification interactive mode, the blind user can select a scene area for sonification by touching the mobile phone screen in the interactive navigation panel. Only the x coordinate that is selected by the user is significant and determines which columns of the U-depth map are presented to the user.

Let $x \in \{0, 1, \ldots, W-1\}$ denotes a column of the map indicated by the user ($W = 320$). Touching the centre of the map triggers retrieval of information about the obstacles directly in front of the user. The sound sonifying the scene depends on the content of the U-depth map. The indicated x coordinate of the map controls left–right panning of the generated sound:

$$v_L = 1 - x/W \qquad v_R = x/W \tag{1}$$

where v_L, v_R are volumes of the output's left and right channels and W is the number of columns of the U-depth map.

It is worth noting that such a method of sonifying the obstacle horizontal position (left/right panning) is intentionally simplified for the user and it is related to the depth values rather than world coordinates.

The y-coordinate location of a 3D scene object in the U-depth map (i.e. the row number) determines the sound frequency that corresponds to the depth information (the higher the pitch the closer the sonified object). The sound signal generated by the system is a packet of sinusoids:

$$s(t) = \sum_{i=0}^{N-1} a_i \sin(2\pi f_i t) w_i(t) \tag{2}$$

where

$$f_i = f_{max} - i\frac{f_{max} - f_{min}}{N-1} \tag{3}$$

$$w_i(t) = \begin{cases} 1, & \text{for } iT \le t < (i+1)T \quad (4a) \\ 0, & \text{otherwise.} \quad (4b) \end{cases}$$

Each sinusoid frequency represents the selected distance range (see Fig. 1d). We define $N = 10$ as the number of different sound frequencies, $f_{max} = 4000$ Hz is the frequency of sound with index 0, which corresponds to the closest objects, and $f_{min} = 400$ Hz is a frequency of sound indexed as $N-1$. Time interval $T = 5$ ms is the sound duration for each distance range. The whole sonification cycle,

therefore, lasts $T \cdot N = 50$ ms. The f_{min} and f_{max} values were selected both to address technical limitation of the selected headphones and are based on testers' preferences about the subjective pleasantness of the selected frequencies. Assigning low frequency sounds to distant objects has its justification. These range of frequencies are better distinguishable by the human ear. This feature is useful in recognizing orientation of large objects with respect to the user e.g. corridor walls or building walls.

Then the amplitude of each sinusoid is calculated as:

$$a_{ie} = \frac{\sum_{j=x_{min}}^{x_{max}} u(j,i)}{H(x_{max} - x_{min} + 1)} \tag{5}$$

where $x_{min} = max\,(0, x - m)$, $x_{max} = min\,(x + m, W - 1)$, H is the number of rows of the depth map, $m = 2$ is the number of columns to the left and right of the selected column (its nearest neighbourhood) that is being sonified. As is apparent, the sound amplitude for a smaller object is smaller than for a larger object. The values for parameter a_{ie} range from 0 to 1, where 1 corresponds to a situation, for which depth values in the range of $\langle x_{min}, x_{max} \rangle$ columns correspond to the same U-depth map values, i.e. it can be a flat wall in front of the user and the ground plane is not visible. To accentuate small size obstacles, which may pose serious danger to the blind users, the amplitudes of the sinusoids defined in Eq. (5) are non-linearly transformed by using the following formula:

$$a_i = a_{ie}\left(1 + C_1 e^{-C_2 a_{ie}^2}\right) \tag{6}$$

where C_1 and C_2 are constants empirically set to $C_1 = 1.2$ and $C_2 = 8.0$. Equation (6) was inspired by the normal (Gaussian) distribution.

Plots of a_i and $g = \frac{a_i}{a_{ie}}$ factors are shown in Fig. 4.

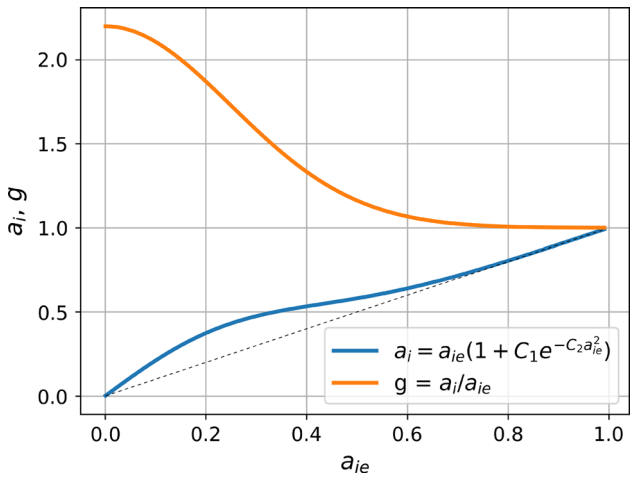

Fig. 4 Plots of a_i and g factors for the parameters used in amplitude transformation defined in Eq. (6)

3.3 Proximity mode

If the user does not touch the screen, by default, the proximity mode is automatically activated. Proximity mode warns the user about the closest obstacles only. For this purpose, merely first 3 out of 10 depth ranges are sonified. The U-depth map is divided into 11 vertical segments. Each segment of columns is jointly analysed in order to detect the number of pixels which are located in the "proximity area" ($z < 1.5$ m). The sonification scheme is identical to the one used in the interactive sonification mode. Each of these 11 parts is played as a stereo sound, the panning of which conveys information about the azimuthal position of the obstacle. However, for the proximity mode only f_{max} frequency is used and the sound volume depends on the number of pixels in each section of the U-depth map. The sound for each part is played with a duration of 3 ms followed by a 3 ms pause. The proximity sonification cycle ends with a 30 ms of silence, so the entire cycle lasts $11 \cdot (3 + 3) + 30 = 96$ ms.

3.4 Verbal mode: distance detector

In the interactive sonification mode all the user's touch gestures are continuously analysed and the U-depth map is sonified. The verbal mode, on the other hand, is a special functionality introduced to the system after consultation with the blind users. If the user performs a vertical fling gesture starting from screen coordinate (x, y), i.e. in x column of the U-depth map the closest obstacle along the fling line is detected. Then the depth value (z coordinate) in centimetres to this obstacle is verbally communicated to the user. Note, that the user does not have to take his finger off the screen to hear the distance to the obstacle. If the user does not move his finger for more than 300 ms, there follows a notification in the form of a short vibration and the fling gesture can be repeated again.

4 Mobility tests of the system: results and discussion

The preliminary trials of the proposed electronic travel aid were the first proof of concept trials with the participation of the authors [23]. A tablet instead of a phone was used to better present the processing and analysis procedures of the depth map. In this trial the sonification sounds were played from a portable wireless speaker.

4.1 Test participants and test plan

The main tests of the system were carried out with the participation of three visually impaired volunteers in an indoor environment. These were a woman aged 48 (further refer-

Fig. 5 Photographs illustrating the test cases: introductory session acquainting Tester 1 with the system (**a**), Tester3 during a walk along the corridor (Test 1) (**b**), Tester 2 performing Test 2, i.e. the walk along the corridor with cardboard box obstacles (**c**), Tester 3 during Test 3, i.e. locating and walking through an open space between cardboard boxes (**d**)

enced as Tester 1), a woman aged 32 (Tester 2) and a man aged 35 (Tester 3). Tester 1 belongs to visual impairment category 6 as defined by the World Health Organization [24], meaning she is totally blind with no light perception. Whereas Tester 2 belongs to category 4 and Tester 3 belongs to category 3. Tester 1 lost her sight at age of 24 and her primary mobility aid is white cane, whereas both Tester 2 and Tester 3 are aided by guide dogs.

All testers are familiar with mobile phones with touch screens. The tests were approved by the Bioethics Commision at Medical University of Lodz, Poland. Mobility tests with the blind users were preceded by an explanation of how the system works and how to operate it. The testers were given time to experiment with the mobile phone and the application. See Fig. 5a showing Tester 1's first hand-on experience of the system. The testers were instructed as how the U-depth map is generated and how to interpret it and were acquainted with the sonification method of the U-depth map generation. This introduction was also carried out using the test image sequences pre-recorded by the system.

During all tests, the testers used only the interactive sonified U-depth mobile navigation system and did not use other assistive aids (e.g., a cane or a guide dog). Below are the three mobility tests carried out to test system usability in real life navigation scenarios:

1. A walk along an empty corridor—to test the user's capability to maintain a straight walking direction alongside walls.
2. A walk along a corridor with obstacles simulated by cardboard boxes—to test the user's efficiency in detecting and avoiding obstacles.
3. Locating an open space between walls of approx. 90 cm width and walking through it—to test how skillfully the users located open doors and walked through them.

The tests were videotaped and, additionally, the mobile application was logging which system operating modes were being used. Also, the completion times of each of the tasks were noted. After each task, the test participants were asked to answer four task related questions. Finally, having finished all the tasks, the participants filled in a system usability questionnaire.

4.2 Walking along an empty corridor: Test 1

The user's task in this test was to walk along a long corridor of 3 m width and overall length of 42 m. In the middle of the path there is a door narrowing the passage to 90 cm. Then, the corridor turns left and then, 3 m farther, it turns right. The testers were instructed to walk along the corridor (see Fig 5b showing Tester 3 performing the test). No details about the corridor topology were disclosed to the testers. During the test, the navigation system was taking logs of user interaction with the application, i.e. which system operating mode was selected by the user (i.e. interactive sonification mode or the proximity mode) and the time stamps of the instances at which the user was activating the verbal distance detector mode.

The data collected from the application logs during the test are presented in Fig. 6 and summarized in Table 4. Figure 6 shows time axes with time intervals of users' interaction activity with the application. Note that Tester 1 was mainly using the interactive sonification mode, i.e. was constantly touching the U-depth panel in search for obstacles (coded grey in Fig. 6). The only longer period during which Tester 1 was using the proximity mode (coded white) was while walking through a narrow passage (see note "door passed" tag on the timeline). Note also, that Tester 1 used the verbal distance detector mode only once. Tester 2, on the other hand, preferred to use the verbal distance detector mode to explore the environment more often. She activated this mode 22 times (see the arrows with indicated distances in Fig. 6). She also frequently switched between the interactive sonification mode and the proximity mode. Finally, Tester 3 has completed the task in the shortest time and walked along the corridor confidently. The other two testers frequently stopped and used the system to scan the corridor. Tester 3 used the verbal distance detector mode while approaching the narrow passage only. Otherwise, he was mainly using the interactive sonification mode with intermittent activation of the proximity mode. All testers have completed this task successfully, however, in considerably different time (see Table 4).

After completion of this task the testers were asked to answer four task related questions. The three questions were

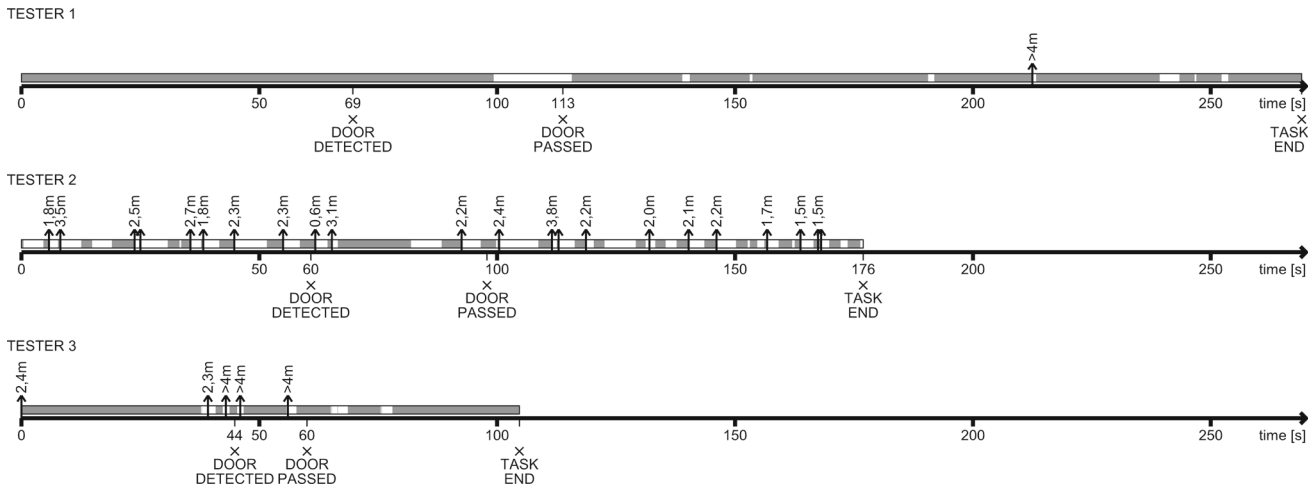

Fig. 6 Timeline of user interactions with the system during Test 1 (walking along an empty corridor), grey—interactive sonification mode, white—proximity mode, arrows indicate activation of the verbal mode distance detector

Table 1 Test 1 related questions (walking along an empty corridor), answers given in the Likert scale: 1—strongly disagree and 5—strongly agree

No	Question	Tester		
		1	2	3
Q1	I do not need much training to complete the task successfully	4	3	5
Q2	I am satisfied with the ease of completing this tasks while using the device	4	4	5
Q3	The system is a helpful tool in solving this type of task	3	4	5
Q4	I am satisfied with the amount of time it took me to complete the task	4	5	5
Average score		3.75	4.00	5.00

based on the questionnaire proposed by Lewis [25]. We have also added one extra question related to the amount of training required to perform each test successfully. Answers (in the Likert scale: 1—strongly disagree and 5—strongly agree) to this task related questionnaire are given in Table 1. Note that the average scores closely correspond to the efficiency with which the test participants completed the task.

4.3 Walking along a corridor with obstacles: Test 2

The scenario for this test was similar to Test 1. The testers' task was to walk along a straight, 18 m long section of the corridor (shorter than in Test 1). In this test, however, there were three obstacles placed along the corridor at random locations (see Fig. 5c). The cardboard boxes sized 40 cm × 40 cm × 160 cm functioned as obstacles. Similarly to Test 1, the application was taking logs of user interaction with the system (see graphical representation of these logs for Test 2 in Fig. 7). Tester 1 similarly as in Test 1 was using mainly the interactive sonification mode and was frequently stopping to scan the environment with the system. However, she also started to use the verbal distance detector mode more frequently. Although, there were two occurrences of box hits in this test, all three testers started to use the system more confidently and completed the task more quickly than in the first test.

In the task related questionnaire (Table 2) the average scores for this task were slightly worse than for Task 1. This is understandable due to increased difficulty of the task. Nevertheless, for Tester 3 the average score neared very good.

4.4 Locating an open space and walking through it: Test 3

The scenario of this test is designed to verify how helpful the system is in a typical mobility task, such as finding an open door and walking through it. For the safety of the testers we have simulated an open door by specially aligned cardboard boxes, as shown in a photo in Fig. 5d. The testers started the test at a 5 m distance from the simulated door and were positioned towards it (but not precisely in the direction of the open space). Their task was to locate the open space between the cardboard boxes and walk through it.

Interestingly, it took all testers quite a long time to complete the task in spite of the short distance to be covered (see again diagrams with timelines of the test shown in Fig. 8). The completion times varied from from $T = 28$ s to $T = 54$ s. The longest time was noted for the tester for whom the sys-

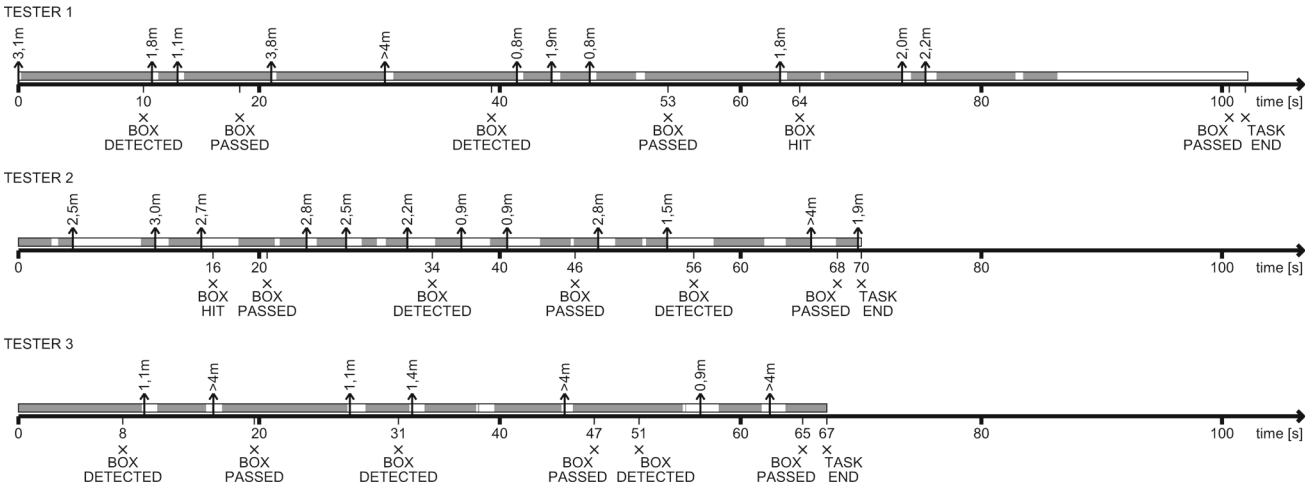

Fig. 7 Timeline of user interactions with the system during Test 2 (walking along a corridor with obstacles), grey—interactive sonification mode, white—proximity mode, arrows indicate activation of the verbal mode distance detector

Table 2 Test 2 related questions (walking along a corridor with obstacles), answers given in the Likert scale: 1—strongly disagree and 5—strongly agree

No	Question	Tester		
		1	2	3
Q1	I do not need much training to complete the task successfully	4	2	5
Q2	I am satisfied with the ease of completing this tasks while using the device	3	4	5
Q3	The system is a helpful tool in solving this type of task	3	3	5
Q4	I am satisfied with the amount of time it took me to complete the task	3	4	4
Average score		3.25	3.50	4.75

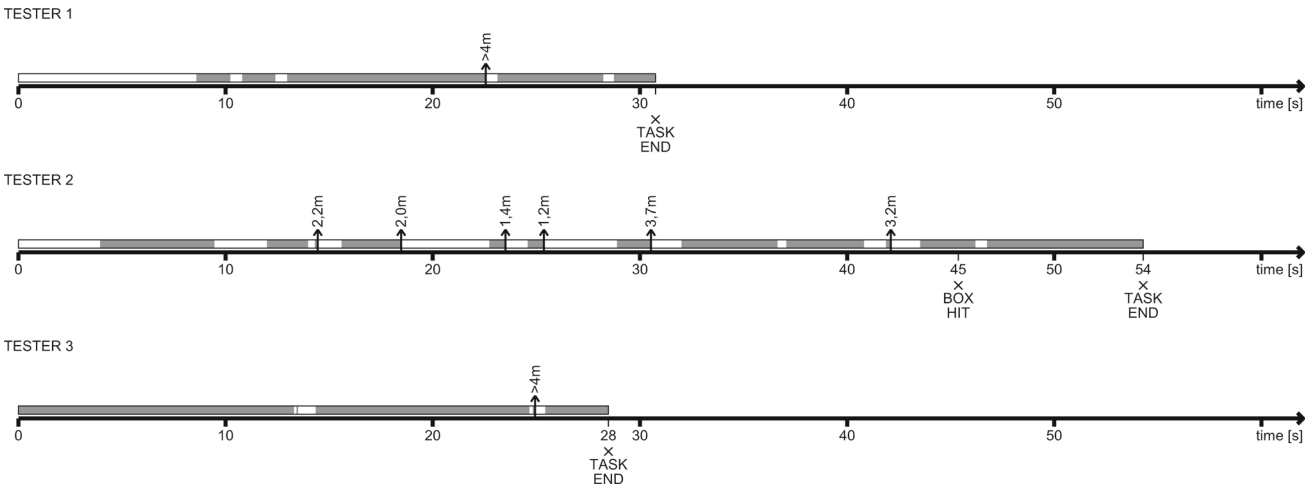

Fig. 8 Timeline of user interaction with the system during Test 3 (door finding), grey—interactive sonification mode, white—proximity mode, arrow—distance detector

tem did not enable him to find the open space quickly and walk through it. After the test, the testers were asked about the long times required to complete the test.

All answers were unanimous and indicated that the test was difficult because, unlike Test 1 and Test 2, the immediate task area (and thus the depth-map) lacked nearby walls. This was the reason given for why it took such a considerable amount of time for the testers to locate the cardboard boxes and a narrow space between the boxes. This observation underlines the importance of objects which can be

Table 3 Test 3 related questions (finding and walking through an open space), answers given in the Likert scale: 1—strongly disagree and 5—strongly agree

No	Question	Tester		
		1	2	3
Q1	I do not need much training to complete the task successfully	2	2	5
Q2	I am satisfied with the ease of completing this tasks while using the device	2	3	5
Q3	The system is a helpful tool in solving this type of task	3	5	5
Q4	I am satisfied with the amount of time it took me to complete the task	1	2	4
Average score		2.00	3.00	4.75

Table 4 Statistics of how the testers performed the mobility tasks while aided by the interactive sonification navigation system

Tester Id	Test time [s]	Modes				
		Interactive		Proximity		Verbal
		Time [s]	Share [%]	Time [s]	Share [%]	Readings [pcs]
Test 1: empty corridor						
Tester 1	269	242	90.14	26	9.86	1
Tester 2	176	80	45.68	96	54.32	22
Tester 3	104	91	87.34	13	12.66	5
Test 2: corridor with boxes						
Tester 1	102	78	76.65	23	23.35	12
Tester 2	70	39	55.69	31	44.31	12
Tester 3	67	55	82.22	11	17.78	7
Test 3: door finding						
Tester 1	30	19	64.90	10	35.10	1
Tester 2	54	32	59.97	21	40.03	6
Tester 3	28	26	94.38	1	5.62	1

detected by the system and serve as markers for the visually impaired. Note that a similarly narrow passage was located by the testers during Test 1 and successfully cleared by the testers in much shorter time. This was because the narrow passage was adjacent to the wall along which the test participants were walking and determining its location presented relatively little difficulty.

The testers were also asked to fill in the task related questionnaire. The answers to the questionnaire (shown in Table 3) indicate the poorest scores in comparison to earlier tests, with the exception of the answers of Tester 3, who ranked this task as not difficult to complete with the aid of the system (in spite of quite a long time actually taken to complete it).

A summary of the timeline diagrams shown in Figs. 6, 7, 8 is given in Table 4. The table contains key information documenting the tests for each tester, i.e. the time required to complete the tests, time proportions between activating interactive sonification and the proximity modes and finally the number of times the users activated the verbal distance detector mode in each task.

Directly after completing all three tests, we conducted another survey which did not focus on particular tasks but on the overall usability features of the tested interactive navigation system for the visually impaired. We used a popular system usability scale (SUS) questionnaire proposed by Brooke [26] to evaluate how useful the system was in helping the blind testers successfully perform the test tasks. The SUS questionnaire consists of 10 questions and uses a five-point Likert-type scale (from 1—strongly disagree to 5—strongly agree) for answers. The questionnaire answers are shown in Table 5. The overall usability score for each participant was calculated according to the procedure proposed by [26] in order to obtain an overall percentage scoring on the 0–100% scale. The users' answers can be summarised as follows:

- the testers find the system easy to use and conclude that it does not require any long and specialised training or a dedicated person support (note that the blind test participants are familiar with mobile phones with touch screens),
- system ergonomic properties were highly graded by the testers (the testers did not complain about the weight of the depth camera placed on the head),
- the testers noted that the system functions were well integrated and easily accessible,

Table 5 System usability questionnaire, answers in the Likert scale [26] (from 1—strongly disagree to 5—strongly agree)

No	Question	Tester		
		1	2	3
Q1	I think that I would like to use this system frequently	3	3	5
Q2	I found the system unnecessarily complex	2	2	1
Q3	I thought the system was easy to use	4	5	5
Q4	I think that I would need the support of a technical person to be able to use this system	2	2	1
Q5	I found the various functions in this system were well integrated	3	3	5
Q6	I thought there was too much inconsistency in this system	2	2	1
Q7	I would imagine that most people would learn to use this system very quickly	4	5	4
Q8	I found the system very cumbersome to use	2	2	1
Q9	I felt very confident using the system	3	3	5
Q10	I needed to learn a lot of things before I could get going with this system	2	2	2
Overall score		67.5%	72.5%	95.0%

- only Tester 3 felt very confident while using the system, the other testers ranked the system functionality as satisfactory,
- the best system usability score was given by Tester 3, i.e. the congenitally blind participant of the conducted tests.

5 Conclusions and future works

An original interactive sonification technique for the purpose of 3D scene representation for the visually impaired people was devised, implemented and tested. The method does not sonify the information represented by the recorded images of 3D scenes directly (as is the case of the vOICe [7]) but employs specially processed depth images termed the U-depth maps. Such a representation allows the user to effortlessly identify distance and angular direction to potential obstacles. An important feature of the system is that a 10 min introduction to the interface appears to be sufficient for the visually impaired testers to use the system efficiently.

The prototype application was tested by three blind users, who managed to successfully complete three indoor mobility tests: (1) a walk along an empty corridor, (2) a walk along a corridor with obstacles and (3) finding and passing through and open space simulating an open door. The system logs collected during the trials helped make interesting quantitative observations about the modes that the testers employed to interact with the system, e.g. Tester 2 made frequent use of the verbal obstacle detector (see right column in Table 4 and arrows in Figs. 6, 7, 8). Additionally, for majority of the tests, the testers were mainly using the interactive sonification mode for obstacle detection rather than the proximity mode, which functions as an automatic noninteractive short distance ($z < 1.5$ m) detector of the very close obstacles. This observation leads to a conclusion that being able to interact is a very welcome feature of the proposed travel aid interface. Finally, the tested interactive sonification system of 3D scenes passed the system usability questionnaire with good and very good scores. As for now the smartphone platform for the assistive device is an accepted solution because of small size factor, weight, considerably high computing power and users' experience in using such a mobile device.

Future study will focus on improving the proposed sonification method in order to be better interpreted by blind people. We have noticed that distance probing system functionality with the verbal mode was relatively often used by the testers to confirm what is the distance to an obstacle that is sonified according to the adopted distance to frequency mapping. Therefore we will focus on fine-tuning this mapping and we will perform more practical sonification tests with a larger group of the visual impaired people. Also, it is considered to assign special alerting sounds for signalling moving objects. The pitch shift in these sounds will inform whether the object is approaching or moving away. Moreover, a number of new sound schemes will be tested for a better recognition of different object classes, e.g. doors, walls, poles. Adding more verbal comments enriching description of the environment geometric properties is also considered. Finally, to limit accidental collisions with obstacles, the point clouds which represent obstacles should be tracked using calculated user ego-motion parameters, when they leave the camera's field of view [16].

Acknowledgements We are grateful to visually impaired persons taking part in the conducted mobility trials.

References

1. Maidenbaum S, Abboud S, Amedi A (2014) Sensory substitution: closing the gap between basic research and widespread practical visual rehabilitation. Neurosci Biobehav Rev 41:3–15
2. Bach-y Rita P (1972) Brain mechanisms in sensory substitution. Academic Press, Cambridge
3. Dakopoulos D, Bourbakis NG (2010) Wearable obstacle avoidance electronic travel aids for blind: a survey. IEEE Trans Systems Man Cybern Part C (Appl Rev) 40(1):25–35
4. Velázquez R, Pissaloux E (2008) Tactile displays in human–machine interaction: four case studies. Int J Virtual Real 7(2):51–58
5. Csapó Á, Wersényi G (2013) Overview of auditory representations in human–machine interfaces. ACM Comput Surv (CSUR) 46(2):19
6. Bujacz M, Strumillo P (2016) Sonification: review of auditory display solutions in electronic travel aids for the blind. Arch Acoust 41(3):401–414
7. Meijer PB. The vOICe, reasearch project webpage. https://www.seeingwithsound.com/. Accessed 21 June 2018
8. Hermann T, Hunt A (2005) Guest editors' introduction: an introduction to interactive sonification. IEEE Multimed 12(2):20–24. https://doi.org/10.1109/MMUL.2005.26
9. Yoshida T, Kitani KM, Koike H, Belongie S, Schlei K (2011) Edgesonic: Image feature sonification for the visually impaired. In: Proceedings of the 2nd augmented human international conference, AH '11. ACM, New York, pp11:1–11:4. https://doi.org/10.1145/1959826.1959837
10. Ramloll R, Yu W, Brewster S, Riedel B, Burton M, Dimigen G (2000) Constructing sonified haptic line graphs for the blind student: first steps. In: Proceedings of the fourth international ACM conference on assistive technologies, Assets '00, ACM, New York, pp 17–25. https://doi.org/10.1145/354324.354330
11. Banf M, Blanz V (2012) A modular computer vision sonification model for the visually impaired. In: Proceedings of the international conference of auditory display. Georgia Institute of Technology
12. Dal Mutto C, Zanuttigh P, Cortelazzo GM (2012) Time-of-flight cameras and Microsoft KinectTM. Springer, Berlin
13. Sell J, O'Connor P (2014) The xbox one system on a chip and kinect sensor. IEEE Micro 34(2):44–53
14. Microsoft Corporation. One Microsoft Way, Redmond. http://www.microsoft.com. Accessed 21 June 2018
15. Occipital Inc. 1801. In: 13th St 202, B.C..S.Z. https://structure.io/. Accessed 21 June 2018
16. Skulimowski P, Strumillo P (2008) Refinement of depth from stereo camera ego-motion parameters. Electron Lett 44(12):729–730. https://doi.org/10.1049/el:20080441
17. Benacer I, Hamissi A, Khouas A (2015) A novel stereovision algorithm for obstacles detection based on UV-disparity approach. In: Proceedings of IEEE international symposium circuits and systems (ISCAS), pp 369–372. https://doi.org/10.1109/ISCAS.2015.7168647
18. Lin Y, Guo F, Li S (2014) Road obstacle detection in stereo vision based on UV-disparity. J Inf Comput Sci 11(4):1137–1144
19. Labayrade R, Aubert D (2003) In-vehicle obstacles detection and characterization by stereovision. In: Proceedings of the 1st international workshop on in-vehicle cognitive computer vision systems, Graz, Austria
20. Labayrade R, Aubert D, Tarel JP (2002) Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. In: Proceedings of IEEE intelligent vehicle symposium, vol 2, pp 646–651. https://doi.org/10.1109/IVS.2002.1188024
21. Arnell F, Petersson L (2005) Fast object segmentation from a moving camera. In: Proceedings of intelligent vehicles symposium IEEE, pp 136–141. https://doi.org/10.1109/IVS.2005.1505091
22. Azevedo VB, De Souza AF, Veronese LP, Badue C, Berger M (2013) Real-time road surface mapping using stereo matching, v-disparity and machine learning. In: The 2013 international joint conference on neural networks (IJCNN), IEEE, pp 1–8
23. Navigating by interactive sonification of U-depth maps—emphatic trials of an electronic travel aid for the blind. Demo movie of the proposed system. (2018). http://eletel.p.lodz.pl/pskul/u_depth_sonification/
24. WHO. International classification of diseases for mortality and morbidity statistics. http://www.who.int/classifications/icd/en/. Accessed 21 June 2018
25. Lewis JR (1991) Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. ACM Sigchi Bull 23(1):78–81
26. Brooke J (1996) SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland AL (eds) Usability evaluation in industry. Taylor and Francis, London, pp 189–194