Image Sonification System Based on Auditory Icons and Stereophonic Presentation Considering Image Scenes and Human Action

Yuichi Mashiba †, Hisham E. Bilal Salih ††, Naoto Wakatsuki †††, Koichi Mizutani †††, Keiichi Zempo †††

Abstract In this paper, we develop an automatic image sonification system using auditory icons and their stereophonic presentation. We evaluated the effects of scene sounds and action sounds on the understanding of the scene and the sense of presence. The results showed that the scene sounds contributed to the understanding of the scene and the sense of presence and that action sounds also improved the sense of presence. On the other hand, it was confirmed that object sounds alone can convey scene information in images and have a high index of sense of presence. Our sonification system is expected to provide not only a more immersive experience for sighted people through images but also an opportunity for people with visual impairments to share the experience through social networking services (SNSs).

Key words: People with Visual Impairment, Image Sonification, Auditory Icon, Stereophonic Presentation, Scene, Human Action

1. Introduction

As of 2021, there were 4.2 billion users of social media; social networking sites have become an indispensable tool in modern life. Interactions on these sites are important for people with visual impairment, as well as sighted people, to maintain social connections. On the other hand, visual content such as images and videos are rapidly increasing on social networking sites. More than 2 billion photos are uploaded daily on Facebook, Instagram, Messenger, and WhatsApp.

However, there are limited ways for people with visual impairments to access such visual content. For example, if the content is an image, it is necessary to provide a description of the image, such as the alt text (that is, the text written by web page authors in the alt field of the Hypertext Markup Language (HTML) tag of images). In addition, screen readers, such as VoiceOver and TalkBack, make this content accessible to people with visual impairments. However, studies on web accessibility indicate that the lack of alt text for images is one of the major barriers to accessibility. In particular, images on social networking sites pose a serious problem, with only 0.1% of images on Twitter having alt text. Despite the increasing importance of image content in social networking sites, people with visual impairments are unable to fully utilize it. Therefore, there is a need for an approach that supports people with visual impairments to communicate with images in the same way as sighted people.

In addition, the number of social networking services (SNSs) that can provide such usability-conscious services is limited. The reason is that service providers are not aware of these accessibility features, and the cost of implementing them is a burden. To solve this problem, it is necessary to provide accessibility functions that can be used by all users, regardless of whether they are people with visual impairments.

Based on the above two points, a method of supporting image communication that is accessible to all users, regardless of visual impairment, is necessary.

In a previous study, we proposed a method of image sonification using stereophonic presentation and auditory icons, which are sounds that are associated with objects or events. Auditory icons are abstract sound and represent information about an object and event (e.g., the engine sound of car, walking sounds). The proposed method is characterized by the use of stereophonic synthesis of nonverbal sounds that repre-
The purpose of this paper is to evaluate the impact of additional sounds that take into account the scene and human actions of an image on the understanding of the scene and the sense of presence, respectively.

2. Related works

2.1 Tactile display

There are numerous approaches for tactile\textsuperscript{15,16} and audio-tactile displays\textsuperscript{17,18}. The advantage of tactile presentation is that it is suitable for comprehending spatial locations on charts and maps, for example. Combined with audio, it is also possible to comprehend the shape and obtain information about the object to describe the object at the location being touched.

On the other hand, the sense of touch enables comprehension of linear shapes, but it is not suitable for comprehending representations of complex shapes such as photographs. Additionally, the tactile system is not very portable, and it is expensive to produce.

2.2 Verbal auditory display

There are systems that use image recognition to automatically generate alt text and make it audible. Methods such as character recognition extract characters in an image\textsuperscript{19} and summarize them into words or sentences and use image recognition to convert objects, settings, image characteristics, etc.\textsuperscript{20}. Other systems combine these technologies\textsuperscript{21,22}, such as Microsoft’s Seeing AI.

While verbal descriptions can present the features of an image in detail, the text is descriptive and does not reflect the elements that people with visual impairments require\textsuperscript{23}. Therefore, it is difficult for these people to obtain the same kind of experience that a sighted person obtains when looking at a photograph. This is because, unlike sighted people who exhibit visual predominance, people with visual impairments experience sensations created from substantial information from senses other than sight when visiting a place. As a result, it is difficult for them to relive the experience with images and alt text.

2.3 Nonverbal auditory display

A method exists that makes an image audible by nonspeech audio by using abstracted sounds. The use of nonspeech audio to convey information is called sonification\textsuperscript{24}, and parameter mapping is a typical method. This approach converts the parameters of certain data into sound parameters such as rhythm and amplitude. The human auditory system has the ability to learn and understand complex acoustic patterns, and these patterns can be used as input to present sounds that
combine multiple parameters and transmit information. For example, one method makes the brightness of a pixel audible based on the amplitude and the position of the frequency and performs mapping from color information (hue, saturation, lightness attribute and HSL-Color) to pitch, timbre, and loudness. Other research addressed combining touch panels to enhance interactivity and express objects and colors at points of tactile contact with sound. For tactile contact with sound, the brightness of a pixel audible based on the amplitude and the position of the frequency is used to weight these bounding boxes. COCO has 91 different labels for 200,000 color images and is a large-scale dataset for object detection, segmentation, and captions. We used the YOLOv5s pretrained model. The mean average precision (mAP) of YOLOv5s is 56.0, and this value applies to the single-model single-scale on the COCO val2017 dataset.

The proposed method uses abstracted sound icons based on the sounds emitted from real objects. Therefore, it can be understood intuitively, and the scene can be recognized. In addition, the three-dimensional (3D) acoustic presentation is based on the positional relationship of sound sources in the image. This approach has the feature of helping people with visual impairments understand the situation of the photograph by immersing them audibly in the situation in which the photograph was taken rather than through exposition.

3. Image sonification system

3.1 System overview

An automatic sonification system was developed to evaluate auditory signals with respect to image scenes and human actions. The user accesses the image to undergo auditory conversion and receives a sound from the system. The sound lasts 10 seconds and is heard using headphones or earphones. Object detection, scene recognition, and action detection are used to obtain image information (Figure 2). This approach is implemented because object detection alone is insufficient for users to recall the scene of an image or to obtain a sense of presence to present image information. Therefore, we hypothesize that scene recognition and action detection can contribute to the user’s understanding of the image and the sense of presence by presenting scene and action sounds that match the contents of the image. Based on the results of image recognition, the most appropriate sound data for the object are selected. In addition, each object’s center coordinates determine the direction in which it is stereophonicized.

3.2 Recognition and detection

For object detection, we use the model of the COCO dataset trained in YOLO. YOLO is an object detection model that performs end-to-end object class and bounding box from images. In YOLO, the image is first divided into areas, and the class probability and bounding box of the object are obtained for each area. This probability is used to weight these bounding boxes. COCO has 91 different labels for 200,000 color images and is a large-scale dataset for object detection, segmentation, and captions. We used the YOLOv5s pretrained model. The mean average precision (mAP) of YOLOv5s is 56.0, and this value applies to the single-model single-scale on the COCO val2017 dataset.

For scene recognition, we use the model in which the Places365 dataset is trained with 50-layer residual network (ResNet). Our learning approach uses 1.8 million images of 365 types. On the Places365 validation set, the top-1 error is 4.14%, and the top-5 error is 14.71%. In this classifier, five scene categories are assigned in descending order of the reliability of classification; however, in this application, the scene category with the highest confidence score is used for sonification.

For action detection, we use the human activity feature extractor and the human body part state (PaSta) detector based on Human Activity Knowledge Engine (HAKE) data. After detecting the person and estimating the posture from the image, the human body part is detected, the state is estimated, and the behavior is classified using the result. The PaSta detector uses a model learned based on PaStaNet. PaStaNet is a dataset containing 118K+ images and 285K+ people. Ten body parts, such as the head, limbs, and hips, are labeled, for example, to see something or eat something if the body part is the head. PaStaNet includes active human and object bounding boxes of 156 activities. We used the pretrained model on Pasta. The mAP on verb detection is 12.23.

3.3 Sound selection

A suitable sound is associated with the label name obtained from image recognition. For the sounds, we prepared sounds for the representative ones that can be made audible from the labels used in each discriminator. We used Freesound for the sound dataset.

In advance, we performed object detection on Flickr30k (a dataset created from images uploaded to Flickr, an image sharing SNS). We prepared sounds for objects that can be expressed by an auditory icon for objects with a larger total number of detections. The scene sounds are binaural recordings. For the sounds of human actions, we did not do the above because of
the classification accuracy but prepared sounds for the actions that can be expressed by the Auditory icon.

### 3.4 Spatialization

Objects and actions recognized by object detection and action detection are made stereophonic according to their positions. The position of the object is assumed to be the center coordinate of the bounding box. Based on the positional information of the object, the sound associated with the object is converted into a stereophonic sound. This is done using the head-related transfer function (HRTF). This function is a frequency-domain representation of the physical characteristics of sound waves transmitted from a sound source to the eardrum, including differences in volume, time, and frequency characteristics. Humans estimate the direction of a sound source by perceiving the changes caused by the reflection and diffraction of the sound emitted from the source on the floor, shoulders, head, and auricles. The HRTF is measured by using a dummy head with a microphone in the ear canal or by placing a microphone in the ear canal. The measured HRTFs of the right and left ears are convoluted into a sound, and the sound is localized by playing each sound into the left and right ears. We used HRTFs described previously.

By listening to this signal in stereo, the user can identify the location of the sound source from the direction in which the sound is heard, without relying on vision. By using the HRTF, the sound is made stereophonic in the range of 90° left and right and 45° up and down from the front direction. The minimum angle is 5°. The stereophonic angle is determined based on the center coordinates of the object that is the sound source. The volume of each sound is normalized so that there is no difference in volume between objects. For the scene sounds, binaural recordings are used, so they are synthesized without stereophonicization. Through the above process, we generated stereophonic sound based on not only the objects in the image but also the scene and human action.

In the next section, we evaluate how sound presentation considering scenes and human actions affects image scene comprehension and presence.

### 4. Sound considering the scene and human action

#### 4.1 Overview

In this experiment, we investigated how the combination of object sounds, scene sounds, and action sounds affects image comprehension and the sense of presence. Participants were asked to answer the question after listening to the audible sound of the image. The image used in the experiment is shown in Figure 3. These images were randomly extracted from Flicker30k. To evaluate the degree of agreement between the generated sound and the visual elements, this experiment was conducted on sighted participants.

The participants were 13 Japanese people (9 males, 4 females). The average age was 28 years.

---

* It is generally recognized that performance evaluation by only a small number of visually impaired people is inappropriate, and therefore sighted participants with blindfolds should be included. Additionally, it is suitable for qualitative user experience evaluations with small numbers of people with visual impairments but not for performance evaluations, such as Ref. [42].

** The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Review Committee of the Faculty of Engineering, Information and Systems, University of Tsukuba (2019R305).
Table 1  List of the sounds. This is also part of the list of options presented to participants in the experiment (Section 3).

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene</td>
<td>Street, Beach, Park, Subway station/platform, Amusement park, Coffee shop, Bus interior, Baseball field, Football field, Swimming location</td>
</tr>
<tr>
<td>Object</td>
<td>Car, Bus, Bicycle, Train, Motorcycle, Truck, Boat, Bird, Dog, Horse</td>
</tr>
<tr>
<td>Action</td>
<td>Walk, Run, Jump, Clap hands, Eat, Fish, Swing, Work on a computer</td>
</tr>
</tbody>
</table>

Table 2  To investigate how scene and action sounds affect the understanding of the image scene and sense of presence, respectively, we conducted experiments in four situations.

<table>
<thead>
<tr>
<th>Action sound</th>
<th>Scene sound</th>
<th>w/o</th>
<th>w/</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o</td>
<td>cond. O</td>
<td>cond. OS</td>
<td></td>
</tr>
<tr>
<td>w/</td>
<td>cond. OH</td>
<td>cond. OSH</td>
<td></td>
</tr>
</tbody>
</table>

4.2  Scene understanding

First, we evaluated the degree of understanding of the image scene. Participants listened to the sound of the images in the following four conditions shown in Table 2.

Ten images including eight different scenes (Figure 3 shows the sonified images) were sonified and presented to participants. Participants were asked to choose from among the options where they could understand the sound and what sound they could hear. The options are those listed in Table 1 plus “I do not know” and “I cannot hear anything”. Participants could not see the image. Participants were asked to listen to all combinations of sounds, and the order in which the sounds were presented was randomized for each.

The sonified images were checked beforehand to determine whether they were correctly labeled, and if not, they were corrected before sonification. This is because the purpose of this experiment is to investigate how the addition of scene and action sounds affects image understanding and sense of presence.

4.3  Sense of presence

For the evaluation of the sense of presence, the question about reality from Witmer and Singer’s Presence Questionnaire (PQ)\(^{43}\) was used. There are two specific questions:

Q1  How inconsistent was the information coming from the sound and the image?
Q2  How well did the sounds match your real-world experience?

Participants were presented with sounds and images and asked to answer a questionnaire. Specific questionnaire items are shown below. Using these questions, participants evaluated themselves on a Likert scale of 1 to 5.

\(^{38}\) Flicker30k
\(^{43}\) PQ

Fig. 3  Images and components of sound used in the experiment. These images were randomly extracted from Flicker30k.\(^ {38}\)
Table 3  Correct answer rate for scenes. In many cases, this rate increased when scene sounds were added. On the other hand, this rate did not increase when action sounds were added.

<table>
<thead>
<tr>
<th>Scene/Image No.</th>
<th>Sound component: p &lt; 0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beach(1)</td>
<td>O 76.9</td>
</tr>
<tr>
<td>Bus interior(2)</td>
<td>O 46.2 OS 0.0 OH 15.4</td>
</tr>
<tr>
<td>Street(3)</td>
<td>O 92.3 OS 7.7 OH 76.9</td>
</tr>
<tr>
<td>Baseball field(4)</td>
<td>O 53.8 OS 7.7 OH 76.9</td>
</tr>
<tr>
<td>Subway station(5)</td>
<td>O 92.3</td>
</tr>
<tr>
<td>Street(6)</td>
<td>O 30.8 OS 92.3 OH 61.5 OSH 69.2</td>
</tr>
<tr>
<td>Watering hole(7)</td>
<td>O 0.0 OS 7.7 OH 0.0</td>
</tr>
<tr>
<td>Football field(10)</td>
<td>O 69.2 OS 0.0 OH 61.5</td>
</tr>
</tbody>
</table>

The Mann–Whitney U test was performed to determine if there was a significant difference in the responses depending on the configuration of the presented sounds. When the p value is less than or equal to 0.01, the null hypothesis that the mean user ratings for each set of sound configurations are approximately the same can be rejected. This means that the mean user ratings for the sound composition of each set are significantly different.

4.4 Result

(1) Comprehension of the scene

The correct answer rate for the scene is shown in Table 3. When scene sounds were added (condition OS), the correct answer rate increased significantly. When human action sound was added (condition OH), there was no significant difference in the correct answer rate.

Additionally, when scene and action sounds were added (condition OSH), there was no significant difference in the correct answer rate compared to the case with scene sounds (condition OS). The correct answer rate also differed depending on the image, and this rate did not change much for the swimming location of image no. 7 even if the scene sound was added. In addition, no one was able to answer correctly with only the object sounds, except for image no. 6.

(2) Sense of presence

The results of the questionnaire are shown in Figure 4,5. In both cases, the evaluation was higher when the sound was combined with other sounds than when only the object sound was presented. The following results were obtained in order of highest score: object and scene sounds, object, scene and action sounds, object and action sounds, and object sounds only. In Q1 (Figure 4), there was a significant difference except for the object and scene sound and the object, scene and action sound.

5. Discussion

5.1 Understanding of the scene and sense of presence

The experimental results showed that the presentation of scene sounds improved the rate of correct responses regarding the image scenes. On the other hand, the human action sound did not contribute to the improvement of the correct response rate of the scene. Therefore, it was confirmed that the sound of human action was not necessarily associated with the understanding of the scene.

In addition, the object and scene sounds (condition OS) was the most highly evaluated for the sense of presence. It was confirmed that adding scene sounds and action sounds can enhance the sense of presence of images and sounds.
On the other hand, in image nos. 2 and 7, the comprehension of the scene was lower when scene sounds were added, but each index of the sense of presence was improved. This suggests that scene sounds do not necessarily improve understanding of the scene, but they can improve the sense of presence. In addition, as shown in image no. 6, the percentage of correct responses to the scene was high even with only object sounds, and the index for the sense of presence was also high. This suggests the possibility of enhancing the scene information and the sense of presence of a certain image even with only object sounds. Furthermore, for images nos. 2 and 7, the evaluation values for the sense of presence when action sounds were added were all lower than those for the other images. This may be because the action sounds were not in accordance with the scene or the scene sounds. In the case of no. 2, the sound of a person walking was added to the sound of a running bus. In the case of no. 7, a person is jumping toward the water, but the action sound is landing on the ground. This may be the reason why the evaluation of realism decreased when scene and action sounds were added.

Image no. 6 was different from the other images in that the correct response rate for the scene was 30.8% even with only object sounds. In this image, unlike the other images, the correct response rate for the scene was 30.8% even with only the object sound, and the correct response rate increased to 61.5% by adding the action sound. The object sounds in this image are a car and a dog, and the action sounds are walking sounds. This suggests that it is possible to present information about a place only by combining the components of the auditory icon, object sounds and action sounds. Similarly, the evaluation value of each index related to the sense of presence was higher in the case of only object sounds. This suggests that the scene sound is not necessary when the image contains enough objects that can be the source of the scene sound.

For the degree to which the sounds deviate from what is experienced in the real world, the number of sounds combined is the largest, and the total variance is the largest for sounds with scene and action sounds. This may be because combining multiple sounds makes the final sound unnatural.

### 5.2 Limitations

We now discuss image recognition and auditory icon presentation.

For image recognition, the accuracy of action recognition is low. This is due to the current limitations of image-based action detection. We plan to solve the discrepancy between the actual action and the presented sound in the future by improving the accuracy of action detection. In addition, if all audible objects in an image are presented as sounds, it may reduce the understanding of the image (for example, a car seen through a window in an indoor image). An optimized discriminator is needed to recognize the audible objects appropriately.

Next, we discuss auditory icons. The proposed method is not able to give the sound of the object that matches the scene of the image. For example, the sound of a person walking outdoors and the sound of a person walking indoors both have the same sound. In this way, the sounds of human actions are also different depending on the characteristics of the person and the scene, but this aspect is not reflected.

There is also a problem of how users interpret auditory icons. This is confirmed by the experimental
results. The user’s interpretation of the auditory icon may depend on the user’s experience. For example, one participant interpreted the sound of a bus running as the sound of a train running. To prevent this, it is necessary to investigate auditory icons optimized for individuals.

6. Conclusion

In this paper, we developed an automatic sonification system that adds scene sounds and human action sounds as auditory elements. We investigated the effects of stereophonic presentation considering the image scene and human action on the user’s understanding of the image and the sense of presence. As a result of the experiment with 13 sighted participants, it was confirmed that the scene sound improved the understanding of the scene depending on the image. On the other hand, action sounds did not lead to a better understanding of the image. Regarding the sense of presence, it was confirmed that both scene sounds and action sounds improved the sense of presence. In addition, the scene sound improved the sense of presence even for images that did not improve the understanding of the image. In addition, it was suggested that scene information and a sense of presence can be obtained even with only object sounds when the sound source that characterizes the scene is included in the image. It is expected that our sonification system provides not only a more immersive experience for sighted people through images, but also an opportunity to share the experience with people with visual impairment through SNS.

References


41) Andrew Sears and Vicki L. Hanson. Representing users in accessibility research. 4(2), 2012.


Yuichi Mashiba received the Master’s degree in Eng. from University of Tsukuba in 2022. He belongs currently to Doctoral Program in Intelligent and Mechanical Interaction Systems, University of Tsukuba.

Hisham Elser Bilal Salih received the Master’s Degree in Disability Sciences from University of Tsukuba in 2010. He is currently a Researcher with the Faculty of Engineering, Information and Systems.