Using Spatialized Audio to Improve Human Spatial Knowledge Acquisition in Virtual Reality

Seraphina Yong
Institute of Information Systems and Applications
National Tsing Hua University
Hsinchu, 30013 Taiwan
seraphinayong1002@gmail.com

Hao-Chuan Wang
Department of Computer Science
National Tsing Hua University
Hsinchu, 30013 Taiwan
haochuan@cs.nthu.edu.tw

ABSTRACT

Picture schematization, the construction of spatial object maps from images, has useful applications ranging from indoor exploration to augmented reality. Since using human spatial knowledge improves schematization, crowdsourcing workflows are introduced for extracting spatial information from pictures. As 360° pictures are now available in virtual reality (VR), crowdsourced 360° picture schematization also becomes essential. Yet, the vergence-accommodation conflict (VAC) in head-mounted displays (HMDs) causes inaccurate spatial perception in VR. We propose integration of spatial audio in VR as a cost-effective and intuitive feature to support spatial perception. This study indicates spatial audio cues in VR are naturally incorporated by humans to significantly improve human spatial knowledge accuracy and, subsequently, crowdsourcing schematization.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; Auditory feedback

KEYWORDS

Spatial memory; virtual reality; spatial audio; crowdsourcing; picture schematization[1]

1 INTRODUCTION

Picture schematization is a process that involves extracting essential spatial relations from the visual details of a scene by identifying the objects in it and illustrating their spatial locations and distances relative to one another on an abstract 2D map. Containing spatial structural information, the schematic maps created from this process can be used to support interior space exploration, automated guidance or object manipulation in augmented reality applications.

However, current methods in computer vision to automate picture schematization are often expensive and do not produce schematic maps of desired quality [3]. Crowd-based picture schematization has been proposed as a more reliable and cost-effective alternative to the above, where crowd workers infer spatial information from the pictures and create schematic maps, which are then merged into the picture's final schematic map [3].

Though this proposed crowdsourcing approach is feasible, human detection of depth information and spatial relations from pictures is still relatively inaccurate. Especially in the case of 360° picture schematization, which would be most feasibly performed in a VR setting via HMD, human spatial knowledge acquisition would then suffer from VAC, an inherent problem of HMDs hindering depth perception.

1.1 Depth perception in VR

The VAC problem decreases fusion accuracy of binocular imagery due to conflicting cues caused by compressing a 3D environment into a stereoscopic display, which decreases the accuracy of egocentric depth perception in VR [1], in turn affecting spatial knowledge accuracy. Saliency of spatial elements as reference cues is important for humans to construct spatial knowledge, since these cues are used as reference nodes to anchor the self in cognitive space [2]. But VAC causes unreliable visual cues if pursuing accurate spatial knowledge construction.

1.2 Audio as a Compensatory Cue

Past research on the effect of multi-sensory input (tactile, olfactory, and audio) in addition to visual cues on object memory in a virtual environment (VE) finds the inclusion of all three non-visual cues improved memory [4]. Amongst the three, spatial audio has lowest computation cost [4] and intuitively encodes spatial information. Audio events are also processed as salient cues. Research has been conducted on effect of spatial audio on presence in VR, but none clearly on its effect on mapping accuracy.

IUI'18 Companion, March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-5571-1/18/03.
https://doi.org/10.1145/3180308.3180360

If including spatial audio in VR improves spatial knowledge accuracy, its use in VR-based interfaces for 360° imagery schematization can improve schematization data quality. To provide the interface with spatial audio support, a surface mesh can be created from multiple image views [5] as a base for scene object-bound spatial audio events to be used with visual information by crowd workers to create a more accurate schematization.

In the following study, we investigated if spatialized audio effectively enhances VR spatial knowledge learning by having subjects explore indoor VR spaces with spatialized or non-spatialized audio cues, then perform spatial mapping tasks to measure spatial knowledge recall.

2 METHOD

A within-subjects experiment was conducted on 26 participants (13 males, 13 females) from ages 20-27 with two conditions (spatialized and non-spatialized audio) counterbalanced by task order and VE arrangement.

2.1 Experiment Material

Conditions were split between two Unity scenes with various pieces of furniture. Indoor setting was chosen for proper spatial audio rendering. These rooms had the same selection of furniture but different arrangement and wall color. In the 12 pieces of furniture, 7 were unique and 5 of the same chair. This was done to simplify the spatial mapping task, which was to sketch the locations of the five chairs in each room. A short audio recording with descriptive content was attached to each piece of furniture and audio source spatiality was manipulated using the Steam Audio SDK. Spatialized audio was rendered binaurally using head-related transfer functions, and non-spatialized audio egocentrically, with no distal or directional information; all audio was heard as centered. Hardware used were HTC Vive HMD and controllers.

2.2 Procedure

To help participants acclimate to VR before performing the tasks, they explored for five minutes the Google Earth VR demo. Each subject was then instructed to explore the first room (with one of the audio conditions) by selecting to play each audio source at least once and to take as much time needed to familiarize themselves with the space. They were told prior to exploration that they would be asked to map the locations of the five chairs in the room afterwards. After participants indicated they finished exploring, they were asked to sketch the chair locations on a provided map of that room, which included the location and footprint of all non-chair furniture. The same process was then repeated for the second room with the remaining condition.

2.3 Measures

No mapping time difference between the two conditions ($F<1$, n.s.). However, a significant difference in the predicted direction was observed between mean distance error for spatialized (M=0.69, SD=0.31) and non-spatialized (M=0.81, SD=0.43) audio conditions, $F[1,25]=4.22$, $p \leq .05$. Distance error means were obtained by averaging error measurements for each of the five chairs.

3 DISCUSSION AND FUTURE WORK

The results of the study indicate that spatial familiarization in VR with spatial audio cues results in higher spatial recall accuracy compared to non-spatial. Participants in the spatial audio condition demonstrated significantly lower recall error when mapping chair locations, and this is based on successful direct sensory integration of the audio cue as they were not told about the presence of spatial audio. Based on this result, incorporation of spatial audio into VR interfaces can be a cost-effective and seamless addition to support more accurate spatial knowledge acquisition in VR.

In future work, to clearly assess effectiveness of applying spatial audio to 360° imagery schematization, we wish to study the effect of spatial audio on spatial knowledge acquisition in a visually flat environment (panoramic image-based VR) to find if accuracy still improves even when coupled with poor visual spatial information.

4 ACKNOWLEDGEMENT

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 106-2633-E-002-001, 105-2628-E-007-004-MY2), National Taiwan University (NTU-106R104045), and Intel Corporation. We thank the anonymous reviewers for their valuable feedback on this work.

REFERENCES

[1] G. Kramida and A. Varshney. 2016. Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* 22, 7 (January 2016), 1912–1931. http://dx.doi.org/10.1109/tvcg.2015.2473855

[2] H. Couclelis, R. G. Golledge, N. Gale, and W. Tobler. 1987. Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology* 7, 2 (June 1987), 99–122. http://dx.doi.org/10.1016/s0272-4944(87)80020-8

[3] H. Rao. 2015. Towards a Crowd-based Picture Schematization System. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion* (IUI Companion '15). ACM, New York, NY, USA, 101-104.

[4] H. Q. Dinh, Neff Walker, Larry F. Hodges, Chang Song, and Akira Kobayashi. 1999. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. *Proceedings IEEE Virtual Reality* (March 1999), 222–228. http://dx.doi.org/10.1109/vr.1999.756955

[5] S. Fuhrmann, F. Langguth, N. Moehrle, M. Waechter, and M. Goesele. 2015. MVE-An image-based reconstruction environment. *Comput. Graph.* 53, PA (December 2015), 44-53. https://doi.org/10.1016/j.cag.2015.09.003