INTERACTIVE SONIFICATION OF THE U-DISPARITY MAPS OF 3D SCENES

Piotr Skulimowski, Mateusz Owczarek, Andrzej Radecki, Michał Bujacz, Paweł Strumiłło

Lodz University of Technology
Lodz, Poland
piotr.skulimowski@p.lodz.pl

ABSTRACT

In this paper we propose a method for real-time, interactive auditory representation of a 3D scene's geometric structure by sonifying its U-disparity maps. The U-disparity is derived from the depth map obtained from stereovision imaging of 3D scenes, and can be interpreted as a bird's eye view of a scene with highlighted scene objects. The user can interactively select the region of the U-disparity map for sonification. Such a representation allows the user to effortlessly identify distance and angular direction to potential obstacles. The prototype application was tested by three blind users, who managed to localize key objects in the sonified 3D indoor and outdoor scenes.

1. INTRODUCTION

The visually impaired people indicate limited mobility as the major problem 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 termed "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 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 for the blind that used tactile modality.

Dynamic development of Information and Communications Technologies (ICT) at the turn of centuries (100 years after seminal efforts by Noiszewski) have opened new prospects for designs of 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 review of wearable obstacle avoidance ETAs is given in [3]. Due to size factor and cost, in the majority of ETAs, the auditory displays are favoured as opposed to haptic interfaces that require complex circuitry to implement haptic stimulations. There are many possible auditory representations of information than can be employed in human-machine interfaces (HMI were widely reviewed in [4]). However, sonification, i.e., non-speech audio, is the method which is 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 [5]. Here it is worth mentioning the vOICe (www.seeingwithsound.com), a widely popularized method for sonifying monochrome images. The sonification method is simple, however, not too intuitive and requires many weeks of training. 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. Such a sonification code is used in a repetitive, 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 a recent decade an important subfield of sonification has emerged, namely: interactive sonification [6]. In such an approach to human-computer auditory interface the user is capable of interacting with the sonification process, e.g. define an image region to be sonified or tune sonification parameters to individual requirements.

In this paper we demonstrate how the technique of interactive sonification can be applied in a simple interface for the blind with an aim to represent spatial 3D geometry of the environment. We use the so-called depth images and their histograms termed "U-disparity" representation of the disparity map obtained from a stereovision system sensing of the environment. The U-disparity maps are locally sonified in response to the user's tactile exploration of such an image model of the environment.

2. STEREOVISION BASICS

2.1. Disparity map calculation

Stereovision is a passive 3D reconstruction method from two (or more) images of the same scene captured from different locations in space. The main advantage of this approach is that it does not need any active lighting and efficient algorithms for reconstructing 3D scene geometry are well developed. After applying a proper calibration procedure of the stereovision system, the depth map representing a 3D structure of the imaged scenes can be reliably calculated. For a calibrated stereovision system the disparity (parallax between left and right image) can be computed as $d = x_l - x_r$, where x_l and x_r are the coordinates of the pixels in the left and right stereovision image being the projections of the same point in space. Having calculated the disparity values for the entire image, depth of scene points can be calculated as:

$$Z = \frac{Bf}{d} \tag{1}$$

where B is the distance between optical centres of the cameras, f is the focal length of the camera. Fig. 1 shows distance as a function of disparity for the selected cameras. Fig. 2 shows the left image captured by the Bumblebee stereovision camera. The corresponding depth map is shown in pseudo-colours (the closer the scene point to the camera the warmer the colour) and a greyscale representation is used for displaying the U-disparity map (explained in the next section).

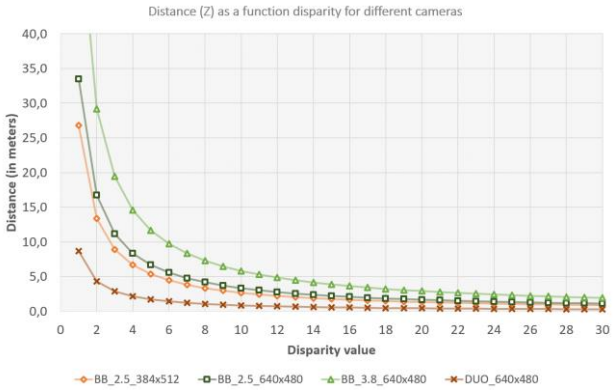

Figure 1. *Distance Z(d) as a function of disparity for different stereo-cameras featuring different image resolution and focal lengths: BB-Bumblebee stereo camera, DUO-Duo MLX stereo camera.*

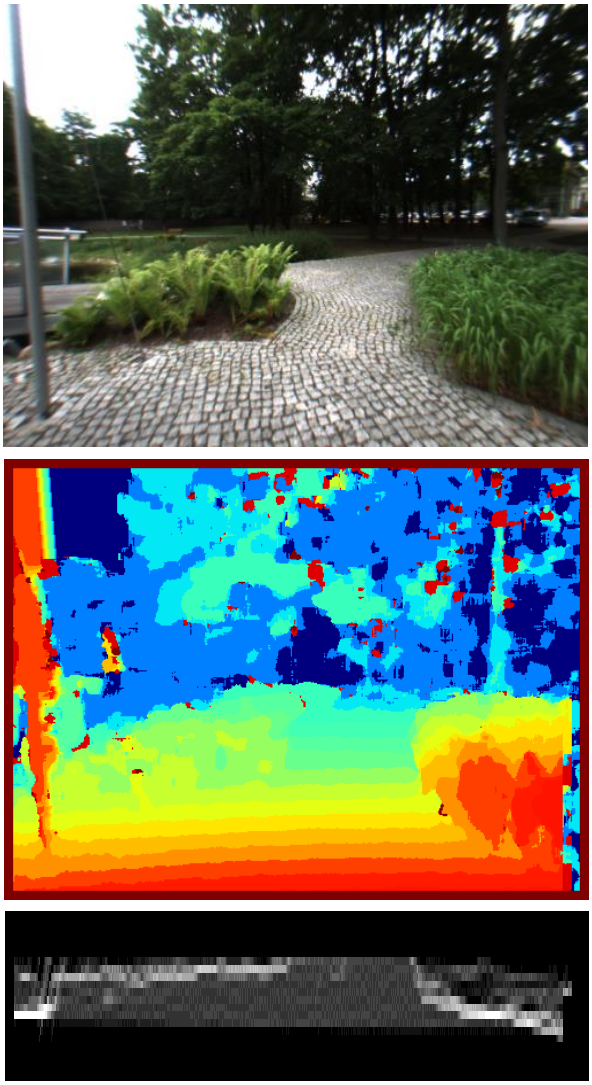

Figure 2. *An image recorded by the Bumblebee stereovision camera (top), the corresponding depth map shown in pseudo-colours (centre) and the corresponding U disparity map (bottom).*

2.2. "U-disparity" representation

Literature review shows that the U-disparity representation of the environment from stereovision can be very effective in obstacle detection for automotive and autonomous robots applications [7,8,9]. The U-disparity representation is built from the depth map by computing histograms of consecutive columns of the depth map. Let the reference image (and disparity image) has pixel resolution $w \times h$. 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 map turns out to be a very efficient representation for localizing scene obstacles (provided the stereovision camera base is parallel to the ground plane [10]). An obstacle located at a well localized distance usually features very many pixels in the disparity map of the same value, which results in high pixel values in the U-disparity map. An example of the U-disparity map for the outdoor scene is shown in the bottom image of Fig. 2. Note that key obstacles, the pole on the left and high grass on the right and left are clearly highlighted in the U-disparity map.

The justification for using the U-disparity map for scene sonification is the representation in which vertical direction denotes depth of scene objects and horizontal direction the azimuth angle of the objects. The U-disparity map is much easier for tactile exploration for the visually impaired user than the depth map, in which vertical direction of the map cannot be associated with depth of scene points (due to different possible positioning of the stereovision camera versus the environment).

3. SYSTEM ARCHITECTURE

A simple schematic of the proposed interactive sonification system is shown in Fig. 3.

Figure 3. *A simplified architecture of an electronic system for interactive sonification of 3D scenes.*

Stereovision camera is mounted on a user's head and is connected to the embedded platform via a WiFi router (Fig. 4).

Figure 4. *Duo MLX stereovision system on a custom-made "glasses"*

The embedded platform (NVIDIA Jetson TK1) is battery powered and attached to a special belt. Image preprocessing procedures and disparity map calculation algorithms are running on the embedded platform. An Android phone is connected to the acquisition system via a Wi-Fi network and the depth map data is transmitted to the mobile phone. The U-disparity map is calculated on the Android-based platform. The mobile phone can be hidden in the user's pocket. The user touches the screen and the selected region of the U-disparity map is sonified. The sonification output stream comes from the mobile phone through the speaker or stereo headphones.

4. INTERACTIVE SONIFICATION OF THE U-DISPARITY MAPS

The blind user can select a scene area for sonification by touching the mobile phone screen on the panel displaying the U-disparity map (see Fig. 5). Let *x* denote a column of the map indicated by the user. The depth map is displayed in the current version of the application for verification purposes. In the release version of the application the U-disparity map will be scaled-up to the full screen width of the mobile device. Touching the centre of the map gives information about the obstacles in front of the user. The sound sonifying the scene depends on the content of the U-disparity map. The indicated column *x* of the map controls left-right panning of the generated sound:

$$a_{Li} = 1 - x/w$$
$$a_{Ri} = x/w \quad (2)$$

where a_{Li}, a_{Ri} are amplitudes (volumes) of the output left and right channels of the *i* sound component.

It is worth noting that such a sonification method of the obstacle horizontal position (left/right panning) is very simplified and it is related to the depth values instead of world coordinates.

The row in the U-disparity map (i.e. the disparity value) in the sonified range determines the sound frequency that codes the depth information (the higher the pitch the closer the sonified object). The sound signal generated by the system is a sum of sinusoids:

$$s(t) = \sum_{i=0}^{i=N-1} a_i \sin\left(2\pi\left(f_{\min} + i\frac{f_{\max} - f_{\min}}{N-1}\right)t\right) \quad (3)$$

Each sinusoid frequency represents the selected distance range. Distance ranges are linked to the discreet values of the disparity map. It was decided to sonify only objects which distance (Z coordinate) from the camera is below a predefined value *Z_max(d_off*) corresponding to disparity *d_off*. We define *N* as the number of different sound frequencies, *f_min* is the frequency of sound with index 0, and *f_max* is a frequency of sound *N*-1 which corresponds to the closest objects. Then the amplitude of each sinusoid is calculated as:

$$a_i = \frac{u[x, d_{off}+1+i]}{h} \text{ for } i < N-1$$
$$a_{N-1} = \frac{\sum_{j=N}^{dmax-d_{off}} u[x, d_{off}+j]}{h} \quad (4)$$

where *h* is the number of rows of the disparity map. It can be noticed that the highest frequency sound source *f_max* represents all objects for which disparity *d* is larger or equal to *d_off+N, i.e. it corresponds to the all closest objects* (please see the example in Fig. 6 and region A indicated by asterisk). Constants *d_off* and *N* in the proposed method depend on the geometric parameters of the selected camera (see Fig. 1) so that a proper control of the sonified depth range can be specified. The *f_min* and *f_max* values were selected to match technical limitation of the speakers built-in into mobile devices that were used for tests.

5. IMPLEMENTATION AND TESTS

5.1. Implementation details on the Android platform

An Android application implementing the described sonification concept acts as a remote control to the system. The application enables to set selected reconstruction parameters like type of disparity calculation algorithm and its parameters such as window size, LED brightness level or confidence parameters for disparity map calculations. Fig. 5 shows the setting panel for the Duo MLX stereovision camera.

For the test purposes the current application also enables to record image sequences with corresponding timestamps for later re-play and analysis. Data transfer between the embedded platform and the mobile phone is provided by the WebSocket protocol, featuring full-duplex communication and a TCP connection. The advantages of using the mobile instead of a dedicated device is its relatively low price and its common use among blind persons.

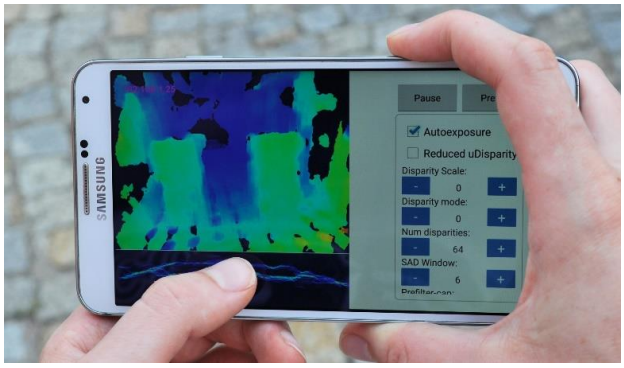

Figure 5. *Depth map of the corridor scene with two obstacles (top-left panel) recorded with the Duo MLX stereovision camera and the corresponding U-disparity map (bottom-left). The depth map is coded using pseudo-colours. The control panel (right) allows to adjust both camera settings and disparity calculation parameters.*

5.2. Tests of the proposed sonification methods

First tests with the blind users, to ensure the repeatability, have been carried out using the pre-recorded test sequences. The sequences were recorded by the DUO MLX stereovision camera, which is used as the main camera in the system prototype. For the development purposes we used the Point Grey colour stereovision camera which is less portable due to its size, but produces much better quality of the images and was used in the tests of the proposed sonification method. The tests were carried out with 3 blind and 2 sighted persons on the chosen outdoor image sequences. All testers were familiar with

the mobile phones with touch screens and they were instructed how the U-disparity map is generated and acquainted with the sonification method of the U-disparity data.

Figure 6. *Offline tests with the use of Android OS tablet. The test scene was captured using the Bumblebee camera. Original reference image is shown in Fig. 7*

Figure 7. *Outdoor scene used in tests with the blind users captured using the Bumblebee camera.*

Fig. 6 illustrates how the trials were conducted. The users were asked to verbally describe the scene based on the sound generated by the mobile device. All users correctly found the obstacles and were able to state, which object is closer to the camera. They were also able to find and indicate directions corresponding to scene spaces devoid of obstacles. It must be noted here that the type of obstacles (e.g. tree, bench, lamp) cannot be communicated to the blind user by means of this scene representation scheme. The user, however, can determine the size, distance and direction of an obstacle. These obstacle features are important for safe mobility in a sonified scene.

Fig. 8 shows spectra of the sounds generated for the selected columns of the U-disparity map. The sounds were recorded from the mobile device using the phone output. It can be noticed that frequency components correspond to the obstacles' disparity values. For region A, a dominant single frequency results from the fact that objects which are very close to the camera (i.e. their disparity values are greater or equal to d_off+N) are coded using a single frequency component (f_{max}), which corresponds to the closest object. Please visit http://eletel.p.lodz.pl/pskul/ison2016 to watch a short video of the proposed method.

The users were asked to express their opinion about the proposed sonification method. They reported no problems in finding and identifying the location of the obstacles and in describing spatial features of the sonified scene. For short time sessions the generated sounds were acceptable for the users, but they noted that for longer sessions listening to such sounds would be tiring.

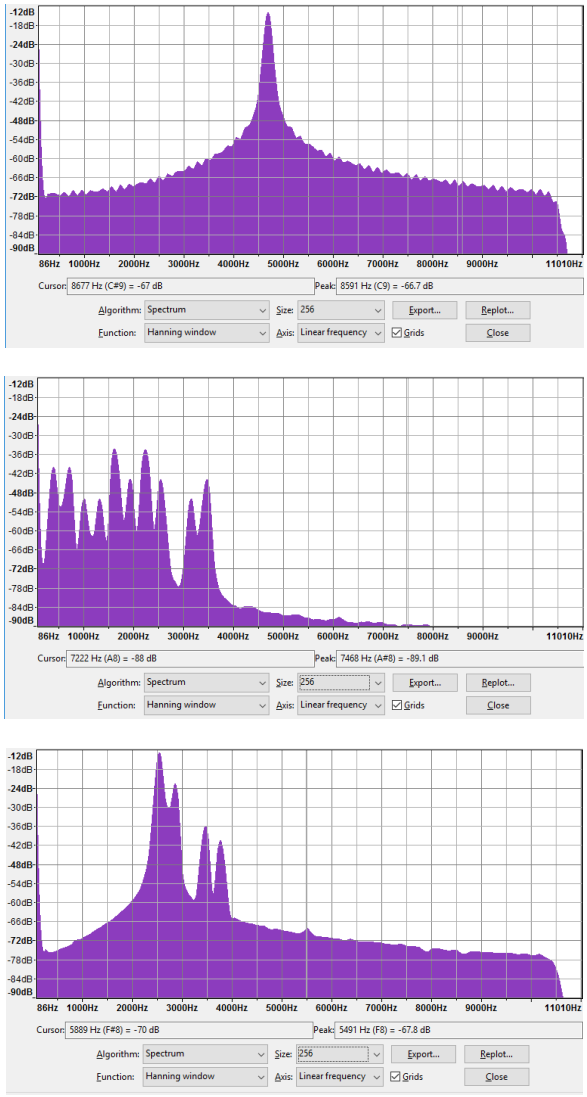

Figure 8. *Fourier spectra of the sounds generated for regions A, B, C of the U-disparity map (see Fig. 6). The sounds were recorded from the phone output of the mobile device.*

The suggested application is not a typical ETA to be used while walking, but as a "look-around" tool that allows a user to interactively study the environment layout. Another possible direction for improving the acceptance of the device by the users is to remove the non-obstacle area from the U-disparity map (see the grey region in the central part of the map in the bottom image of Fig. 2) to reduce the complexity of the generated sounds.

Although, the presented sonification method is very simple, it can serve as a useful tool for aiding the visually impaired in space orientation tasks. Its main disadvantage is that the scene needs to be explored by touch in a serial manner. This may take some time before the user fully comprehends the image content and spatial layout. Our current work is concentrated on developing a sonification system that will combine interactive

sonification and classic audio description. Such an approach is termed *sonic description*. It introduces a complete set of sounds with unique time-frequency characteristics that describe an image sonically in a parallel form after a prior segmentation and recognition of scene elements. This will allow, after a proper machine learning process, a quick interpretation of the observed environment.

6. CONCLUSIONS

An original interactive sonification technique for the purpose of 3D scene representation for the visually impaired people was devised and implemented. The method does not sonify the information represented by the recorded images of 3D scenes directly (as is the case of the vOICe) but employs the processed depth images termed the U-disparity maps. Such maps allow the blind user to interactively explore depth-azimuth space so facilitating search and localization of obstacles. First trials of such an interactive sonification scheme with three blind volunteers shows a potential use of the system as a spatial orientation aid for the visually impaired.

Acknowledgements: This work was partially supported by the National Science Centre of Poland under grant no 2015/17/B/ST7/03884 in years 2016-2018 and by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 643636 "Sound of Vision."

7. REFERENCES

[1] S. Maidenbaum, S. Abboud, A. Amedi, "Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation," in *Neuroscience and Biobehavioral Reviews*, 2014, 14, 3–15.

[2] P. Bach-y-Rita, *Brain Mechanisms in Sensory Substitution.* Academic Press, 1972.

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

[4] A. Csapo, G. Wersenyi, "Overview of auditory representations in human-machine interfaces" in *ACM Computing Surveys*, 2013, 46(2), 19:1–19:23.

[5] M. Bujacz, P. Strumillo, "Sonification: review of auditory display solutions in electronic travel aids for the blind", in *Archives of Acoustics*, 2016, 41(3), 401–414.

[6] T. Hermann, A. Hunt, "An Introduction to Interactive Sonification" in *IEEE Multimedia*, April–June 2005, vol. 12, no. 2, pp. 20–24.

[7] I. Benacer, A. Hamissi, A. Khouas , "A novel stereovision algorithm for obstacles detection based on U-V-disparity approach" in *International Symposium on Circuits and Systems*, 2015.

[8] Y. Lin, F. Guo, S. Li, "Road Obstacle Detection in Stereo Vision Based on UV-disparity", Journal of Information & Computational Science, 2014, 11, (4), pp. 1137–1144.

[9] R. Labayrade, D. Aubert, "In-vehicle obstacles detection and characterization by stereovision", in *Proceedings the 1st International Workshop on In-Vehicle Cognitive Computer Vision Systems*, 2003, pp.13–19.

[10] V. Azevedo, A. Souza, L. de Paula Veronese, C. Badue, M. Berger, "Real-time Road Surface Mapping Using Stereo Matching, V-Disparity and Machine Learning" in *International Joint Conference on Neural Networks*, 2013.