Audible Panorama: Automatic Spatial Audio Generation for Panorama Imagery

Haikun Huang*
University of Massachusetts, Boston

Michael Solah*
University of Massachusetts, Boston

Dingzeyu Li
Adobe Research
Columbia University

Lap-Fai Yu
George Mason University

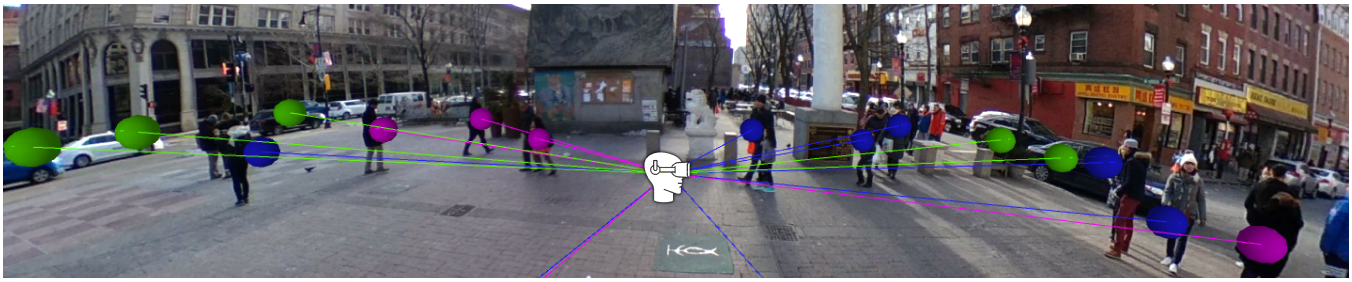


Figure 1: Our approach automatically places audio files as sound sources on a 360° panorama image to enhance the immersion when experienced through a virtual reality headset. In this example, a background audio file with sounds for a town is placed by our approach. Audio files for people chatting (in blue), walking (in purple), and of cars (in green) are automatically assigned as sound sources for the detected objects, and are placed at estimated depths in the scene with respect to the user. Please refer to the supplementary material for the audio results.

ABSTRACT

As 360° cameras and virtual reality headsets become more popular, panorama images have become increasingly ubiquitous. While sounds are essential in delivering immersive and interactive user experiences, most panorama images, however, do not come with native audio. In this paper, we propose an automatic algorithm to augment static panorama images through realistic audio assignment. We accomplish this goal through object detection, scene classification, object depth estimation, and audio source placement. We built an audio file database composed of over 500 audio files to facilitate this process.

We designed and conducted a user study to verify the efficacy of various components in our pipeline. We run our method on a large variety of panorama images of indoor and outdoor scenes. By analyzing the statistics, we learned the relative importance of these components, which can be used in prioritizing for power-sensitive time-critical tasks like mobile augmented reality (AR) applications.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**;
• **Computing methodologies** → *Virtual reality* .

KEYWORDS

immersive media; spatial audio; panorama images; virtual reality; augmented reality

*The two authors contributed equally to this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300851

ACM Reference Format:
Haikun Huang, Michael Solah, Dingzeyu Li, and Lap-Fai Yu. 2019. AudiblePanorama: Automatic Spatial Audio Generation for Panorama Imagery. InCHI Confernce on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, NewYork, NY, USA, 11 pages.https: //doi.org/10.1145/3290605.3300851

1 INTRODUCTION

Sound has been demonstrated to be an integral aspect of immersion [12, 35], so it is no surprise that there have been numerous attempts to produce realistic sound for 3D environments [2, 4].

Given a 360° panorama image, we present an approach to create a realistic and immersive audible companion. We start by detecting the overall scene type as well as all the individual sounding objects. Leveraging scene classification and object recognition, we then assign a background audio of the scene and customized audios for each object.

We use object and scene tags assigned during scene classification and object recognition for audio assignment. These tags are matched with audio tags from our audio file database and audio associated with the audio tags are assigned as audio sources. We estimate the depths of detected objects by comparing relative heights of objects and pre-established knowledge on the average heights of different types of objects.

We have three major contributions.

- We proposed a new framework for automatically assigning realistic spatial sounds to 360° panorama images based on object recognition and depth estimation.
- We constructed a large dataset of panorama images and audio files. The panorama images are made audible by running our approach on them. The dataset, the results obtained by running our approach, as well as the tool for experiencing the audible panorama will be publicly released for research purposes.
- We conducted statistical analysis that evaluates the importance of various factors in our proposed framework.

2 RELATED WORK

As the camera hardware for 360° contents are significantly improving, there is an increasing amount of interests to facilitate better interaction with the 360° contents. However, most existing work focused on the visual aspect, for example, sharing content playback [30], improving educational teaching [1], assisting visual focus [16], augmenting storytelling [8, 23], controlling interactive cinematography [22], enhancing depth perception [13], and collaborative reviewing [19].

On the audio side, Finnegan et al. made use of audio perception to compress the virtual space, in addition to conventional visual-only approaches [5]. Rogers et al. designed and conducted a user study on sound and virtual reality (VR) in games, exploring the player experience influenced by various factors [25]. Schoop et al. proposed HindSight that can detect objects in real-time, therefore greatly enhancing the awareness and safety with notifications [27]. While previous methods all rely on accompanying audio signals, our project tries to enable better interactive experience in 360° images by synthesizing realistic spatial audio from only visual content. We achieve this by constructing a comprehensive audio dataset which we will discuss later.

To enhance the sense of presence enabled by immersive virtual environments, high-quality spatial audio that can convey a realistic spatial auditory perception is important [2, 3, 10, 29]. Many researchers have studied realistic computer-generated sounds. One widely researched excitation mechanism is rigid body sound [20]. To model the sound propagation process, efficient wave-based [24] and ray-based simulation methods [21] have been proposed. More closely related to our method are scene-aware audio for 360° videos [14] and automatic mono-to-ambisonic audio generation [18], both of which require audio as part of the input. We draw inspirations from existing sound simulation algorithms and synthesize plausible spatial audio based on only visual information without any audio input to our system.

In existing virtual sound simulations, scene information, such as the number of objects, their positions, and motions, is assumed to be known. However, in our project, we compute this information automatically through object detection and recognition. In computer vision, robust object detection has been a grand challenge for the past decades. Early work detects objects rapidly, for example, human faces, using carefully designed features [33]. Recently, more robust methods leveraging deep neural networks have been shown to achieve high accuracy [9, 11]. We match the scene in a panorama image with an audio file, but also detect and recognize objects in the image and estimate their depth to place audio sources for those objects at convincing positions.

Since the production of sound is a dynamic process, we need to classify not only the objects, but also their movements, their actions, and their interaction with the surroundings. For example, a running pedestrian and a pedestrian talking on the phone should produce different sounds in the audible panorama. To this end, accurate action recognition can guide the generation of more realistic sounds. Most existing action analysis research requires video as input since the extra temporal information provides strong hints as to what certain human actions are [31, 36]. Human action recognition on still images remains a challenging task. One approach is to use word embeddings and natural language descriptions to recognize actions [28].

Traditional sound texture synthesis can generate credible environmental sounds, such as wind, crowds, and traffic noise. Harnessing the temporal details of sounds using time-averaged statistics, McDermott et al. demonstrated synthesizing realistic sounds that capture perceptually important information [17]. In addition to sound texture, natural reverberation also plays an important role in the perception of sound and space [32]. An accurate reverberation conveys the acoustic characteristics of real-world environments. However, in these ambient sound synthesis methods, the spatial

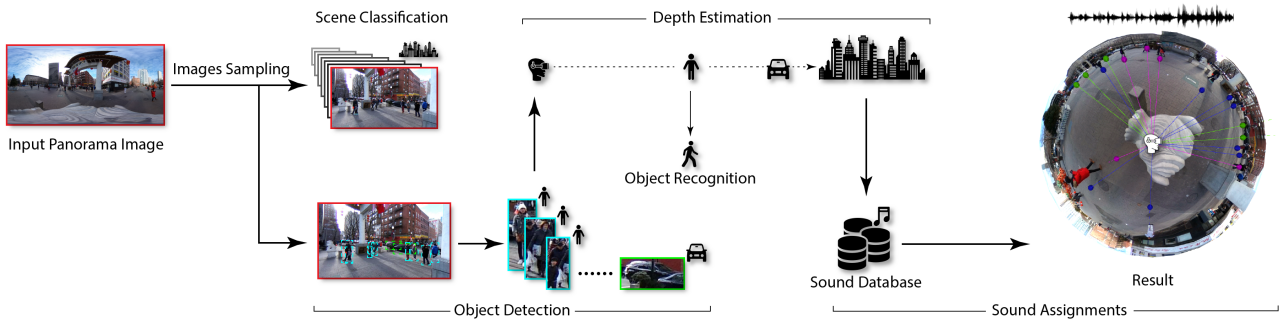


Figure 2: Overview of our approach. Given a panorama image, our approach performs scene classification and object detection on images sampled horizontally from the panorama image. Next, it performs object recognition. Based on the scene knowledge, it assigns a corresponding background sound for the scene; it also places appropriate sound sources at the estimated object locations accordingly.

information is missing since the environment is treated as a diffuse audio source. We build upon these existing ideas and augment panorama images with spatial audios.

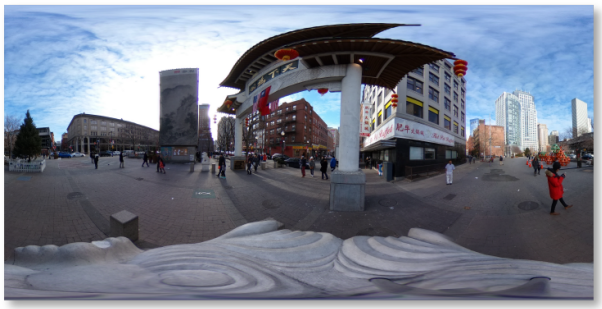
3 OVERVIEW

Figure 2 depicts an overview of our approach. Our method takes a 360° image as input. After scene classification, object detection, and action recognition, the image is labeled with what we will call here on out a background tag, which matches the type of scene in the image (for example, "Town"), and also the objects are labeled with object tags. Object tags either are object names or, if the object detected is a person, words for actions.

To start with, we built an audio file database. These files are organized into two types: background and object audio files. Each audio file is associated with an audio tag, which we set as the object tags for object audio files and scene tags for background audio files.

Our approach then estimates, with a single user interaction for inputting the depth of one object in the scene, the depth and hence the 3D location for each of the detected objects by using estimates of the real-life average height of objects and the relative height of objects recognized in the scene. Based on the detection and recognition results, our approach assigns appropriate audio sources at the calculated depth for each object by comparing the object tags with the audio tags in the database. If there is a match between the tags, we randomly select an audio file from our database that is labeled with the matching audio tag. These audio tags are effectively the different categories of sound that we have in our database.

For getting the audio file for the background audio of the scene, we use the same approach except that we use the scene tag instead of the object tags. The output of our approach is an audible panorama image, which can then be experienced using a virtual reality headset.



(a) A representative panorama image used for the scene before being rendered in our viewer.

(b) A sample image and the corresponding tags assigned by our automatic scene classifier. In this case, some scene tags assigned were "Crowd", "Transport", and "Town". "Town" was ultimately the highest scored and most frequently occurring tag across all tested image segments, so it was selected as the scene tag for the *Chinatown* scene.

Figure 3: An illustrative example, *Chinatown*.

4 APPROACH

Sound Database

To ensure that our sound database is comprehensive, we select the sounding objects based on 1,305 different panorama images found on the internet. By running scene classification and object detection and recognition on these panorama

Figure 4: To perform object detection, our approach samples images of the scene by rotating the camera horizontally by 36° each time until the whole scene is covered. Here we illustrate the process for *Chinatown*, with the bounding boxes in the enlarged image showing the detected objects.

images we were able to detect repeatedly occurring objects and scene types, which we also use as the corresponding object and scene tags. We then set the categories for all the sounds and build a database with background and single audio files. The audio files represent background sounds, human actions, sounds of motor vehicles, etc. Our sound database constitutes a total of 512 different audio sources, which are organized into two types: audio sources for objects and sounds for background ambient scenes:

- Object Sounds: There are 288 different object sounds, which include human chatting and walking, vehicle engine sounds, animal yelling and others in 23 categories. Each category, which matches previously mentioned object tags, is used for audio source assignment for objects in the scene. We normalize the volume of all sounds and generate versions of the audio files for various configurations.
- Background Audio: There are 224 different background audio files in total. We catalog these using 22 tags that match the different scene tags such as "City", "Room", "Sea", etc. These audio tags are used for selecting the audio source for the background based on the scene tag.

For a list of representative audio tags, refer to Table 1. A complete list can be found in our supplementary material. All audio files were obtained from the website "freesound.org" [6]. To generate 3D spatial audio from a 2D image, we estimate the relative depth of different objects in order to place the sounds reasonably in 3D space to simulate spatial audio.

Tag	Type	#	Tag	Type	#
City	Bg	14	Car	Obj	6
Crowd	Bg	12	Cat	Obj	2
Library	Bg	12	Chatting	Obj	17
Room	Bg	15	Dog	Obj	13
Sea	Bg	12	Walking	Obj	2

Table 1: A partial list of the audio tags used in our database. Each audio tag is a label for audio files in our database. So the "Chatting" tag, for example, tags different files for audio of people chatting. The type refers to audio files either being for background audio (*bg*) or object audio (*obj*). Refer to our supplementary materials for a complete list of all audio tags. Our database contains a total of 512 audio files.

Scene Classification

The goal of scene classification is to assign a scene tag to the scene, which matches one of the background audio tags in our audio database. Beginning with a panorama image and a camera viewing horizontally from the center of the rendered image, we rotate the viewpoint horizontally 36° to capture different segments of the image.

Our system also splits the captured samples vertically. If desired, the user may increase the number of samples adaptively to provide more input data for scene classification. Each segment is assigned a list of five scene categories, which we use as the scene tags, ranked in decreasing order of classification confidence scores.

We combine the classification results on all slices and select the most frequently-occurring, highest-scored tag as the tag for the overall scene for the panorama image. So, for example, that the two most common occurring scene tags for an image are "Library" and "Living Room", and the confidence score of each is 0.8 and 0.92 respectively, then "Living Room" will be selected as the scene tag. Refer to Figure 3(b) for an example.

Object Detection and Recognition

We use TensorFlow Object Detection, which is based on a Convolutional Neural Network (CNN) model pre-trained on the COCO dataset [15]. Our approach slices the panorama image the same as in scene classification, and we run object detection on each slice. If there are any overlapping objects

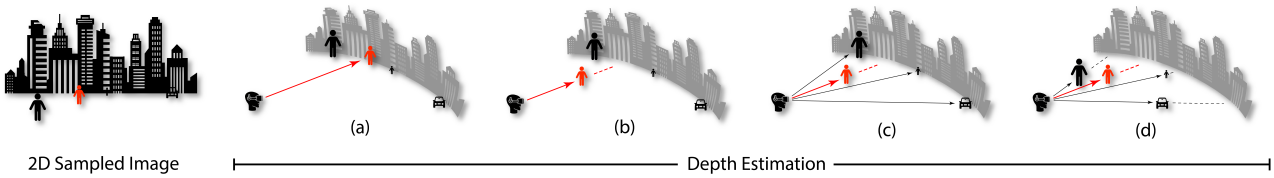


Figure 5: Depth estimation. We assume the average real-world heights of different objects categories based on estimation and previously recorded averages, which are shown in our supplementary materials. In this example, we show both people (average height: 5.3 ft) and a car (average height 5.0 ft). The person highlighted in red represents the reference object and the red line represents the baseline depth specified by the user. The depth estimation estimates the depth of the black objects by these inputs. (a) The system chooses the reference object, which corresponds to the object that received the highest confidence score during object detection. (b) The designer specifies the baseline depth to the reference object. (c) The system estimates the depths of the other objects. (d) audio sources corresponding to the objects will be placed at their estimated depths to create spatial audio.

from one slice to the next, we count the detected objects as the same object. Once objects have been detected, we use Google Vision for object recognition, feeding in cropped images of all detected objects to the object recognition algorithm. We assign object tags to the objects in the same way that we tag the scene. Figure 4 depicts the object detection process.

Depth Estimation

Our depth estimation technique requires comparing two objects detected on a panorama image. The first object is a reference object r which we use as a baseline to estimate the depths of all other detected objects in the image. The depth of the reference object, d_r, is set by the user. This is the only user interaction required during the depth estimation process. By default, the reference object is chosen as the object with the highest score (which corresponds to the highest confidence) of running object recognition on the image.

Our goal is to estimate the depth d_i for each object i detected on the panorama image. Let R_r and R_i be the estimated real-world heights of the reference object r and the detected object i respectively (e.g., the average height of a "person" is 5.3 ft [34], and that of a "car" is 5.0 ft).

The average heights for all objects in the database were either estimated by us or taken from real-world data. Please refer to our supplementary materials for a complete list of the average heights for all the object tags. Savva et al. offer a technique for automatically estimating the appropriate size of virtual objects [26], which is complementary to our approach. Let N_r be the normalized height of the reference object r (i.e., the object's height in the image divided by the image height) and N_i be the expected normalized height of the detected object i. By similar triangles, we have the following relationship:

$$\frac{N_i}{N_r} = \frac{R_i}{R_r} \tag{1}$$

Here, N_i is the only unknown variable, which we solve for. The final step is to calculate the estimated depth d_i. To do this, we compare the expected normalized height (N_i) with the actual normalized height (N_i') of the object in the image, whose depth we are trying to estimate. Then, by similar triangles, we obtain the estimated depth d_i of object i by:

$$d_i = \frac{N_i}{N_i'} d_r \tag{2}$$

This depth estimation process is applied for every detected object in the image. Figure 5 illustrates this process.

Audio Source Assignment

Once objects in the image have been detected and tagged, the next step is to assign an adequate audio source to each one. We accomplish this by comparing the 5 tags assigned to each object during object recognition to the tags in our database. We assign a sound if one of the tags matches a tag in the database. The tags of each object are compared to the tags in the database in order of highest to lowest confidence scores.

For objects in the image detected as persons, tags for actions are automatically assigned by the object recognition algorithm, so our approach handles action recognition for the assignable object tags for actions. In our approach, these are "Chatting", "Chatting on Phone", "Typing", and "Walking". Some object tags for actions recognized by object recognition including "Sitting" and "Standing" are ignored by our approach because they are not audible.

Audio Equalization Processing

As a post-processing step, all the assigned audio files can be equalized using Adobe Audition. This sound equalization is done in order to match the quality of the sound to fit either an indoor or outdoor environment according to the recognized scene type. In our implementation, we first normalized the volumes of the sounds from different sources

Figure 6: We run our algorithm on a dataset of 1,305 images, plus the 4 images (*Chinatown* which shown in Figure 1, *Seashore, Cafeteria, Living Room*) we took for the user study. Here we display some of the panorama images used. Refer to our supplementary material for more examples. "Bus" by Artem Svetlov/CC BY 2.0; "Museum", "Campus" and "Dock" by Kaleomokuokanalu Chock/CC BY-NC-SA 2.0; "Park" by tadayoshi527/CC BY-NC-ND 2.0;

before scaling them by distance, and applied an equalization matching algorithm to create indoor and outdoor versions for each sound [7].

5 EXPERIMENTS

Our experiments were conducted with a 3.3GHz Intel Core i7 processor, an NVIDIA Quadro M5000 GPU, and 32GB of RAM. To conduct the experiments, we created a 360° panorama image viewer with the Unity engine, which supports spatial audio. We release our code, data, and viewer for research purposes. The data includes 1,305 panorama images obtained from flickr.com and four images which we captured with a Ricoh Theta V 360° camera.

Sound Placement Results

We discuss the results of running our approach on 4 different panorama images, namely, *Chinatown*, *Seashore*, *Living Room*, and *Cafeteria*, which are depicted in Figure 1 and Figure 6. For audible versions of these results for the *Chinatown* scene, please refer to the supplementary video.

***Chinatown*:** Our approach automatically assigned an audio file that matched the "Town" tag as the background audio and identified many people. These people were assigned object tags like "Walking" and "Chatting". Cars were also detected and recognized, and the object tag "Car" was assigned to these objects. The database then assigned audio files for the objects that matched these tags.

The locations of the people and vehicles are easily discernible, with the background audio making the scene experienced in VR feel like an actual city center. The background audio "Town" matches sound that one could expect to hear in a town center including background vehicle sounds and construction sounds.

***Seashore*:** Our approach automatically assigned an audio file matching the "Sea" tag from our database as the background audio. Our approach also detected one person far off in the distance with a cellphone so an object tag of "Chatting on Phone" was assigned to that object. This object tag matches the audio tag "Chatting on Phone" in the database, so an audio file associated with that tag was randomly chosen.

Since the person is far off in the distance, the audio source was placed accordingly, which can barely be heard. We used this image to test results of using background audio with few object audio sources.

The background audio assigned consists of the sounds of waves reaching the shore. This mimics what one would expect to hear at a beach shore, as the sound of waves tends to drown out other sounds, especially in a scene like this, which is not crowded.

Living Room: The background sound assigned from our database matches the background tag "Room" and consists of quiet background noise, which mimics the background noises heard inside city apartments. Only a single audible object in the room was detected and recognized with the object tag "TV". By matching the object tag with the same audio tag in the database, we randomly selected an audio file with sounds from a television. In our case, music plays from the television. When viewing this scene with the Oculus Rift headset, it is immediately recognizable where and from what object the audio is coming from.

***Cafeteria*:** Our approach assigned an audio file matching the audio tag "Restaurant" as the background sound of this scene. It also identified people with the tag "Chatting", so audio files for chatting were also assigned.

The restaurant audio file used as background audio, combined with the placed audio files for people chatting, produces the effect of being in an indoor crowded space.

Set	Audio Configurations
1. Space	**spatial** & stereo & mono audio
2. Background	**with** & without background audio
3. Synthesized	recorded audio & **our synthesized**
4. Correctness	**correctly-** & incorrectly-detected objects
5. EQ	with & **without audio equalization**
6. Depth	**our depth** & uniform & random depth
7. No. of objects	10% & 50% & **100%** of all objects

Table 2: Audio configurations used in the user study. The bolded configuration is the standard configuration used in each set. The standard configuration is the result of running our approach on a particular scene without any modifications. We created these sets to investigate what features of the audio were important in delivering a realistic audio experience for the scenes.

Scene	Set						
	1	2	3	4	5	6	7
Chinatown	**0.011**	0.722	0.589	**< 0.001**	0.596	0.123	0.288
Cafeteria	**0.004**	0.221	**0.005**	**< 0.001**	0.782	0.148	0.186
Seashore	0.124	**< 0.001**	0.126	**< 0.001**	0.838	N/A	N/A
Living Room	**< 0.001**	1.000	0.055	**< 0.001**	0.914	N/A	N/A

Table 3: The p-value for each scene and set of audio configurations calculated from the user study data. The p-values smaller than 0.05, which reject the null hypothesis H_0, are bolded. We performed this statistical analysis to study which aspects of our system affect the user-perceived quality of the audio assigned in each case.

6 USER STUDY

To evaluate our approach, we conducted an IRB-approved user study with 30 participants. The users were aged 18 to 50, consisting of 17 males and 13 females, with no self-reported hearing or vision impairment. They were asked to wear an Oculus Rift headset with Oculus Touch controllers and headphones to view four panorama images. The audio was spatialized to stereo via Unity's built-in spatializer plugin (Oculus Spatializer HRTF). The 4 scenes are *Chinatown*, *Seashore*, *Cafeteria* and *Living Room*.

Study Design

The goal of the user study was to test how different characteristics of the audio assigned affected user ratings. To this end, we set out to test 7 sets of different audio configurations. The goal of this setup is to investigate which aspects of the synthesized audio had the largest effect on the subjectively perceived quality. Please refer to Table 2 for a description of the audio configurations included in each set.

Users were asked to view each scene while the audio was played with a certain audio configuration. For each set of audio configurations, the user experienced the same image under the different configurations belonging to that set. The 7 sets were tested in random order, with within-set audio configurations also being played in random order. The initial viewing angle of the scene was randomized before each scene and audio configuration were shown to avoid bias. Users experienced each scene and audio configuration combination once so that they could give a rating for each audio configuration under each set on each scene.

Rating Evaluation

Users rated each configuration using a 1-5 Likert scale, with 1 meaning that the audio did not match the scene at all and 5 meaning that the audio was realistic. The audio for each configuration played for approximately 10-20 seconds. We also calculate the p-value for each audio configuration comparison using the Analysis of Variance (RM-ANOVA) test for sets with 3 configurations and using the T-Test for sets with only 2 configurations. The tests were run independently for each scene. We chose Repeated Measures ANOVA and the T-Test since each participant did all configurations under each set. Note that the audio synthesized by our approach without any modification (referred as standard configuration) was included in each set. Any p-value below 0.05 indicates that we can reject the null hypothesis H_0, which refers to the situation that the average user ratings for the audio configurations in each set are about the same. So, whenever we reject H_0 for configurations in a set, it means that the difference in rating caused by the different configurations in that set is significant.

Results

Figure 7 shows a comparison of the different ratings given to each audio configuration by the users. The p-values calculated are shown in Table 3. Our supplementary materials contain the individual scores of each participant. We discuss the results for each set of audio configurations:

Set 1 (Space): For the space set, full spatial audio (the standard configuration) received the highest average score. If we look at the p-values for each scene, we see that they are all below the threshold of 0.05 for rejecting H_0, except for the *Seashore* scene. *Seashore* is assigned only background audio and one object audio for a person who is far away in the scene; the realism brought about by spatial audio may not be apparent or important for this scene. We can conclude that the spatial positioning of audio sources by our approach produces more realistic results than only using stereo or mono audio.

Set 2 (Background): For every scene except *Seashore*, configurations with background audio have scores about the same or slightly higher than the scores obtained by only including object sounds. The p-values for all scenes except *Seashore* are above 0.05. What we can conclude from this is that the effect of adding background audio is not significant when sufficient object audio sources are placed in the scene. This could be explained by the fact that for *Chinatown* and *Cafeteria* there are many distant, individual objects whose sounds may constitute a realistic background sound

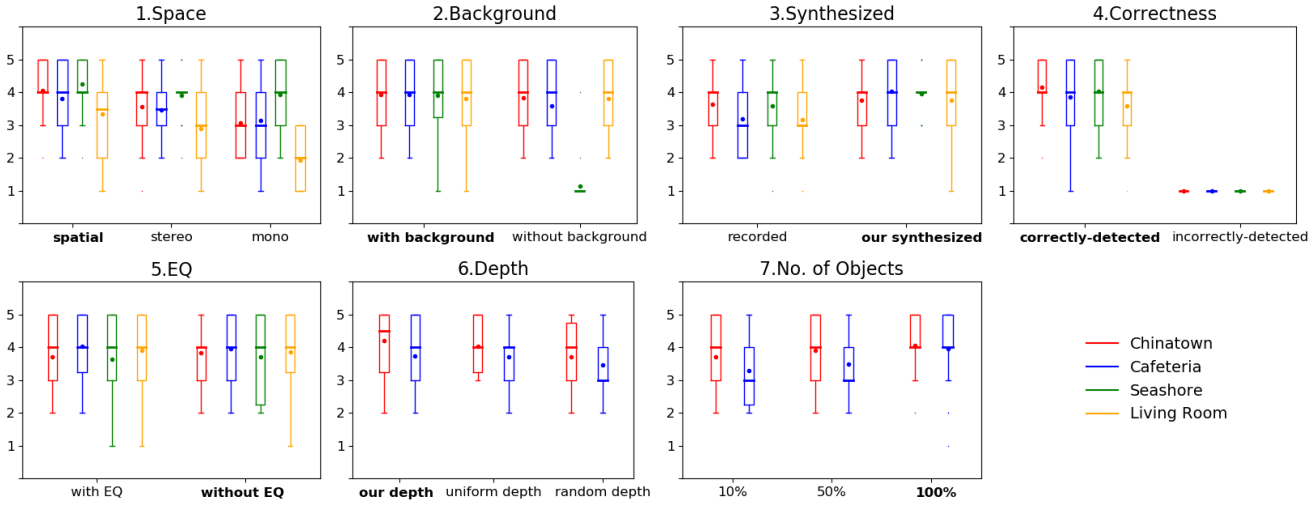


Figure 7: Results of the user study. Each plot corresponds to a set of audio configurations. The colored boxes represent the IQRs; the colored dots represent the means; the thick bars at the middle of the IQRs represent the medians; the upper and lower whiskers extend from the hinge by no further than 1.5* IQR, which show an approximately 95% confidence interval for comparing medians; and the bold texts represent the standard configurations. For each set, the average user rating under each audio configuration for each scene is shown. Note that for the *Seashore* and *Living Room* scenes, sets 6 (Depth) and 7 (Number of Objects) were not tested since each scene only has the background audio and audio for one object assigned.

when mixed together, while for *Living Room* the environment is supposed to be quiet. For *Seashore*, turning off the background audio renders the scene almost silent because all the expected background beach sounds (e.g., sea waves sound) are gone while only the sound of a single detected person chatting on a phone can be heard. In conclusion, we believe that the background audio's importance depends on the scene's setting, while using both background and object audios may provide a more realistic experience in some scenes.

Set 3 (Synthesized): For all the scenes, the configuration with our synthesized audio had a higher average score than that of the recorded audio. However, only the p-value for *Cafeteria* scene was significant (below the threshold of 0.05). We conclude that our synthesized audios are perceived to be at least as realistic as those recorded with a 360° camera.

Set 4 (Correctness): Audio placed for correctly recognized objects received higher average ratings across the board. Configurations produced by our approach with correct audio placement scored higher across all four scenes. The p-values were all below 0.05, so we can conclude that using audio files that match correctly with the objects of the scene is important.

Set 5 (EQ): Across all four scenes, the average scores for configurations with and without audio equalization were approximately the same and the p-values were all above 0.05. We can conclude that the effect brought about by sound equalization is negligible for our scenes at least.

Set 6 (Depth): On average, the configuration for audio placed at depths calculated by our approach scored higher than the configurations for uniform and random depths. While this is the case, the p-values for all scenes were above 0.05. We can conclude that while placing audio sources at proper depth may enhance realism, the effect is not significant in general. As shown in set 1, the positions of the sound sources are more important than their depths with regard to the realism perceived by the users. We conclude that while being able to recognize the depths of audible objects is important, there can be some flexibility in positioning the sound sources at their exact depths.

Set 7 (Number of Objects): On average, users preferred the configurations with all (i.e., 100%) audio sources used. However, the p-values were all above 0.05. We conclude that while using all detected audio sources produces more realistic results on average, the realism enhancement brought about by using all sources compared to using only some of the sources is not significant. In other words, having a few properly placed sound sources is enough to create the sense that the sound is realistic, even if other objects in the scene are not assigned audio files.

Post-hoc Tests: We also run pairwise post-hoc tests on sets 1, 6, and 7 for the four scenes, with a total of 24 post-hoc tests. For set 1 (space), there is significant difference (p-value smaller than 0.05) in ratings between the standard configuration (spatial) and the mono configuration in the *Living Room* scene. For set 6 (depth), there is significant difference in ratings between the standard configuration (using

our estimated depth) and the random depth configuration in the *Chinatown* scene. For set 7 (number of objects), there is significant difference in ratings between the standard configuration (using all detected objects) and configuration using 10% of the detected objects. Please refer to our supplementary documents for full results of the post-hoc tests.

User Feedback

Most users commented that they found the synthesized sounds to be realistic and immersive. However, some users commented that they found some audio sources unrealistic because they could not see moving objects. This was especially true in the *Chinatown* scene, where some users complained that they found the sound of cars unrealistic since no vehicles were moving. While this is a limitation posed by static panorama images, it does relate to the most common suggestion that users had on extending our approach to videos.

For the *Living Room* scene, some users stated that while the TV sound source was realistic, the audio for the sound source was too centralized. In other words, when turning their head away from the TV, they expected to hear more sound waves bouncing back across the room. We could explore incorporating sound waves bouncing off surfaces in our approach, which would require semantic understanding of what surfaces are in the scene.

Many users claimed that they could clearly tell the difference between the configurations that had audio sources positioned with respect to objects in the image and those that did not. Overall, most users were able to identify where audio sources were placed in the images. Most said that such 3D placement of sounds enhanced the scenes. Refer to our supplementary material for information on the frequency of certain types of comments made by users.

Discussion

Out of the 7 sets of audio configurations tested, audio placed by our approach with no modifications tested best in most experiments. From our user study, we see that our configuration received the same or highest average score among all audio configurations in each set. From the parameters we tested, the most relevant ones are spatial audio and correct objects. In comparison, audio equalization, accurate depth, and using all detected objects are not as important as the spatialness and correct detection. As for the background audio, our results show that its importance depends on the scene complexity and it can enhance the realism in some cases (e.g., *Seashore*).

One application of these findings is for prioritizing in power-sensitive or real-time computing, for example, mobile AR applications where certain computations can be conducted with a lower priority without significantly deteriorating overall user experience. We advocate an emphasis on a high-quality ambisonic audio engine and robust object detection algorithm. On the other hand, estimating accurate, high-resolution depth and performing audio equalization could be given a lower priority.

7 CONCLUSION AND FUTURE WORK

We presented Audible Panorama, an approach for automatically generating spatial audio for panorama images. We leveraged scene understanding, object detection, and action recognition to identify the scene and objects present. We also estimate the depths of different objects, allowing for realistic audio source placement at desired locations. User evaluations show that the spatial audio synthesized by our approach can bring realistic and immersive experience for viewing a panorama image in virtual reality. We are open-sourcing the audiovisual results we ran on Flickr panorama images, the viewer program, and the manually curated audible sound database.

Limitation and Future Work

Our current approach only applies to 360° panorama images. As an early attempt, we focus on panorama images which are abundant but usually lack an accompanying audio. A useful and natural extension would be to make our approach compatible with 360° videos with temporal consistency.

As with other data-driven synthesis approaches, one inherent issue with our approach is generalization. In our current study, only 4 scenes are comprehensively evaluated.

While we synthesized and release the audios for all the panorama images we collected on our project website, we did not conduct a full-scale evaluation on all the 1,305 results. One future work is to determine whether those results are as good as the 4 evaluated ones. Our approach may not perform well on panorama images with a) moving objects with dynamic sound; b) tiny objects too small to be detected (e.g., a bird); and 3) partially occluded objects that result in object detection failure. For example, while a user may expect a partially occluded car on a road to give car engine sounds, an object detection algorithm may fail to detect the car due to partial occlusion and hence no car engine sound is assigned by our approach. We are interested in developing a more systematic way of measuring audio quality and perceptual similarity, especially for immersive audiovisual contents.

The user feedback we received also hints that exploring how to synthesize sounds for moving objects could help improve our approach. Inferring semantic behavior from still images is inherently challenging due to the lack of temporal information which carries important object movement and action cues. However, as humans can infer the object motions on a single image in many cases, with the advancements of computer vision techniques, we believe it would be possible for computers to infer similarly, perhaps by leveraging a more sophisticated understanding of the scene context (e.g.,

a car near the curb rather than in the middle of the road is more likely to be parked and static) or by analyzing the subtle details (e.g., motion blur) on different regions of the image.

Another interesting extension is to apply our approach for panoramic cinemagraphs: still panorama images in which a minor and repeated movement occurs on a few objects, giving the illusion that the viewer is watching an animation. Our approach could be applied to assign sounds of this repeated motion to the moving objects on a cinemagraph.

We have created a sound database containing audio sources for many types of scenes and objects observed on common panorama images. One further augmentation is to devise sound synthesis algorithms that can synthesize novel audios adaptively based on observations from new images. Such synthesized audios may match with the scene even more closely and realistically, as well as introducing more natural variations. By releasing our results and working toward a more comprehensive research toolkit for spatial audio, we look forward to user experience enhancement in virtual reality.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their useful comments and suggestions. We would also like to thank the user study participants, and we are also thankful for the free audio files from freesound.org. The authors would also like to thank all the Flickr users for sharing their panorama images. This research project is supported by the National Science Foundation under award number 1565978.

REFERENCES

[1] Fathima Assilmia, Yun Suen Pai, Keiko Okawa, and Kai Kunze. 2017. IN360: A 360-Degree-Video Platform to Change Students Preconceived Notions on Their Career. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017, Extended Abstracts*. 2359–2365.

[2] Durand R Begault and Leonard J Trejo. 2000. 3-D sound for virtual reality and multimedia. (2000).

[3] Doug A Bowman and Ryan P McMahan. 2007. Virtual reality: how much immersion is enough? *Computer* 40, 7 (2007).

[4] Janki Dodiya and Vassil N. Alexandrov. 2007. Perspectives on Potential of Sound in Virtual Environments. *HAVE 2007. IEEE International Workshop on Haptic, Audio and Visual Environments and Games* (2007), 15–20.

[5] Daniel J. Finnegan, Eamonn O'Neill, and Michael J. Proulx. 2016. Compensating for Distance Compression in Audiovisual Virtual Environments Using Incongruence. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 200–212.

[6] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound Technical Demo. In *ACM International Conference on Multimedia (MM'13)*. ACM, ACM, Barcelona, Spain, 411–412.

[7] François G. Germain, Gautham J. Mysore, and Takako Fujioka. 2016. Equalization matching of speech recordings in real-world environments. In *IEEE ICASSP 2016*.

[8] Jan Gugenheimer, Dennis Wolf, Gabriel Haas, Sebastian Krebs, and Enrico Rukzio. 2016. SwiVRChair: A Motorized Swivel Chair to Nudge Users' Orientation for 360 Degree Storytelling in Virtual Reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 1996–2000.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Claudia Hendrix and Woodrow Barfield. 1996. The sense of presence within auditory virtual environments. *Presence: Teleoperators & Virtual Environments* 5, 3 (1996), 290–301.

[11] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. 2017. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 3296–3297.

[12] FZ. Kaghat, C. Le Prado, A. Damala, and P. Cubaud. 2009. Experimenting with Sound Immersion in an Arts and Crafts Museum. *In: Natkin S., Dupire J. (eds) Entertainment Computing - ICEC 2009*. 5709 (2009).

[13] Arun Kulshreshth and Joseph J. LaViola Jr. 2016. Dynamic Stereoscopic 3D Parameter Adjustment for Enhanced Depth Discrimination. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 177–187.

[14] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. 2018. Scene-Aware Audio for 360°Videos. *ACM Trans. Graph. (SIGGRAPH)* 37, 4 (2018).

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr DollÃąr. 2014. Microsoft COCO: Common Objects in Context. *eprint arXiv* 1405.0312 (2014).

[16] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell Me Where to Look: Investigating Ways for Assisting Focus in 360° Video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. 2535–2545.

[17] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. 2013. Summary statistics in auditory perception. *Nature neuroscience* 16, 4 (2013), 493–498.

[18] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-Supervised Generation of Spatial Audio for 360°Video. In *Neural Information Processing Systems (NeurIPS)*.

[19] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. CollaVR: Collaborative In-Headset Review for VR Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, UIST 2017, Quebec City, QC, Canada, October 22 - 25, 2017*. 267–277.

[20] James F O'Brien, Chen Shen, and Christine M Gatchalian. 2002. Synthesizing sounds from rigid-body simulations. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM, 175–181.

[21] Masashi Okada, Takao Onoye, and Wataru Kobayashi. 2012. A ray tracing simulation of sound diffraction based on the analytic secondary source model. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 9 (2012), 2448–2460.

[22] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, UIST 2017, Quebec City, QC, Canada, October 22 - 25, 2017*. 289–297.

[23] Vanessa C. Pope, Robert Dawes, Florian Schweiger, and Alia Sheikh. 2017. The Geometry of Storytelling: Theatrical Use of Space for 360-degree Videos and Virtual Reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. 4468–4478.

[24] Nikunj Raghuvanshi, John Snyder, Ravish Mehra, Ming Lin, and Naga Govindaraju. 2010. Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 68.

[25] Katja Rogers, Giovanni Ribeiro, Rina R. Wehbe, Michael Weber, and Lennart E. Nacke. 2018. Vanishing Importance: Studying Immersive Effects of Game Audio Perception on Player Experiences in Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*. 328.

[26] Manolis Savva, Angel X. Chang, Gilbert Bernstein, Christopher D. Manning, and Pat Hanrahan. 2014. On Being the Right Scale: Sizing Large Collections of 3D Models. In *SIGGRAPH Asia 2014 Workshop on Indoor Scene Understanding: Where Graphics meets Vision*.

[27] Eldon Schoop, James Smith, and Bjoern Hartmann. 2018. HindSight: Enhancing Spatial Awareness by Sonifying Detected Objects in Real-Time 360-Degree Video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018*. 143.

[28] K. Sharma, A. C. Kumar, and S. M. Bhandarkar. 2017. Action Recognition in Still Images Using Word Embeddings from Natural Language Descriptions. In *2017 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 58–66.

[29] Jonathan Steuer. 1992. Defining virtual reality: Dimensions determining telepresence. *Journal of communication* 42, 4 (1992), 73–93.

[30] Anthony Tang and Omid Fakourfar. 2017. Watching 360° Videos Together. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*. 4501–4506.

[31] Y. Tian, Y. Kong, Q. Ruan, G. An, and Y. Fu. 2017. Hierarchical and Spatio-Temporal Sparse Representation for Human Action Recognition. *IEEE Transactions on Image Processing* PP, 99 (2017), 1–1.

[32] James Traer and Josh H McDermott. 2016. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences* 113, 48 (2016), E7856–E7865.

[33] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1. IEEE, I–I.

[34] Peter M Visscher. 2008. Sizing up human height variation. *Nature genetics* 40, 5 (2008), 489.

[35] Jiulin Zhang and Xiaoqing Fu. 2015. The Influence of Background Music of Video Games on Immersion. *Journal of Psychology and Psychotherapy* 5, 191 (2015).

[36] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Dahua Lin, and Xiaoou Tang. 2017. Temporal Action Detection with Structured Segment Networks. *CoRR* abs/1704.06228 (2017).