MUSIC SCOPE PAD

Masatoshi Hamanaka
RIKEN

ABSTRACT

We present Music Scope Pad, an application for efficiently discovering what one wants to view among many unknown music videos (MVs). Current video players only enable the user to view one MV at a time, so to explore what the user wants to view among many unknown MVs, they need to play each MV individually, which requires a large amount of manipulation. Our app features artificial-intelligence processing of video acquired with the device's front camera to detect the natural movements of the user's head and hands while listening to music, enabling the user to explore MVs. Ten MVs are played simultaneously on the iPad screen and through the spatial acoustics of the device. The user can then explore the MVs they want to view by moving their head left or right. The volume of each MV through the spatial acoustics is automatically adjusted so that the MVs closer to the center of the screen are louder. If the user then cups their hands around their ears, as if they were listening carefully, they can hear the MVs directly in front of them on the screen through the spatial acoustics with more emphasis. If the user keeps their focus on one MV for more than three seconds, that MV will be selected and only that video will be played from the beginning.

1. INTRODUCTION

Streaming music video (MV) services on the Internet have made it possible to view a vast number of MVs, but opportunities to discover unknown MVs are limited, and many users view only a limited number of MVs. Therefore, we developed an application called Music Scope Pad that enables users to efficiently discover MVs among many unknown MVs. Current search and recommendation applications make it more efficient for users to find the MVs they want [1–13]. They save time by previewing songs from a generated ranking list.

In contrast, Music Scope Pad makes it possible to select an MV from the many available without the need for screen touch or other visual manipulations by detecting natural movements when users are listening to music and focusing on a particular MV that they want to hear. Music Scope Pad provides a novel MV-selection interface with the following three functions.

Scoping function: This enables users to preview many MVs allocated in three-dimensional (3D) space by moving their heads left or right. The function enables them to landscape MVs, saving time in previewing them.
Focusing function: This highlights a particular MV that users want to view by having them cup their hands around their ears as if they are listening to something carefully.
Switching function: This seamlessly switches MVs in 3D space by user gestures that show the palm of the hand.

Previously reported headphones with sensors to detect the direction users are facing or the location of the head can improve the sense of musical presence and create a realistic impression but cannot highlight what users want to focus on [14–17]. However, it is difficult to clearly hear a particular musical source from many other sources with these headphones, including those that users may have preferred not to hear. Music spatialization systems [18] enable users to control the localization of each musical source through a graphical interface. However, it is difficult to control each source's location through such an interface.

We previously proposed Music Scope Headphones, which provided the aforementioned functions by using a digital compass that detects the orientation of the face and distance sensors that measure the distance between the hand and ear [19–22]. As face-to-face exhibitions have become difficult due to social distancing in the wake of COVID-19, we investigated implementing similar functions in Music Scope Pad so that a wider audience can experience our demonstration by using a tablet.

2. MUSIC SCOPE PAD

Music Scope Pad has two modes, video selection and video viewing. In the video-selection mode, multiple MVs are played back simultaneously in a virtual reality space depicted on the tablet screen (Fig. 1). In the video-viewing mode, one selected MV is played (Fig. 2). We developed Music Scope Pad to enable users to save time in previewing MVs on the basis of the following three policies.
Reduced operations: When selecting MVs with a computer, we generally have to play them individually with many mouse or screen touch operations, which interrupt the process of listening. To solve this, Music Scope Pad enables operations without mouse clicks or screen touch by detecting natural movements when people listen to music and using those movements to control a tablet.
Easy preview of many MVs: We wanted to increase the number of opportunities for encountering unfamiliar MVs in collections. However, the number of MVs that can be previewed is limited within a fixed amount of time. This is

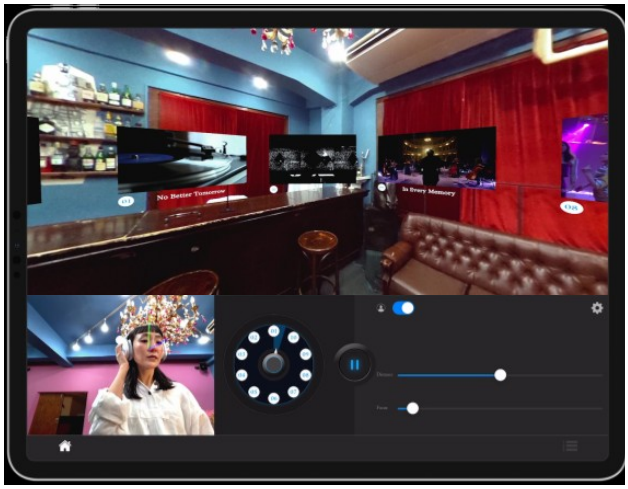

Figure 1. Video-selection mode

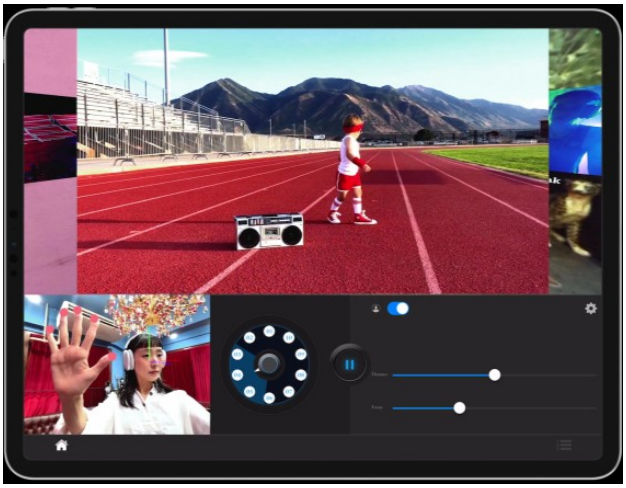

Figure 2. Video-viewing mode

because the larger the number of MVs to be previewed, the shorter the time to watch each MV. To solve this, Music Scope Pad playing many MVs at the same time.

Enjoy discovering MVs: Finding one's favorite MV among many is not an easy task. Therefore, Music Scope Pad uses a virtual 3D space for audio and visual, so that discovering MVs becomes a new form of entertainment.

We will explain the problems and solutions we encountered with Music Scope Pad on the basis of these policies.

How MVs are previewed: Music Scope Pad automatically adjusts sound volume of each MV so that the sound of one MV can be distinguished from those from others. That is, it increases that MV's volume while decreasing that of the others. Therefore, a user can easily preview a particular MV.

How motion is detected: The head direction is detected using artificial intelligence (AI) to analyze the image acquired with the front camera of the tablet. Since the orientation of the iPad is constantly changing, the angle information from the gyro sensor on the tablet and head-direction information acquired with the front camera are integrated to detect the user's head direction in real space. Similarly, the AI analyzes the image acquired with the tablet's front camera to detect the user's finger.

How function and motion are linked: How usable Music Scope Pad is depends on the quality of the links between the functions and the users' natural movements when watching MVs. Let us imagine the following scenario.

•We receive several MVs from a childhood friend.
•It sounds like the friend is playing a saxophone on one of these recordings.

In such a case, we would ordinarily search for MV sound with a saxophone part then might want to hear the saxophone playing more clearly. We use the following three links to achieve this.

Link for scoping function: The head direction detected by the AI analyzing the image from the front camera is linked to the head direction in the virtual 3D space of the spatial sound. Therefore, when users move their head left (right), the sound normally heard from the left (right) side can be heard from the frontal position. This enables users, through natural movements, to preview the MV sound they want to hear most clearly and hear it from the in front of them in virtual space.

Link for focusing function: The AI analyzing the image from the front camera detects the motion of users cupping their hands around an ear while they are listening to sound coming from the frontal position. The distance between the hand and ear determines the area in which sounds are audible. For example, when users place their hand close to their ear, they can only hear the MV sound from the frontal position (Bottom left of Fig. 1). When they remove their hand, they can hear all the MV sounds except those behind them. When they put their hand in the middle position, they can hear the MV sounds located in the front half position. By adjusting the distance between their hand and ear in this manner, they can control the focus level and highlight the MV sounds of interest.

Link for switching function: In the app's video-selection mode, we can search and preview 10 MVs allocated in 3D space from those listed in order with a music retrieval system. When converging on several MV sounds using the focusing the video-selection mode, users can delete unfocused sounds. If they want more convergence, they only need to adjust the focus level. When the user focuses on the sound of one MV for 3 s, the mode switches to video-viewing mode and that MV is played from the beginning. When the user shows their palm, it returns to the video-selection mode (Bottom left of Fig. 2).

3. CONCLUSION

We introduced Music Scope Pad, an application that enables users to quickly search and select from many music videos. Music Scope Pad and an introductory video explaining how to use it can be downloaded at `https://gttm.jp/hamanaka/en/musicscopepad/`

4. REFERENCES

[1] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.

[2] D. Zeng, Y. Yu, and K. Oyama, "Audio-visual embedding for cross-modal music video retrieval through supervised deep cca," in *2018 IEEE International Symposium on Multimedia (ISM)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2018, pp. 143–150.

[3] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-based video recommendation system based on stylistic visual features," *Journal on Data Semantics*, vol. 5, pp. 1–15, 06 2016.

[4] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, *Music Recommender Systems*. Boston, MA: Springer US, 2015, pp. 453–492.

[5] S. Oramas, O. Nieto, M. Sordo, and X. Serra, "A deep multimodal approach for cold-start music recommendation," in *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, ser. DLRS 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 32–37.

[6] S. Oramas, V. C. Ostuni, T. D. Noia, X. Serra, and E. D. Sciascio, "Sound and music recommendation with knowledge graphs," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, oct 2016.

[7] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Music video emotion classification using slow-fast audio-video network and unsupervised feature representation," *Scientific Reports*, vol. 11, 10 2021.

[8] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[9] E. Pampalk, "A Matlab Toolbox to Compute Music Similarity from Audio." in *Proceedings of the 5th International Conference on Music Information Retrieval*. Barcelona, Spain: ISMIR, Oct. 2004.

[10] F. Vignoli and S. Pauws, "A music retrieval system based on user-driven similarity and its evaluation," in *Proceedings of the 6th International Conference on Music Information Retrieval*. ISMIR, 2005, pp. 272–279.

[11] T. Jehan, P. Lamere, and B. Whitman, "Music retrieval from everything," ser. MIR '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 245–246.

[12] A. L. Uitdenbogerd and R. G. van Schyndel, "A review of factors affecting music recommender success," in *Proceedings of the 3rd International Conference on Music Information Retrieva*, 2002, p. 204–208.

[13] M. Goto and T. Goto, "Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces." in *Proceedings of the 6th International Conference on Music Information Retrieval*. London, United Kingdom: ISMIR, Sep. 2005, pp. 404–411.

[14] W. Oliver and E. Gerhard, "Listen - augmenting eeveryday environments through interactive soundscapes," in *Proceedings of IEEE Workshop on VR for public consumption, IEEE Virtural Reality*, 2004, pp. 268–275.

[15] J.-R. Wu, C.-D. Duh, M. Ouhyoung, and J.-T. Wu, "Head motion and latency compensation on localization of 3d sound in virtual reality," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '97. New York, NY, USA: Association for Computing Machinery, 1997, pp. 15–20.

[16] C. Goudeseune and H. Kaczmarski, "Composing outdoor augmented-reality sound environments," in *International Computer Music Conference*. International Computer Music Association, 2001, pp. 83–86.

[17] Y. Vazquez-Alvarez, I. Oakley, and S. Brewster, "Auditory display design for exploration in mobile audio-augmented," *Personal and Ubiquitous Computing*, vol. 16, pp. 1–13, 12 2011.

[18] F. Pachet and O. Delerue, "A mixed 2d/3d interface for music spatialization," in *Virtual Worlds*, J.-C. Heudin, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 298–307.

[19] M. Hamanaka, "Music scope headphones: Natural user interface for selection of music," in *Proceedings of ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006,*, 2006, pp. 302–307.

[20] M. Hamanaka and S. Lee, "Sound scope headphones," in *ACM SIGGRAPH 2009 Emerging Technologies*, ser. SIGGRAPH '09. New York, NY, USA: Association for Computing Machinery, 2009. [Online]. Available: https://doi.org/10.1145/1597956.1597977

[21] S. Miyashita, M. Hamanaka, and S. Lee, "Concert viewing headphones," ser. ACE '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 108–109. [Online]. Available: https://doi.org/10.1145/1971630.1971665

[22] M. Hamanaka and S. Lee, "Concert viewing headphones," in *SIGGRAPH Asia 2013 Emerging Technologies*, ser. SA '13. New York, NY, USA: Association for Computing Machinery, 2013. [Online]. Available: https://doi.org/10.1145/2542284.2542288