The Visual Computer
https://doi.org/10.1007/s00371-020-01918-1

ORIGINAL ARTICLE

Audio–visual object removal in 360-degree videos

Ryo Shimamura[1] · Qi Feng[1] · Yuki Koyama[2] · Takayuki Nakatsuka[1] · Satoru Fukayama[2] · Masahiro Hamasaki[2] · Masataka Goto[2] · Shigeo Morishima[3]

Abstract

We present a novel concept *audio–visual object removal* in 360-degree videos, in which a target object in a 360-degree video is removed in both the visual and auditory domains synchronously. Previous methods have solely focused on the visual aspect of object removal using video inpainting techniques, resulting in videos with unreasonable remaining sounds corresponding to the removed objects. We propose a solution which incorporates direction acquired during the video inpainting process into the audio removal process. More specifically, our method identifies the sound corresponding to the visually tracked target object and then synthesizes a three-dimensional sound field by subtracting the identified sound from the input 360-degree video. We conducted a user study showing that our multi-modal object removal supporting both visual and auditory domains could significantly improve the virtual reality experience, and our method could generate sufficiently synchronous, natural and satisfactory 360-degree videos.

Keywords Audio–visual object removal · 360-degree video · Human perception · Signal processing · Virtual reality

1 Introduction

360-degree videos, or spherical panoramic videos, have become popular among end-users thanks to consumer-level 360-degree cameras [9,19] as well as video-sharing platforms that support 360-degree videos [4,7]. With the increasing popularity of 360-degree videos, a new problem has appeared: unlike traditional cameras with *framing* capabilities, it is challenging for 360-degree cameras to capture only focused subjects, and thus, the captured videos often include unwanted objects such as passing cars. We thus consider that the demand for removing unintentionally included objects from original videos, namely, *object removal* from 360-degree videos, has been increasing.

Removing specific targets from an image is a well studied topic within the field of computer vision and computer graphics. *Image inpainting* can be used to remove a specified target object visually from an original image and replace it with the appearance of something else that fits the image (*e.g.,* background) [1,2,11,23,28]. Recently, several methods have been proposed that are capable of handling video clips rather than just still images [25,26]. Nevertheless, to our knowledge, only few papers [17,23] have discussed 360-degree videos, and, more importantly, none of these methods take auditory information into consideration. It is crucial that the visual and auditory information remain well-synchronized even after the object removal process; otherwise, a disparity between the visual and auditory domains could provide unnatural experiences.

We propose a novel concept of *audio–visual object removal* in 360-degree videos, in which a user-specified object in the target 360-degree video is removed in both the visual and auditory domains synchronously (see Fig. 1). The key idea is to effectively incorporate information acquired from the video inpainting process into the audio removal process. This multi-modal approach can reduce mismatches between visual and auditory domains, and we expect that

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00371-020-01918-1) contains supplementary material, which is available to authorized users.

✉ Ryo Shimamura
s-ryo@akane.waseda.jp

Shigeo Morishima
shigeo@waseda.jp

1 Waseda University, Tokyo, Japan

2 National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

3 Waseda Research Institute for Science and Engineering, Tokyo, Japan

Published online: 31 July 2020

Fig. 1 Concept of *audio–visual object removal* in 360-degree videos. **a** For a 360-degree video with sounds in the standard 4-channel format, the user specifies a target object, that will be removed, in the initial frame. In this case, a skateboarding man was chosen. **b** The target object is visually tracked across all frames, so that the sound generated by the object is identified. **c** The target object is removed in both, the visual and audio domains

it provides better viewer experiences of the resultant edited videos compared with when the object is removed only in the visual domain.

To achieve this concept, we present a method consisting of two sequential subprocesses: the visual cue removal and the auditory cue removal. During the visual cue removal process, our method tracks the location of the target object in the input video, crops a small area containing the tracked object, and then removes the object by applying an existing video inpainting technique [10] while taking care to keep temporal consistency. During the audio removal process, our method estimates directions of the sound sources in the spatial audio data of the target video using a technique called *Directional Audio Coding* (DirAC) [18], identifies the sounds generated by the target object, and then synthesizes a spatial audio data by subtracting the identified sounds. This target sound identification is enabled by the directional hint acquired in the video cue removal process. Finally, our method combines the processed visual and auditory information to produce a resulting 360-degree video, where a synchronized audio–visual object removal is achieved.

To validate this concept, we captured multiple test scenes in varied conditions and then conducted a user study using these test scenes and our implementation. The result showed that the proposed multi-modal audio–visual removal method could offer satisfactory 360-degree videos for the test scenes. It also indicated that the multi-modal approach could offer better experiences than single-modal (*i.e.,* visual- and audio-only) ones.

To summarize, our contribution is two-fold:

- We propose a novel concept of audio–visual object removal in 360-degree videos and a method to implement this concept. This is useful for users to remove unintentionally included objects with sounds from 360-degree videos.
- We validated our concept through a user study that our audio–visual removal method is superior to single-domain removal methods in terms of perceived synchronization, naturalness and satisfaction.

2 Related work

2.1 360-Degree videos

As 360-degree cameras capture their surroundings omnidirectionally, some projections are required to store the visual information in traditional 2-dimensional video formats. The *equirectangular* projection is the most used; however, there are two difficulties when one applies typical image/video processing techniques to the equirectangular format. First, it inevitably results in severe distortion [23,27]. Second, stepping over the boundaries of an equirectangular projection may lead to unexpected output [29]. Our method is designed to handle these issues properly.

Special care must be taken with the audio as well as the visual data. Many 360-degree cameras (*e.g.,* [19]) capture and store the surrounding soundscape as *spatial audio* for immersive experiences. In practice, this is implemented using the *Ambisonics* framework [6], in which the audio is stored as a 4-channel audio data format, called *B-format*, and then decoded for each speaker setting in play time. Our method works with this format and thus is compatible with standard 360-degree cameras.

2.2 Image and video inpainting

Image inpainting is a technique for filling a "hole" on a target image with a plausible and natural appearance, and it can be used for object removal for images. Many methods have been proposed, including recent learning-based [1,11,23,28], partial differential equation (PDE)-based [3], exemplar-based [13], and wavelet transform-based [2] methods. However, the distortion caused by equirectangular projection is prone to poor results when directly applying these methods proposed

for flat images. To cope with this issue for 360-degree images, Upenik et al. [23] presented the approach of first undistorting the area of interest, applying inpainting techniques, and then projecting it back to the equirectangular representation. However, their method requires manually specifying pixel-perfect masks of the target object and also does not consider frame consistency when applied to videos, which may produce flickering artifacts. We take a similar approach, for *video* inpainting by incorporating *SiamMask* [25], which is an optical flow-based tracking algorithm. Kim et al. [10] proposed a learning-based video inpainting method that maintains both long-term and short-term consistency; however, their method only works with traditional 2-dimensional videos. We combined Upenik et al. 's method and Kim et al. 's method to remove a specified target object visually in 360-degree videos.

2.3 Sound localization and separation

There are several approaches to separate sounds originating from different sources, including beamforming-based [22], learning-based [5,8,14,15,20], and parametric-spatial-sound-processing-based [12] approaches.

By taking the beamforming-based approach, Ruochen et al. [21] realized an application called *audio zooming* with B-format data. With multiple mono-microphone arrays, Nair et al. [16] extended this idea to an audio–visual zooming application that enhances the sound of a target object. The idea of using both visual and auditory information is similar to ours, but the purpose and necessary techniques are totally different; they focused on enhancing the sound from a single narrow area, not removing the specified sound completely.

Learning-based methods [5,8,14,15,20] have recently been successful even for challenging scenes such as two lecturers talking at similar frequencies. Morgado et al. [14] proposed a method that separates individual sound sources and localizes them in a 360-degree video. Yet, their applicability is greatly restricted by the expensive cost of acquiring a large amount of training data, and they cannot work well when the training data do not contain scenes sufficiently similar to the target scene.

DirAC [18,24] is one of the parametric-spatial-sound-processing-based methods. This method separates the target B-format data into several frequency bins with Short-Time Fourier Transform (STFT) and a mel filter bank, and then estimates parameters of the sound direction and sound diffuseness for every frequency bin. Thus, this parametric representation has the potential to be used for removing specific auditory information if relevant parameters can be appropriately identified.

Our audio removal process is built on the DirAC method and utilizes the directional hints acquired from the video inpainting process to identify the frequency bins that match the target object direction and then remove the relevant auditory information.

3 Method

This section describes the proposed method for an audio–visual object removal in 360-degree videos. The input consists of a 360-degree video with 4-channel audio (*i.e.,* B-format) and a bounding box selection of the target object in the initial frame. As shown in Fig. 2, our method consists of two subprocesses: 360-degree video inpainting and direction-based audio removal. To achieve a satisfying video inpainting quality for equirectangular videos, our method elaborately avoids distortions and inconsistencies across video boundaries. With the help of the extracted accurate and robust visual cues, it then identifies and removes the audio of the target object from the original video.

3.1 360-Degree video inpainting

Directly applying existing video inpainting techniques to equirectangular videos causes two specific problems. First, compared to traditional 2-dimensional images, severe distortions caused by equirectangular projections will lead to poor quality when directly applying 2-dimensional inpainting techniques. Second, naively splitting 360-degree images into 2-dimensional images will lead to unconnected boundaries (see Fig. 3), which brings a wrong result in the object tracking task. To solve the distortion problem in image inpainting, our method crops an area of interest that contains the target object, projects the cropped patch back to the traditional 2-dimensional image, applies existing inpainting techniques, and then restores the patch back to the area of interest. Even though a similar idea is shared in [23], our method achieves an automatic distorting process through all frames instead of manually creating a distorting mask. To automatically create such masks, our solution for boundary inconsistency is essential, which is described as follows.

At first, once the target to be removed is specified as a bounding box in the initial frame ($t = 0$), a mask of the target at the initial frame is generated by SiamMask [25]. Then, our method horizontally "rotates" the equirectangular panoramic image at the next frame ($t = 1$), so that the center of mass of the mask at $t = 0$ is located at the center of the rotated image at $t = 1$. The rotated image is then used to generate the mask of that frame ($t = 1$), and the generated mask is then used to rotate the image at the next frame ($t = 2$). Similarly, our method rotates every following frame in an accumulative way to generate masks, by which our method prevents the target object from moving across boundaries and thus robustly tracks the target object. This process is formally described as follows. The equirectangular image at

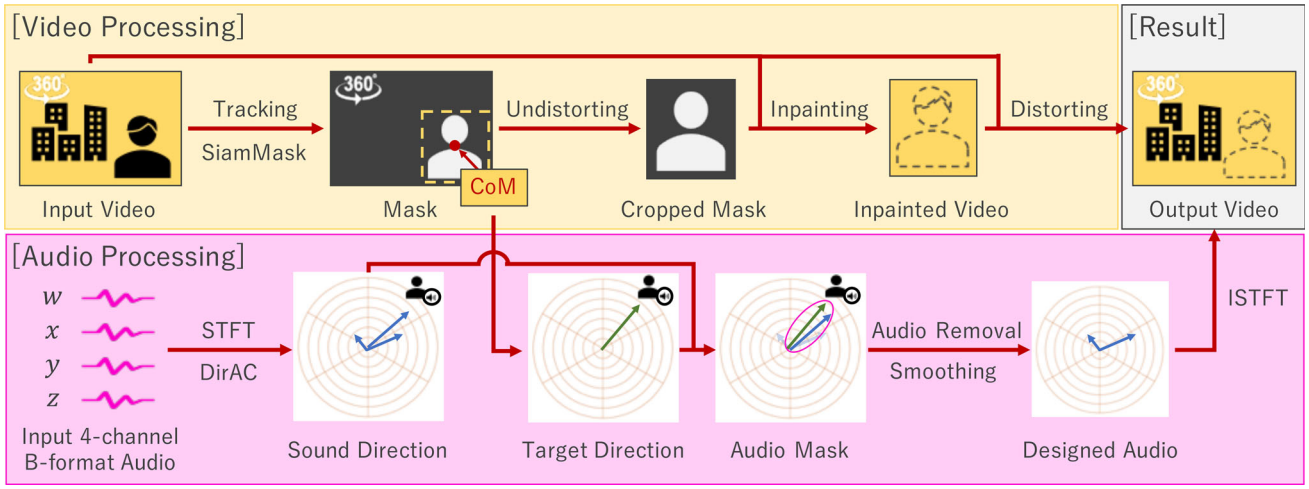

Fig. 2 Workflow of our method. In video processing, we begin by tracking the center of mass (CoM) of the target using SiamMask [25]. Then, the area around the target is cropped. For this cropped video, *Deep Video Inpainting* [10] is then performed to remove the object. The inpainted video is then returned to the original place. In audio processing, the 4-channel audio data are first separated into several frequency bins. With *DirAC* [18,24], the direction of the sound pressure of each frequency bin is estimated. By using the direction obtained during the video processing, the frequency bins that should be removed are identified, and then an audio removal mask is created. Then, this mask is smoothed to avoid noisy outputs. Finally, with this audio removal mask, an object-removed 4-channel audio data are synthesized, and the audio data are transformed back into time domains by inverse STFT (ISTFT)

Fig. 3 An example where special care is necessary in the perspective of boundaries. The removed target (circled by a red line) steps over the boundary of an equirectangular image (orange lines). In this case, visual tracking does not work well

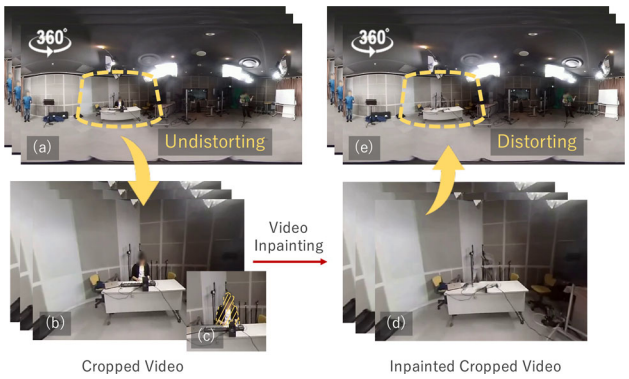

Fig. 4 Video inpainting flow. **a** An area surrounding the target is cropped and undistorted through all frames, and then a cropped video is created. **b** The cropped video is inpainted by a video inpainting method. **c** A mask area in a specific frame. **d** The result of the inpainting of the cropped video. **e** The inpainted cropped video is substituted back to its original place

the t-th frame, I_t, is horizontally rotated to synthesize an equirectangular image, I_t^{rot},

$$I_t^{\mathrm{rot}} = \mathrm{Rotate}\left(I_t, \frac{H}{2} - c_{t-1}\right), \quad (t = 1, 2, \ldots), \tag{1}$$

where H is the width of the equirectangular image, $c_{t-1} \in [0, H-1]$ is the horizontal position of the center of mass of the mask at the $(t-1)$-th (before-rotated) image, and $\mathrm{Rotate}(I, \Delta)$ is an operator that synthesizes an image by rotating the input image I by Δ (where $\Delta = H$ means a 360-degree horizontal rotation). Our method then generates the mask at the t-th frame using I_t^{rot}, instead of I_t.

Using the automatically generated masks across all frames, our method then performs video inpainting while considering the distortion issue to achieve a high-quality output (see Fig. 4). To handle the distortion caused by equirectangular projections, by tracking the center of the masks through all frames, an area of interest based on the maximum widths and heights of the masks is cropped (Fig. 5). hFOV (the field of view on the horizontal axis) and vFOV (the field of view on the vertical axis) of the cropped patch are calculated by

$$\mathrm{hFOV} = \frac{h}{H} \times 360^\circ, \tag{2}$$

$$\mathrm{vFOV} = \frac{v}{V} \times 180^\circ, \tag{3}$$

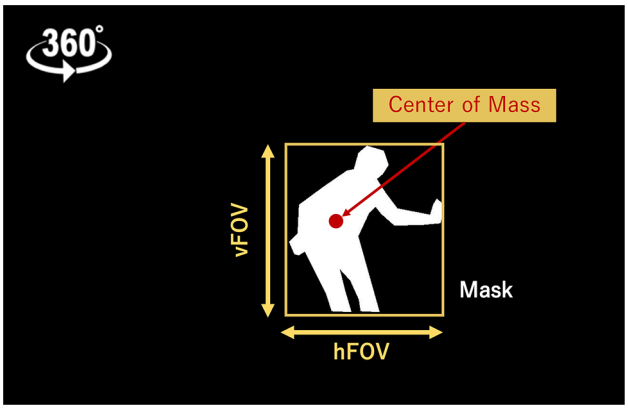

Fig. 5 The cropping area is decided using the center of mass and the size of the mask. With the center of mass, the center point of the cropping area is decided. With this size, the hFOV (field of view (FOV) on the horizontal axis) and vFOV (FOV on the vertical axis), which is used for undistorting the cropping area (see Fig. 4) are determined

respectively, where h and v are the maximum width and height of the masks in all frames, and V denotes the height of the equirectangular image. The center of the mask represents the directional information with two parameters: θ (elevation-degree), ϕ (azimuth-degree). Using the tracking information, we calculated undistorted images for the cropped area around the target mask to be able to apply the 2-dimensional video inpainting method. After the inpainting process, the processed patch is then placed back to the original location in the equirectangular images.

It is worth mentioning that although visual artifacts caused by inpainting techniques may not be as noticeable in traditional 2D formats, 360-degree video with its nature of sharing a limited resolution over all directions, are more prone to such a problem. Although it is beyond the scope of this work, we believe the quality of our method can be greatly improved with a better image inpainting algorithm.

3.2 Direction-based sound removal

To identify and separate the target audio source is a challenging task. In the case of audio input only, it is difficult to identify the sound emitted by the target object. The proposed method solves this problem by using the visual cues in combination with 4-channel microphone data.

In the Ambisonics framework [6], 4-channel microphone data (called A-format) are converted into four directional components based on spherical harmonics (called B-format). B-format data are represented as a tuple of four signals, (w, x, y, z), where w is the omni-directional sound pressure, and (x, y, z) are the left-right, front-back, and up-down sound pressure gradients, respectively. Given a B-format data, our method first performs the DirAC method [18] to obtain a parametric representation of the spatial audio. Specifically, DirAC separates a B-format audio data into several frequency bins using STFT and mel filter bank, and then estimates the primary sound direction of each bin by referring to the instantaneous intensity vector $\boldsymbol{I}_{k,n} \in \mathbb{R}^3$ as follows.

$$\boldsymbol{I}_{k,n} = \frac{1}{2}\,\mathrm{Re}\left(P_{k,n}\boldsymbol{U}^*_{k,n}\right), \tag{4}$$

$$P_{k,n} = w_{k,n}, \tag{5}$$

$$\boldsymbol{U}_{k,n} = \frac{1}{\sqrt{2}Z_0}\left(x_{k,n}\boldsymbol{e}_x + y_{k,n}\boldsymbol{e}_y + z_{k,n}\boldsymbol{e}_z\right), \tag{6}$$

where k is the index of the frequency bins, n is the index of the time steps, and Z_0 is the acoustic impedance. The complex values $w_{k,n}$, $x_{k,n}$, $y_{k,n}$, and $z_{k,n}$ denote the results of applying STFT to the signals w, x, y, and z in the B-format, respectively. Moreover, $\boldsymbol{e}_i \in \mathbb{R}^3$ represents the unit vector in Cartesian coordinates, $\mathrm{Re}(\cdot)$ is the real part of the complex value, and $(\cdot)^*$ is its complex conjugate. Refer to [18] for details.

For an accurate estimation of the target direction, the tracking result in the video inpainting process is available. Comparing the directions which are based on the center of mass from the video processing and the ones estimated by DirAC, an audio-removal mask $\boldsymbol{A}$ is calculated via the following steps:

$$\boldsymbol{A}_{k,n} = \begin{cases} 0, & \text{if } \cos^{-1}\left(\frac{\boldsymbol{r}(\theta_n,\phi_n)\cdot\boldsymbol{I}_{k,n}}{\|\boldsymbol{I}_{k,n}\|}\right) < \psi, \\ 1, & \text{otherwise}, \end{cases} \tag{7}$$

where ψ is the threshold of the interior angle between the visual and audio directions. θ_n and ϕ_n are the elevation and azimuth, which indicate the estimated direction from the video processing at time step n and $\boldsymbol{r}(\theta_n, \phi_n)$ is a directional unit vector in the direction indicated by θ_n and ϕ_n. To obviate the artifact caused by instant changes between $\{0, 1\}$, the smoothed mask $\boldsymbol{A}^{\mathrm{smooth}}_{k,n}$ further smooths the acquired binary masks with an average filter on the time axis as follows:

$$\boldsymbol{A}^{\mathrm{smooth}}_{k,n} = \frac{1}{s}\sum_{m=-\lfloor\frac{s}{2}\rfloor}^{\lfloor\frac{s}{2}\rfloor} \boldsymbol{A}_{k,n+m}, \tag{8}$$

where s is the size of the window (here we assume s is an odd number). After multiplying the frequency bins with the smoothed mask, the B-format audio data, which were synthesized, are finally transformed back into time domains by ISTFT.

4 Experiments

In this section, we describe subjective studies conducted for validating the concept of audio–visual object removal in

Table 1 Scenes for the experiment

Scene	Audio sources	Removed target	Movement	Frequency	Field	Ambience
1	Speaker/skateboarder	Skateboarder	**Dynamic**	Almost static	Outdoor	On (mainly birdsong)
2	Piano/maracas/violin	Piano	Static	**Dynamic**	Indoor	Off
3	**Speaker/speaker**	Speaker	Dynamic	Almost static	Outdoor	On (mainly birdsong)

Fig. 6 Video processing results of a certain frame of scenes in the main study. Left: scene 1, Middle: scene 2, Right: scene 3. Up: original input, Down: visual removal output

360-degree videos. More specifically, our goal is to evaluate the experience of object-removed 360-degree videos edited in the visual and auditory domains synchronously by our method compared to those edited in either the visual or auditory domain only.

4.1 Apparatus and scenarios

We captured multiple scenes under varied conditions with a 360-degree camera equipped with a 4-channel microphone. Specifically, we prepared three scenes to validate the concept in different conditions: (1) a person plays a skateboard around a speaker; (2) people play different musical instruments; and (3) a man interrupts another one's talk with his voice. Details of each scene are described in Table 1. We chose a variety of scenes in the perspectives of the change of the target's position, the change of the target's frequency, and the domain of audio sources. The original input scenes and audio–visual removal output scenes are demonstrated in Figs. 6 and 7. As highlighted in Table 1, the choice of each experimental condition was motivated by the following reasons: scene 1 investigates the effectiveness of the proposed method against dynamically moving objects, scene 2 investigates the capability of removing audio sources with a wide range of frequencies, such as musical instruments, and scene 3 investigates the performance of removing sounds from the same auditory domain. Each video clip lasts approximately 15 s.

For capturing these video clips, we used RICOH THETA V (30 frames-per-second in Full HD) as the 360-degree camera and RICOH TA-1 (4 channels with a 48 kHz sampling rate) as the microphone. In the experiments, we used a desktop computer with Intel i9-9900K and NVIDIA GeForce RTX 2080Ti and HTC Vive Pro (refresh rate: 90 Hz; FOV: 110°; resolution: 1440 × 1600) with its controller.

4.2 Pilot study

Before the main study, we conducted a pilot study. The purpose of this pilot study was to empirically select appropriate parameters in the audio removal processing. The parameters to be determined were the number of frequency bins, the threshold of the interior angle between the visual and auditory directions of a target object, and the size of a smoothing window. In this study, the combination of parameters was empirically selected: {50, 100} for the number of frequency bins, {20°, 40°, 60°} for the threshold of the interior angle, and {1, 3} for the smoothing window size. We recruited 7 participants, ages 22–51 (Mean = 27.7) without visual and aural disorders in the pilot study. We asked each participant to rate the top three audio clips in terms of their naturalness out of $2 \times 3 \times 2 = 12$ candidates generated using different combinations of parameters for each scene. Based on these votes, we selected the parameter combinations to be used in the main study (see Table 2).

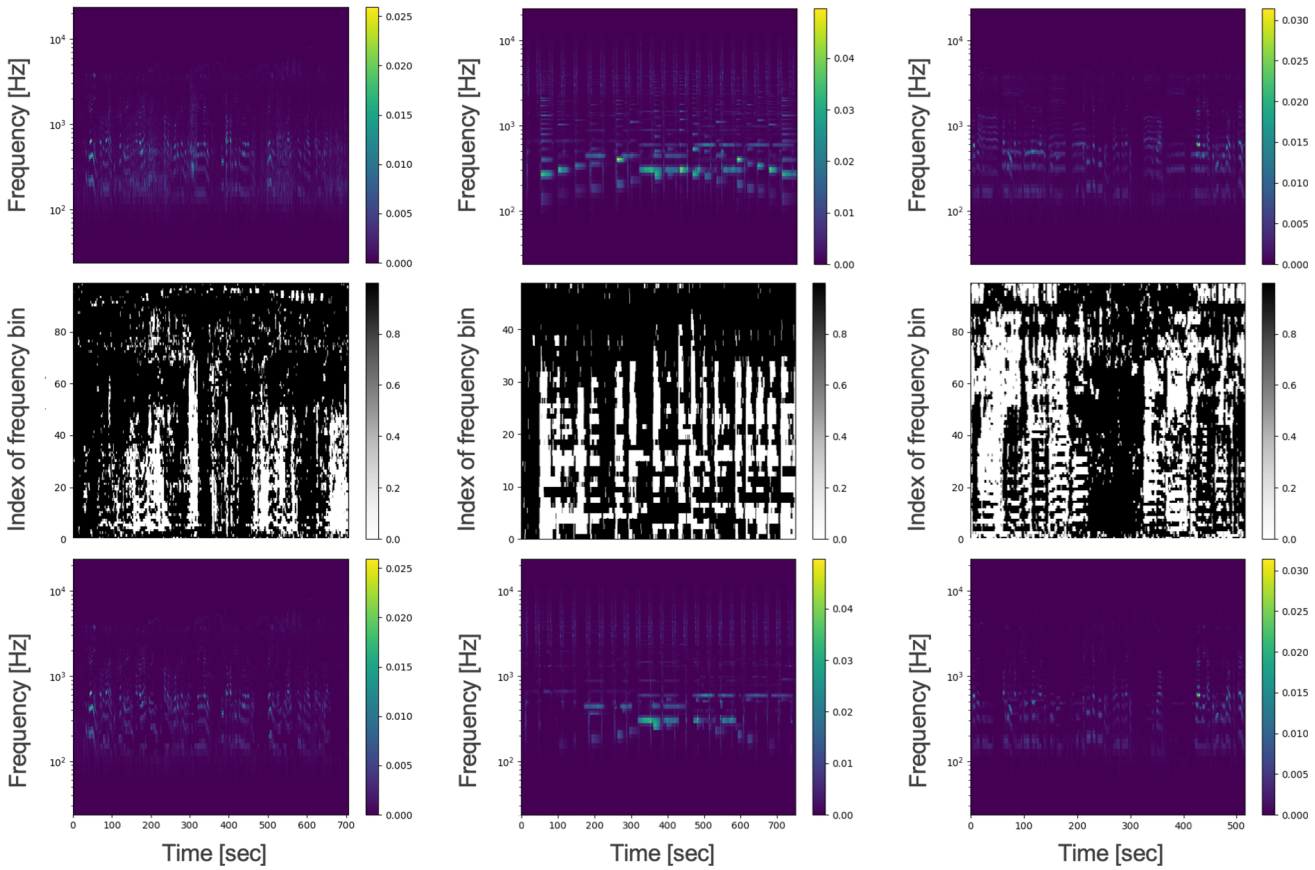

Fig. 7 Audio processing results of omni-directional channel data w as a representative of B-format in each scene in the main study. Left: scene 1, Middle: scene 2, Right: scene 3. Up: original input, Middle: audio removal mask, Down: audio removal output. The viridis color bar expresses the STFT magnitude. The gray color bar stands for the binary value of the smoothed audio removal mask (0: identified frequency bins, 1: non-target's frequency bins)

4.3 Main study

4.3.1 Goal

The specific goal of the main study is to evaluate whether the audio–visual object removal in a 360-degree video is preferable to single-domain (*i.e.,* either auditory or visual) removals through subjective user evaluation. Our hypothesis is that audio–visual removal is the superior editing method compared with one-domain removal from the perspectives of synchronization, naturalness, and satisfaction.

4.3.2 Experimental design

We recruited 17 new participants, ages 21–28 (Mean = 23.2) in the main user study. All of them have no visual and aural disorders. Twelve of them had experienced virtual reality technology before this experiment. Before the experiment started, we briefly explained the tasks to the participants and also how to use the controller. We allowed the participants to take a break whenever they wanted to do so.

Table 2 The results of the pilot study

Scene	Frequency bins	Directiony threshold	Smoothing size
1	100	20°	3
2	50	40°	3
3	100	60°	3

These are the best sets of parameters for each scene

We compared three approaches that remove a target (a) visually, (b) aurally, and (c) both of them synchronously from 360-degree videos. The order of the trials (a)–(c) and the order of the scenes was randomized to cancel bias in the results. Note that we showed the participants the original video clips without any editing before showing the edited videos, to simulate the real usage of editing the target object out. For each trial, the participants were allowed to replay the video as often as they wished to do so. After completing all

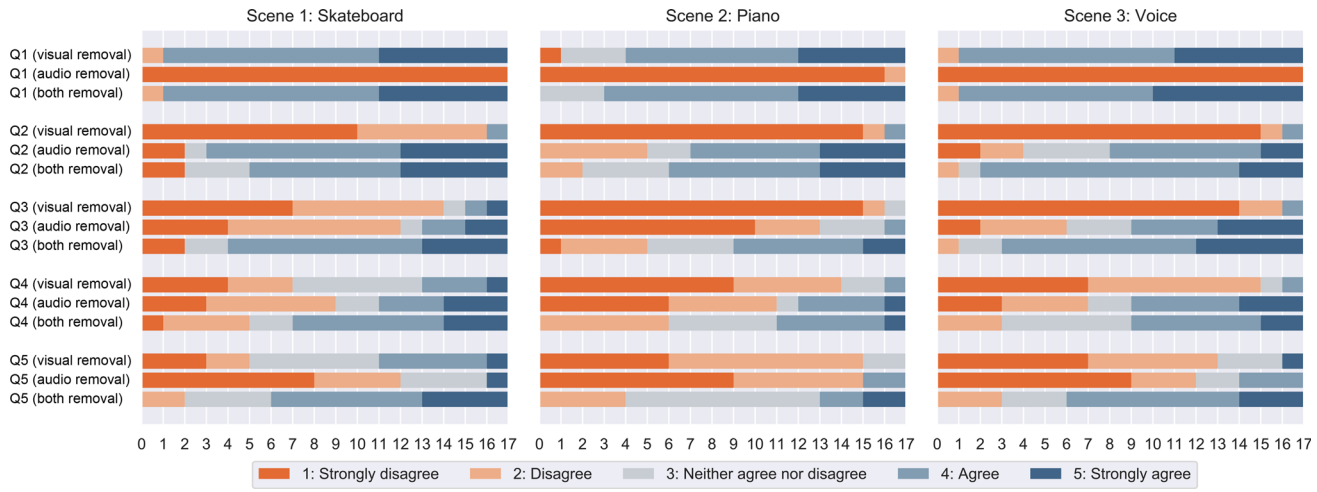

Fig. 8 Results of the questionnaires in the main study. For each scene, the participants answered five questionnaires (Q1–Q5) three times with different conditions (visual removal, audio removal, and removal of both)

Table 3 The average value (mean) and standard deviation (SD) for each scene and questionnaire

Scene	Type	Q1		Q2		Q3		Q4		Q5	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	Audio–visual removal	4.2	0.75	3.8	1.3	3.8	1.2	3.4	1.2	3.8	0.97
	Visual removal	4.2	0.75	1.5	0.80	1.9	1.1	2.6	1.2	2.9	1.2
	Audio removal	1.0	0.0	3.9	1.2	2.4	1.3	2.8	1.4	1.9	1.1
2	Audio–visual removal	4.1	0.70	3.8	0.97	3.2	1.1	3.1	0.97	3.1	0.92
	Visual removal	3.9	1.0	1.2	0.75	1.2	0.53	1.7	0.92	1.8	0.66
	Audio removal	1.1	0.24	3.5	1.2	1.7	0.99	2.4	1.4	1.7	0.99
3	Audio–visual removal	4.3	0.77	4.0	0.71	4.1	0.83	3.4	0.94	3.6	1.0
	Visual removal	4.2	0.75	1.2	0.75	1.3	0.77	1.8	0.83	1.9	1.1
	Audio removal	1.0	0.0	3.3	1.2	3.2	1.4	3.1	1.4	1.9	1.2

trials for each scene, we asked the participants to fill in the following questionnaires with a 5-point Likert-scale (from 1: Strongly disagree to 5: Strongly agree):

Q1: "The target was removed visually."
Q2: "The target was removed aurally."
Q3: "The view was synchronized with the sound."
Q4: "This edited video looks and sounds natural."
Q5: "I am satisfied with the result of this edited video."

After all trials were completed, an additional semi-structured interview was conducted for each participant.

4.3.3 Results of questionnaires

Figure 8 reports the results of the questionnaires in the main study. The means and standard deviations (SD) of each scene and question are summarized in Table 3. To evaluate whether there are statistical differences in the scores between the conditions (a) and (c), and between the conditions (b) and (c), we performed the Wilcoxon signed-rank test since the data are paired; the p values of the test are shown in Table 4. For Q1 (visual removal), the conditions of the visual removal (a) and audio–visual removal (c) received high scores for all scenes. For Q2 (audio removal), the conditions of audio removal (b) and audio–visual removal (c) received high scores for all scenes. For Q3 (synchronization), Q4 (naturalness), and Q5 (satisfaction), the condition of audio–visual removal (c) received higher scores than the other two conditions (a) and (b).

In addition to the results reported in Table 4, we also found significant differences in other data pairs as follows. Between visual removal and audio removal, there are significant differences in several cases. At scene2 and scene3, there is a significant difference too between visual removal and audio removal (scene 2: $p = 0.040$, scene 3: $p = 0.0035$). There are significant differences among scenes too. In scene 1, visual removal received relatively high scores compared

Table 4 The p value of each scene and questionnaire

Scene	Value	Q1	Q2	Q3	Q4	Q5
1	Against visual removal	1.0	< **0.001**	**0.0014**	**0.0021**	**0.0025**
	Against audio removal	< **0.001**	0.67	0.062	0.18	< **0.001**
2	Against visual removal	1.0	< **0.001**	< **0.001**	**0.0022**	< **0.001**
	Against audio removal	<**0.001**	0.28	**0.0059**	0.053	< **0.001**
3	Against visual removal	1.0	< **0.001**	<**0.001**	< **0.001**	< **0.001**
	Against audio removal	< **0.001**	0.066	**0.013**	0.35	< **0.001**

All values are compared with those of audio–visual removal (p values are shown in bold when they are smaller than 0.05)

with the other two scenes (between scene 1 and scene 2 at Q4: $p = 0.017$, between scene 1 and scene 3 at Q4: $p = 0.018$, between scene 1 and scene 2 at Q5: $p = 0.0037$, between scene 1 and scene 3 at Q5: $p = 0.0081$). Between the outdoor scenes (scene 1, and scene 3) and the indoor scene (scene 2), there is a tendency that users are more satisfied with audio–visual removal outputs for outdoor scenes compared with indoor scenes. For Q5, there are significant differences (between scene 1 and scene 2: $p = 0.047$, between scene 2 and scene 3: $p = 0.025$).

The results of the Likert-scale questionnaires showed that our multi-modal method could successfully provide better experiences than the baselines. Although the number of scenes was not large and their complexity was only moderate, we believe that this reasonably confirms our hypothesis that synchronized audio–visual removal is superior to single-domain removal in terms of user experience.

4.3.4 Discussions

In Q3 (synchronization), Q4 (naturalness), and Q5 (satisfaction), the condition of audio–visual removal surpassed the visual removal for all scenes. According to the feedback given in the interview, confusing sounds coming from nowhere are most commonly perceived for visual removal (*e.g.,* "When the target was removed only visually, I was confused by a sound coming from a place where there was nothing."). However, as the participants had no idea of the target object's original loudness (*e.g.,* "Sometimes it is natural, even if the target is removed only aurally because it is possible that the target did not generate any sounds."), there were no significant differences between the audio–visual removal and the audio-only removal for all the scenes in the perspective of naturalness (Q4). In addition, differences in scene properties seemed to lead to notable differences; as verified by Q5, the scores of the audio–visual removal from the outdoor scenes (scene 1 and scene 3) were higher than those from the indoor scene (scene 2). One possible reason is that the outdoor scenes had more sound sources and also stronger ambient sound (*e.g.,* "Compared with indoor scenes, I did not care about the quality of the audio removal output because there is an ambient soundscape present outdoors."). Besides, as the target object's sound was originally weak in scene 1, the visual-only removal satisfied some of the participants, and so the visual-only removal condition for scene 1 gained significantly higher scores in Q4 and Q5 than those for the other two scenes.

5 Limitations and future work

Our current implementation has several limitations. While our method was successful for the scenes used in the experiment, it does not work well for some more challenging scenes. For example, for a scene containing an audio source whose frequency is closer to that of the target object, our method would fail in separating these sounds because our method is based on STFT separation. Another challenging case is to track the target visually and to remove only the sound generated by the target, when the target and another audio source pass each other. This is because our method is based on optical-flow-based visual tracking, and therefore occlusion causes tracking errors and this direction-based method used to identify the target sound results in not only removing target sounds but also other audio sources close to the target.

We used a 360-degree video and a 4-channel microphone data this time. In video inpainting process, the limitation of input video's resolution with existing video inpainting methods amplifies the problem of visual artifacts in outputs. We believe the quality can be easily improved with a better video inpainting algorithm. In audio removal process, we plan to improve the quality by using more channels or referring to upcoming higher-quality audio separation and localization methods. In this paper, the result of the estimation of audio direction is not used in the process of estimating the direction with visual cue. In the future, the feasibility of using audio-directional information for visual tracking in difficult scenes (*e.g.,* the removed target is occluded) is worth investigating. In addition, we will consider trying audio inpainting (*i.e.,* filling the edited soundscape with plausible synthesized sounds) instead of audio removal. Another future study would be to

implement an end-user-friendly interface for audio–visual removal. We validated the concept through the user study with several test scenes. However, there are still rooms to improve our method for more conditions considering daily scenes. Therefore, we plan to add new scenes which include challenging scenes and conduct comprehensive studies with various complicated scenes taken by end users.

6 Conclusion

In this paper, we introduced a novel concept, the audio–visual removal of unwanted objects in 360-degree videos. We realized this concept using a two-stage approach. Instead of solely removing visual cues, our method incorporated visual information acquired in the video inpainting process to eliminate corresponding auditory cues synchronously. Our user study with several test scenes indicated that our multimodal approach could offer more synchronous, natural, and satisfactory user experience. We envision conducting next experiments against more challenging scenes to validate the concept more extensively. In addition, we plan to propose a higher quality method for broad range of potential applications in 360-degree video editing.

Funding This work was supported by JST ACCEL Grant No. JPMJAC1602, JST-Mirai Program Grant Number JPMJMI19B2 and JSPS KAKENHI Grant No. JP19H01129, Japan.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Akyazi, P., Frossard, P.: Graph-based inpainting of disocclusion holes for zooming in 3d scenes. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 867–871. IEEE (2018)
2. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier–Stokes, fluid dynamics, and image and video inpainting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, vol. 1, pp. I–I. IEEE (2001)
3. Bertalmío, M., Caselles, V., Haro, G., Sapiro, G.: Pde-based image and surface inpainting. In: Handbook of Mathematical Models in Computer Vision, pp. 33–61. Springer (2006)
4. Facebook: How do I upload a 360 video on Facebook?—Facebook Help Center. Retrieved November 19, 2019 from https://www.facebook.com/help/828417127257368
5. Feng, W., Guan, N., Li, Y., Zhang, X., Luo, Z.: Audio visual speech recognition with multimodal recurrent neural networks. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 681–688. IEEE. (2017)
6. Gerzon, M.A.: Periphony: with-height sound reproduction. J. Audio Eng. Soc. **21**(1), 2–10 (1973)
7. Google: Upload 360-degree videos—YouTube Help. Retrieved November 19, 2019 from https://support.google.com/youtube/answer/6178631
8. Hershey, J.R., Casey, M.: Audio-visual sound separation via hidden Markov models. In: Advances in Neural Information Processing Systems, pp. 1173–1180 (2002)
9. Insta360.com: Insta360 360 Camera—Insta360, the leader in 360 cameras. Retrieved November 19, 2019 from https://www.insta360.com/
10. Kim, D., Woo, S., Lee, J.Y., So Kweon, I.: Deep video inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5792–5801 (2019)
11. Korman, S., Avidan, S.: Coherency sensitive hashing. IEEE Trans. Pattern Anal. Mach. Intell. **38**(6), 1099–1112 (2015)
12. Kowalczyk, K., Thiergart, O., Taseska, M., Del Galdo, G., Pulkki, V., Habets, E.A.: Parametric spatial sound processing: a flexible and efficient solution to sound scene acquisition, modification, and reproduction. IEEE Signal Process. Mag. **32**(2), 31–42 (2015)
13. Le Meur, O., Gautier, J., Guillemot, C.: Examplar-based inpainting based on local geometry. In: 2011 18th IEEE International Conference on Image Processing, pp. 3401–3404. IEEE (2011)
14. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: Advances in Neural Information Processing Systems, pp. 362–372 (2018)
15. Mroueh, Y., Marcheret, E., Goel, V.: Deep multimodal learning for audio-visual speech recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2130–2134. IEEE (2015)
16. Nair, A.A., Reiter, A., Zheng, C., Nayar, S.: Audiovisual zooming: what you see is what you hear. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1107–1118. ACM (2019)
17. Paredes, D., Rodriguez, P., Ragot, N.: Catadioptric omnidirectional image inpainting via a multi-scale approach and image unwrapping. In: 2013 IEEE International Symposium on Robotic and Sensors Environments (ROSE), pp. 67–72. IEEE (2013)
18. Pulkki, V.: Spatial sound reproduction with directional audio coding. J. Audio Eng. Soc. **55**(6), 503–516 (2007)
19. Ricoh Company, Ltd.: 360-degree camera RICOH THETA. Retrieved November 19, 2019 from https://theta360.com/
20. Rivet, B., Wang, W., Naqvi, S.M., Chambers, J.A.: Audiovisual speech source separation: an overview of key methodologies. IEEE Signal Process. Mag. **31**(3), 125–134 (2014)
21. Ruochen, W., Yuhong, Z., Wei, Z.: Acoustic zooming based on real-time metadata control. In: 2014 4th IEEE International Conference on Network Infrastructure and Digital Content, pp. 338–342. IEEE (2014)
22. Van Veen, B.D., Buckley, K.M.: Beamforming: a versatile approach to spatial filtering. IEEE ASSP mag. **5**(2), 4–24 (1988)
23. Upenik, E., Akyazi, P., Tuzmen, M., Ebrahimi, T.: Inpainting in omnidirectional images for privacy protection. In: ICASSP 2019-

2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2487–2491. IEEE (2019)
24. Vilkamo, J., Lokki, T., Pulkki, V.: Directional audio coding: virtual microphone-based synthesis and subjective evaluation. J. Audio Eng. Soc. **57**(9), 709–724 (2009)
25. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: a unifying approach. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
26. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
27. Yang, W., Qian, Y., Kämäräinen, J.K., Cricri, F., Fan, L.: Object detection in equirectangular panorama. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2190–2195. IEEE (2018)
28. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514 (2018)
29. Zioulis, N., Karakottas, A., Zarpalas, D., Daras, P.: Omnidepth: Dense depth estimation for indoors spherical panoramas. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 448–465 (2018)

Ryo Shimamura received a B.E. degree in the Department of Applied Physics from Waseda University. He is currently a student in Waseda University of the Department of Pure and Applied Physics. His research interests are virtual reality, augmented reality, human–computer interaction, and signal processing.

Qi Feng received the B.E. and M.E. in Applied Physics from the Graduate School of Advanced Science and Engineering at Waseda University, Tokyo, Japan in 2017 and 2019, respectively. He is currently pursuing his Ph.D. degree in Waseda Univerity. His main research area includes deep learning applications, computer vision, computer graphics, virtual and augmented reality.

Yuki Koyama is a Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. He received his Ph.D. from the University of Tokyo in 2017. His main research field is the intersection of computer graphics and human–computer interaction. In particular, he is interested in developing computational techniques for enabling new interactions, producing creative artifacts, and enhancing design processes.

Takayuki Nakatsuka received a B.E. degree in the Department of Applied Physics and an M.E. degree in the Department of Pure and Applied Physics from Waseda University, Japan. He is currently a student in the Graduate Schools of Advanced Science and Engineering and Graduate Program for Embodiment Informatics at Waseda University. His primary research interests are in physically-based animation, human motion analysis, human computer interaction, and virtual reality.

Satoru Fukayama received his Ph.D. degree in information science and technology in 2013 from the University of Tokyo. He is currently a senior researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His interests are in music information retrieval, especially music generation with probabilistic models. He has received awards, including IPSJ Yamashita SIG Research Award, several Best Presentation Awards, and Specially Selected Paper Award from the Information Processing Society of Japan.

Masahiro Hamasaki received his Ph.D. degree in Informatics in 2005 from Soken University. He is currently a group leader of Media Interaction Group, at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. He is also an associate professor at the University of Tsukuba. His research interests cover online community assistance, knowledge construction, social media analysis, and web intelligence.

Masataka Goto received the Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. Over the past 28 years he has published more than 270 papers in refereed journals and international conferences and has received 51 awards, including several best paper awards, best presentation awards, the Tenth Japan Academy Medal, and the Tenth JSPS PRIZE. In 2016, as the Research Director he began OngaACCEL Project, a 5-year JST-funded research project (ACCEL) on music technologies.

Shigeo Morishima was born on August 1959. He received the B.S., M.S. and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, in 1982, 1984, and 1987, respectively. He was a visiting professor of University of Toronto from 1994 to 1995 and an invited researcher of Advanced Telecommunication Research institute from 1999 to 2011. Currently, he is a professor of School of Advanced Science and Engineering, Waseda University. He was a General Chair of ACM VRST 2018 and VR/AR adviser of SIGGRAPH ASIA 2018. He received many awards and takes an administration board member of several societies.