Auditory Augmented Reality: Object Sonification for the Visually Impaired

Flávio Ribeiro [#1], Dinei Florêncio [*2], Philip A. Chou [*3], and Zhengyou Zhang [*4]

[#] *Electronic Systems Eng. Dept., Universidade de São Paulo, Brazil*
[1] `fr@lps.usp.br`

[*] *Microsoft Research, One Microsoft Way, Redmond, WA*
[2,3,4] `{dinei,pachou,zhang}@microsoft.com`

***Abstract*—Augmented reality applications have focused on visually integrating virtual objects into real environments. In this paper, we propose an auditory augmented reality, where we integrate acoustic virtual objects into the real world. We sonify objects that do not intrinsically produce sound, with the purpose of revealing additional information about them. Using spatialized (3D) audio synthesis, acoustic virtual objects are placed at specific real-world coordinates, obviating the need to explicitly tell the user where they are. Thus, by leveraging the innate human capacity for 3D sound source localization and source separation, we create an audio natural user interface. In contrast with previous work, we do not create acoustic scenes by transducing low-level (for instance, pixel-based) visual information. Instead, we use computer vision methods to identify high-level features of interest in an RGB-D stream, which are then sonified as virtual objects at their respective real-world coordinates. Since our visual and auditory senses are inherently spatial, this technique naturally maps between these two modalities, creating intuitive representations. We evaluate this concept with a head-mounted device, featuring modes that sonify flat surfaces, navigable paths and human faces.**

***Index Terms*—augmented reality, natural user interface, sonification, spatialization, blind, visually impaired.**

I. Introduction

According to the World Health Organization, there are an estimated 39 million blind people in the world [1]. In the United States, there are an estimated 1.3 million legally blind individuals [2], with approximately 109,000 who use long canes and 7,000 who rely on guide dogs [3]. Since vision impairments hinder a wide variety of human activities, assistive devices have been designed to facilitate specific tasks or enhance mobility.

We start from the observation that our two highest bandwidth senses – vision and hearing – have spatial structure. Using spatialized (3D) audio, we synthesize virtual acoustic objects, placing them at specific real-world coordinates. Thus, by leveraging the innate human ability for 3D sound source localization, we can relay location-dependent information without having to explicitly encode spatial coordinates.

In this paper, we combine this approach with computer vision techniques to create natural high-level scene representations for the visually impaired. For example, we use face recognition to detect known individuals, who can then virtually identify themselves using recordings of their own voices, which appear to originate from their real-world locations. On one hand, computer-vision adds a layer of cognition which promotes relevance and produces tremendous bandwidth savings. On the other hand, spatial audio eliminates the overhead of encoding, transmitting and decoding spatial coordinates.

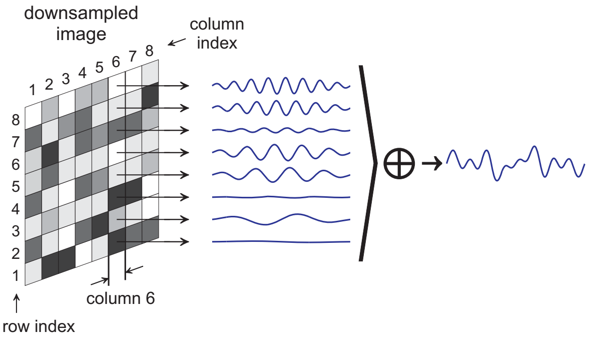

Figure 1. Encoding used by The vOICe

In recent years, there have been several proposals for translating visual information into audio by encoding low level features. For instance, several methods were proposed to encode a bitmap image one pixel column at a time, using frequency-domain multiplexing for each column. The vOICe [4] was the first proposal for a wearable device of this kind. A camera acquires a bitmap image of up to 64x64 pixels with 16 shades of gray per pixel, and the system encodes columns in left-to-right order. Given a column, each pixel controls a sinusoidal generator, with its value determining amplitude, and its coordinate being proportional to the frequency (see Fig. 1). At a given moment, the user hears the superposition of all the sinusoids from a column. After all columns have been rendered, a synchronization click is generated and the process restarts.

The literature features several variations of this approach. In [5], an RGB camera image was first reduced to 1 bit per pixel, and pixels were associated with musical notes. A black and white image was then be mapped into a melody. SVETA [6] was a more recent proposal which transduced a disparity image obtained from stereo matching. To reduce user fatigue, pixels were associated with major chords instead of pure sinewaves.

Unless the image is very sparse, encoding every pixel

978-1-4673-4572-9/12/$31.00 ©2012 IEEE

generally produces an overwhelming combination of sounds. The representation is difficult to interpret, and despite attempts to use more pleasant sounds, even short-term use produces significant user fatigue. To reduce the amount of information to be encoded, See ColOr [7], [8] used segmentation to create a cartoon-like image, and objective saliency methods to estimate the most relevant regions. To transduce a simplified RGB-D image, hue was encoded by the timbre of a musical instrument, saturation by one of four possible notes, and luminosity with bass or a singing voice. Pixel locations were spatialized using 9 ambisonic channels. To avoid overwhelming the listener, only a small window (chosen by the user on a tactile tablet) was encoded.

Even though segmentation and saliency reduce the amount of information, color adds another dimension which must be encoded. Since current state-of-the-art objective saliency methods are only based on low-level features such as luminance, contrast and texture, they completely neglect cognitive aspects which often determine regions of interest. These issues are reflected in a See ColOr user study, where participants required an average of 6.6 minutes to locate a red door on an image which had no other red features [8].

Thus, it becomes clear that given the bandwidth limitations of audio, it is important to avoid low-level representations and arbitrary encodings. In this paper, we present our preliminary work intended to address these issues. We acquire and transduce data in real-time using a helmet mounted RGB-D camera, an inertial measurement unit and open-ear headphones. We illustrate the concept by applying computer-vision methods for plane decomposition, navigable floor mapping and object detection. This representation carries more cognitive content and is much more summarized than a raster scan, and thus can be relayed without overwhelming the user. By sonifying high-level spatial features with 3D audio, users can use their inherent capacity for sound source localization to identify the position of virtual objects. Thus, we intuitively represent coordinates from visual space in auditory space, avoid using explicit arbitrary encodings, and the representation is summarized further.

This paper is organized as follows. Section 2 presents the designed system, which includes an audio spatialization engine and components for plane detection, floorplan estimation and face recognition. For each component, we describe how outputs are sonified into 3D audio. Section 3 describes experiments used for evaluating our auditory mapping techniques. Finally, Section 4 has our conclusions and directions for future work.

II. System description

Figure 2 shows the block diagram for the proposed device. Visual input is captured at 640x480 pixels with the RGB-D camera module used in the Microsoft Kinect (see Figure 3). The RGB and depth cameras were calibrated to produce an accurate correspondence between depth and RGB pixels.

This proof of concept version implements modules for plane detection, floorplan estimation and face recognition. High-level visual information is represented using a combination of pre-recorded wave files, a text to speech synthesizer and a musical instrument synthesizer. Every sound is spatialized in 3D space using HRTFs from the CIPIC database [9]. As described below, head tracking is implemented with a 3D gyroscope. A 3D accelerometer is used to estimate the gravity vector, and infer the location of the floor plane.

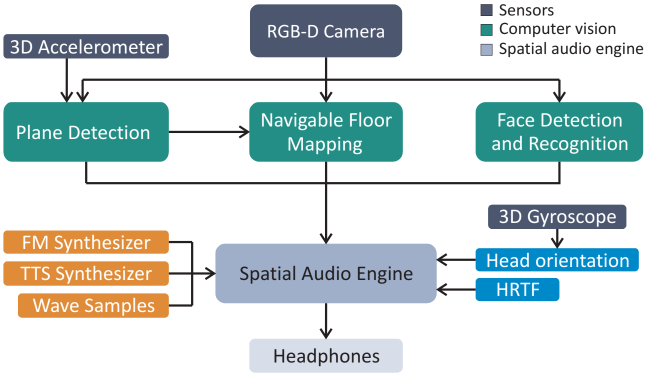

Figure 2. System block diagram

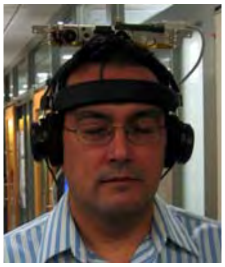

Figure 3. Prototype device

A. Audio spatialization

When representing real-world elements using audio, one must establish how to sonify real-world coordinates. Several previous proposals have relied on arbitrary encodings (for example, using frequency to represent vertical position [4]–[6]). Instead, we render the location of an object by synthesizing a virtual sound source at its corresponding real-world coordinates. The source is spatialized with 5° resolution in azimuth and elevation, by filtering with HRTFs from the CIPIC database [9]. We used HRTFs for the KEMAR with small pinnae, interpolated to obtain 5° resolution.

Since the CIPIC HRTFs are measured at a fixed distance of 1.0 m, they are not range dependent and effectively act as a far-field model. We represent range by attenuating the source by 6 dB for each doubling in distance, and by adding a fixed, location-independent reverberation component (see Fig. 4). Intensity and direct-to-reverberant energy ratio are known to be the two primary cues for range, and have complementary roles for relative and absolute range perception [10]–[12]. We use a time-invariant virtual room for synthesizing reverberation, with a reverberation time of 300 ms. Having a fixed virtual room is useful, because users are known to learn the acoustic

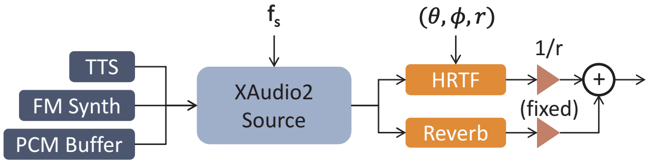

Figure 4. Spatialized source

characteristics of an environment, and improve their sound source localization performance with time [13].

HRTF filtering is performed with an FFT accelerated convolution engine, with a typical latency of 10 ms. Each audio source is associated with a playlist of wave samples, speech utterances (produced by a text-to-speech synthesizer) or musical notes (produced by an FM synthesizer). To facilitate the description of real-world features, a source can also describe parameterized curves in real-world coordinates.

Head tracking is an integral component of audio spatialization. If the user moves, the virtual audio sources associated with real-world objects should not be dragged along with him. Thus, during the acoustic rendering of a scene (which lasts approximately 1 second) we perform head tracking, and update the coordinates of all 3D audio objects. To simplify tracking, we only estimate the relative rotation between computer vision updates. This rotation is given by the composition of rotation matrices of the form

$$R\left(\theta_x, \theta_y, \theta_y\right) = \begin{bmatrix} 1 & -\theta_z & \theta_y \\ \theta_z & 1 & -\theta_x \\ -\theta_y & \theta_x & 1 \end{bmatrix}. \qquad (1)$$

While the composition of rotations is not commutative, the composition of infinitesimal rotations is. We sample a 3D gyroscope at 40 Hz, such that a relative rotation matrix can be updated by iteratively multiplying by (1) and reorthonormalizing the result.

B. Plane detection

Plane detection is used for two purposes: to identify the floor, and to provide an environmental decomposition into flat surfaces. Indeed, planes are the dominant primitives in man-made environments, and can be conveniently used to identify walls and furniture. Our underlying assumption is that given the decomposition of an environment into planes of known sizes and locations, a user is able to infer which classes of objects they belong to, given contextual clues. For instance, the location of a table could be inferred from the description of a large horizontal plane. Likewise, a chair could be associated with a small pair of horizontal and vertical planes.

Our algorithm for fast plane detection is described in the Appendix. Figure 5 shows an example for plane segmentation, where each plane is drawn with a different color. The sampling rectangles used for each rectangle are also represented.

Figure 6 shows how planes are represented acoustically. The plane detector associates each detected plane with its point cloud. Using the eigendecomposition of this flat set of points, we approximate it as a quadrilateral and produce an estimate of its 4 corners. A vertical plane is represented by a clockwise sequence of musical notes, rising in pitch, with each corner rendered by a virtual source at the corner's real world location. A horizontal plane is distinguished by being represented by a counterclockwise sequence of musical notes, falling in pitch. While other representations are certainly possible, this proved sufficient to relay the concept of a plane.

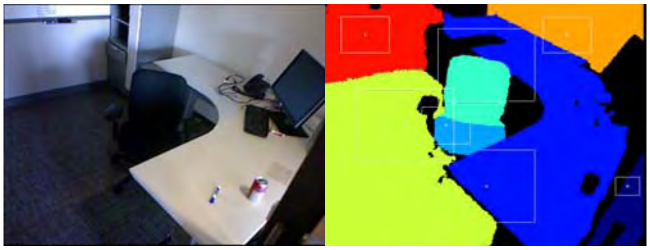

Figure 5. Plane segmentation example

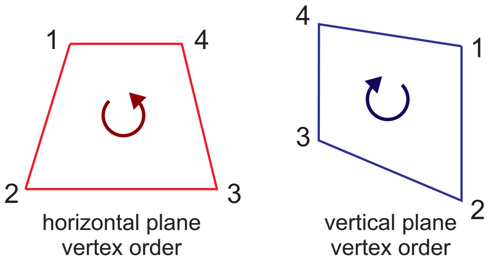

Figure 6. Acoustic rendering for plane representation

C. Floorplan description

Several devices for the visually impaired have relied on ultrasound for describing local geometry. In comparison, depth cameras provide a dramatic improvement in terms of spatial resolution. Furthermore, their range can reach 10 m with current technology. Thus, one can perform local mapping to complement the short range and tactile feedback of a white cane. With this in mind, our device implements a floorplan description mode, which is intended to relay on demand a fast description of the navigable floor.

We define the navigable region to be the visible floor, bounded by obstacles. Visibility is important for safety reasons, as it prevents instructing the user to walk on potentially nonexistent ground. Depth cameras relying on infrared also produce offscale high pixels for glass and black surfaces, effectively classifying them as distant objects. By treating these offscale regions as non-navigable, we prevent collisions with undetectable obstacles.

The plane detector is first used to locate the floor, which is the largest plane with an orientation consistent with the gravity vector (given by the accelerometer). After plane detection, we rotate the entire point cloud so $y = 0$ for all points in the floor. We then project the entire point cloud onto the xz plane, creating a 2D floorplan. We ignore points which are small in height, and fall under the error threshold of the camera. We also ignore obstacles which the user can walk under. The obstacle floorplan is then convolved with a human-like

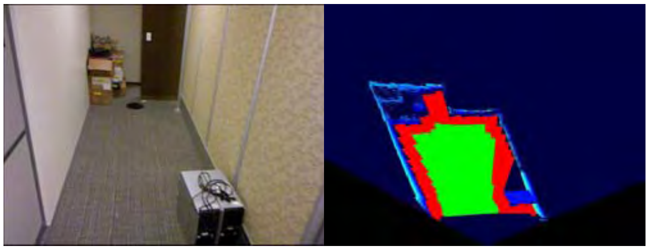

Figure 7. Navigable floorplan example

cross-section, defining a navigable floorplan. From the camera coordinates, we cast a ray for every degree in azimuth, and store the maximum distance that can be traveled before hitting an obstacle.

Figure 7 shows an example of this polar floorplan, which we sonify. The red regions indicate the visible floor which can be reached by line of sight. The green regions are reachable by walking along a straight line.

Figure 8 illustrates how the polar floorplan is acoustically rendered. Even rendering cycles start with a synchronization click, spatialized at $-90°$ in azimuth (the user's absolute left), and describe the polar floorplan in left-to-right order. Odd rendering cycles start with a synchronization click at $+90°$, and describe the floorplan in right-to-left order. The synchronization clicks are important for giving an indication of when the description starts, and for giving the user a reference for absolute left and right. This reference is always useful, and becomes more important if the CIPIC HRTF is significantly mismatched with respect to the user's personal HRTF.

This sequential description was shown experimentally to be preferable to a random sampling, especially in the presence of HRTF mismatch. Indeed, it appears that with a sequential sweep, the spatial sense from a click can be integrated with neighboring clicks, providing a clearer spatial sense.

Following the synchronization click, the floorplan is sampled at every $2°$, generating a low pitched click if an obstacle has been detected at less than 1 m, and a high pitched click otherwise (indicating a navigable direction). Low pitched clicks have constant amplitude, while high pitched clicks are stronger for longer navigable paths[1]. This modulation provides an intuitive cue for navigability, since the user becomes accustomed to walk into the direction of loud high pitched clicks. We note that this agrees with the convention used worldwide for crosswalks.

D. Face detection and recognition

During interviews with visually impaired and blind users, face detection and recognition were suggested as desirable features for an assistive device. Indeed, for a blind individual, not knowing the identity of an approaching person generally implies a missed opportunity for social interaction. Thus, our device integrates face detection and recognition, implemented with [14], [15].

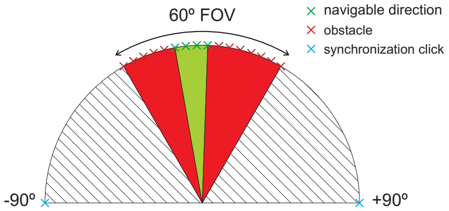

Figure 8. Acoustic rendering for navigation

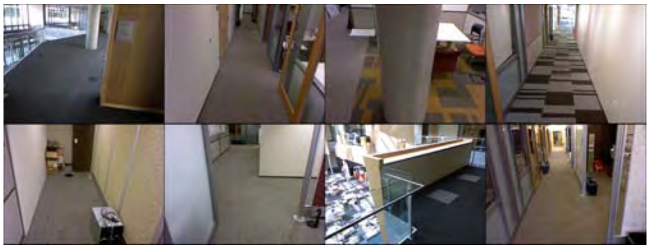

Figure 9. Training set

During start-up, the device loads a database of faces, which are used to recognize individuals who appear on the RGB frame. Each face is represented by the spatialized name of the person, rendered as a virtual source located at the real-world coordinates of the face. When a face is detected but not recognized, the device uses a musical note fallback. With an enhanced face detector, one could potentially use voices for more descriptive fallbacks (male/female, adult/child, etc.).

III. Evaluation

While the mapping from a scene to the auditory representation is straightforward, it produces a summarized description. Thus, it remained to be seen whether the sonifications could be interpreted with sufficient accuracy, enabling the identification of environmental features. To obtain quantifiable and repeatable results, we designed a scene classification task, where sighted users listened to an auditory rendering, and then were asked to choose the scene which best matched it. For this task we used pre-captured still frames, allowing the same data to be presented to multiple participants.

The classification experiment began with a training session, where users were shown an RGB image of a scene and then listened to its associated acoustic rendering. For each example, participants were instructed to notice how specific spatial features corresponded to synthesized sounds. The training set consisted of 8 diverse indoor environments (see Figure 9), which captured a wide variety of detail. During this step, participants were free to ask any questions.

The evaluation phase used 3 sets of 8 scenes (see Figure 10). For each participant, 2 scenes were drawn randomly from each of the 3 sets, without replacement, for a total of 6 scenes. For each scene, the participant only knew which set it came from. His task was to identify which of the 8 set members

[1] High pitched clicks are spatialized at the correct azimuth, with amplitude proportional to $6.0 - r$, where $r \in [0, 5.0]$ is the distance to the nearest obstacle (in meters).

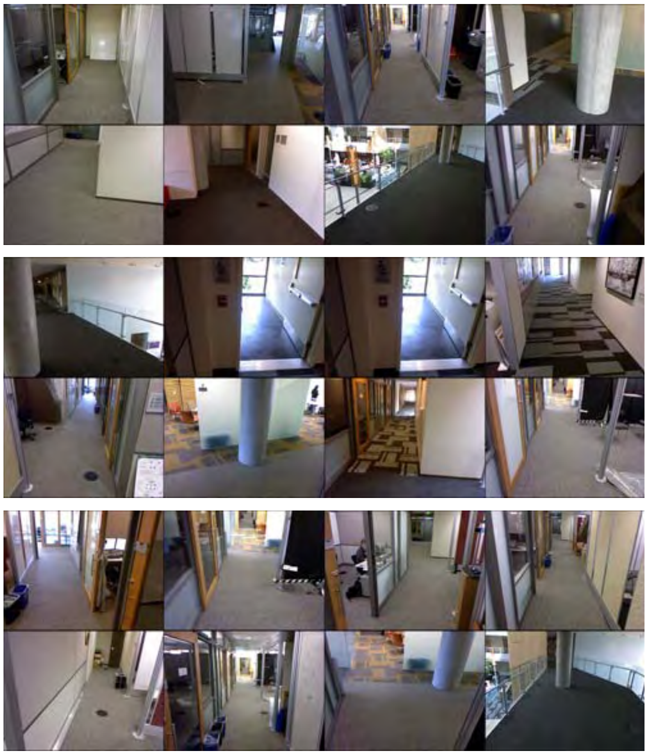

Figure 10. Test sets

best described the acoustic rendering. Participants typically required approximately 1 minute, and no more than 2 minutes to make decisions. We note that a significant fraction of this time was actually spent looking at the features of each of the 8 photographs, and mentally comparing them to the auditory rendering. We expect these times to improve through the use of personalized HRTFs.

Our user study featured 14 participants, who had received no previous training. Figure 11 shows the score histogram for their answers. By observing the experiments, it was clear that some individuals were extremely adept at matching our acoustic representation with visual information. In general, users showed very good results, with 9 participants correctly associating at least 5 of the 6 test scenes. Some users complained about having difficulty localizing spatialized sounds, which could be due to a mismatch between their personal HRTFs and the KEMAR HRTF, or due to physiological constraints (some individuals have poor sound source localization, even in real-life scenarios). Some mistakes were clearly due to limited training, because participants were hearing the representation correctly, but making incorrect inferences (for example, confusing the meaning of the clicks from Figure 8).

With an ideal natural representation, training would only involve learning which sound is assigned to each type of object. Training with a low-level (e.g., pixel-based) representation is much more involved, due to the wide diversity of encoded sounds and to the lack of a one-to-one map. Like a natural representation, our approach sonifies high-level objects. Nevertheless, it still performs some high-level encoding due to the absence of scene interpretation. With advances in computer vision, we foresee the use of progressively more natural descriptions (for instance, a chair could have a unique sound, instead of being represented as two planes).

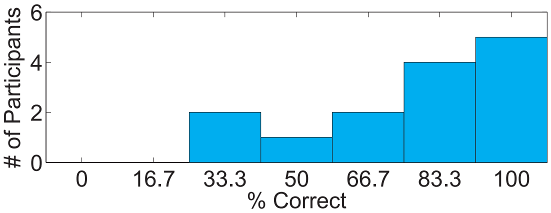

Figure 11. Participant score histogram

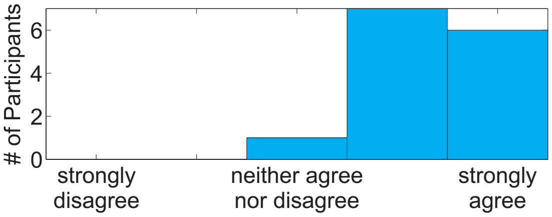

Figure 12. Answers to the survey question "I believe my performance will improve with training"

A detailed study with visually impaired participants is being planned, and initial feedback has been very positive. According to a blind user, "it's a very intuitive device, so I think that you would get used to it very quickly". Plane decomposition was considered useful for unknown environments: "the flat surface mode – I really liked that, because you can detect objects around you and kind of get a feel for how the room is laid out".

Blind interviewees also noted the importance of training. Visually impaired individuals learn to interpret the world through suble cues (for example, a blind person can detect a telephone pole by noticing how traffic sounds diffract around it). In contrast, this device renders environments using a very explicit representation. Thus, a blind person would be expected to adapt to this new language.

IV. Conclusion

In this paper, we described a new approach for representing visual information with spatial audio. This method was implemented using a head-mounted device with a RGB-D camera module, an accelerometer, a gyroscope and open-ear headphones. In contrast with previous proposals, we rely heavily on computer vision for obtaining summarized environmental models, and on audio spatialization for representing spatial locations, thus circumventing the need to encode coordinates.

Preliminary results show that most users can interpret the representations, effectively building mental maps from the acoustic signals, and associating them with spatial data. This is an encouraging result, since these users received little training and personalized HRTFs were not used. For a practical device, we envision measuring personalized HRTFs, similarly to fitting a hearing aid or glasses.

Our proof of concept device features modes for plane decomposition, floorplan estimation and face detection. Plane decomposition could be generalized with other primitives. A practical device would benefit from additional operating modes, such as optical character recognition followed by text-to-speech synthesis, and barcode recognition followed by product lookup. Outdoor urban use could benefit from crosswalk and traffic sign detectors, and GPS integration. Specific context-dependent tasks could also be modeled. In particular, entertainment applications could involve games such as bowling and billiards.

Appendix

Plane detection is performed using the 640x480 point cloud produced by the depth camera. We implemented a fully deterministic approach based on multi-scale sampling, designed to be computationally efficient and robust to noise. While RANSAC and its variants [16], [17] are very effective for fitting a wide range of primitives, plane detection can benefit from a more specialized approach. Our proposal samples a depth frame using uniform rectangular grids which are gradually refined. This approach promotes the fast and robust removal of large flat regions, and progressively searches for smaller plane sections.

At each sampling scale, the depth frame is divided into rectangles with 50% overlap. For each rectangle, a gradient fill is applied at its center, effectively identifying a connected region in 3D space. For sufficiently large connected regions with points $\{(x_i, y_i, z_i)\}_{i=1}^{N}$, we estimate the least-squares plane $ax + by + cz + d = 0$ using the least-squares solution to

$$\underbrace{\begin{bmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_N & y_N & z_N & 1 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \mathbf{0}. \tag{2}$$

Note that the optimal plane is given by the least significant right singular value of $\mathbf{A}$, which is also the least significant eigenvector of $\mathbf{A}^T\mathbf{A}$. In practice, one can subsample the connected region to reduce N.

Similarly to RANSAC, we obtain the set of inliers, i.e., the subset of points which are close to a least-squares plane. Unlike RANSAC, we do not iterate over multiple plane candidates in search for the largest plane. Instead, we either accept a plane candidate at the current scale if the ratio of inliers to the total number of points in the sampling rectangle is sufficiently large, and reject it otherwise. Using a model for the depth camera, we define the inliers such that this ratio test produces plane estimates with a given false positive probability.

For the Kinect depth camera, we assume that the depth map noise is Gaussian, with a variance given by [18]

$$\sigma_z^2 = \frac{\sigma_0^2 z^4}{f_d^2 B^2}, \tag{3}$$

where z is the depth coordinate, $f_d = \frac{f_x + f_y}{2}$ is the mean focal length and σ_0 and B are constants. Let t_0 be the inlier distance threshold at a reference depth z_0. One can use (3) to determine the probability p_0 of having a depth error exceeding t_0. We use (3) to produce a depth-dependent inlier threshold $t(z)$, such that for all z, the depth error exceeds $t(z)$ with constant probability p_0. Assuming the independence of depth errors, for sufficiently large N one should have approximately Np_0 inliers. By comparing Np_0 with the actual number of inliers, one can accept or reject a plane candidate. After extracting inliers for the connected region, we recompute the least-squares estimate, and extract inliers for the entire depth map. The least-squares estimate is computed again, and produces the accepted plane estimate after a final inlier extraction.

This method is more computationally efficient than RANSAC, since it does not require testing a large number of randomly sampled plane candidates. By using a multiscale approach, it is guaranteed to extract large planes first, dramatically reducing the size of the remaining point cloud.

References

[1] W. H. O. (WHO), "Fact sheet 282: Visual impairment and blindness," Oct. 2011, available: http://www.who.int/mediacentre/factsheets/fs282/en/.

[2] N. F. for the Blind, "Blindness statistics," available: http://www.nfb.org/nfb/blindness_statistics.asp.

[3] A. F. for the Blind, "Facts and figures on adults with vision loss," available: http://www.afb.org.

[4] P. Meijer, "An experimental system for auditory image representations," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 2, pp. 112–121, 1992.

[5] J. Cronly-Dillon, K. Persaud, and R. Gregory, "The perception of visual images encoded in musical form: a study in cross-modality information transfer," *Proc. R. Soc. B*, vol. 266, no. 1436, p. 2427, 1999.

[6] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and S. Yaacob, "Wearable real-time stereo vision for the visually impaired," *Engineering Letters*, vol. 14, no. 2, 2007.

[7] G. Bologna, B. Deville, T. Pun, and M. Vinckenbosch, "Transforming 3d coloured pixels into musical instrument notes for vision substitution applications," *EURASIP J. Image Video Process.*, 2007.

[8] G. Bologna, B. Deville, and T. Pun, "On the use of the auditory pathway to represent image scenes in real-time," *Neurocomputing*, vol. 72, no. 4-6, pp. 839–849, 2009.

[9] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. WASPAA*, 2001.

[10] D. Mershon and L. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Attention, Perception, & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975.

[11] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. Acoust. Soc. Am*, vol. 111, no. 4, pp. 1832–1846, 2002.

[12] ——, "Direct-to-reverberant energy ratio sensitivity," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2110–2117, 2002.

[13] B. Shinn-Cunningham, "Localizing sound in rooms," in *Proc. ACM SIGGRAPH*, 2001.

[14] C. Zhang and P. Viola, "Multiple-instance pruning for learning efficient cascade detectors," in *Proc. NIPS*, 2007.

[15] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. CVPR*, 2010.

[16] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer Graphics Forum*, vol. 26, no. 2. Wiley Online Library, 2007, pp. 214–226.

[18] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang, "3d deformable face tracking with a commodity depth camera," *Computer Vision–ECCV 2010*, pp. 229–242, 2010.