Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12914-z

Dataset and semantic based-approach for image sonification

O. K. Toffa[1,2] · **M. Mignotte**[1,2]

Received: 22 October 2021 / Revised: 22 December 2021 / Accepted: 9 March 2022

Abstract
This paper presents an image-audio dataset and a mid-level image sonification system that strives to help visually impaired users understand the semantic content of an image and access visual information *via* a combination of semantic audio and an easily decodable audio generated in real time, both triggered by sliding, taping, holding actions when the users explore the image on a touch screen or with a pointer. Firstly, we segmented the original image using a label fusion model and based on the user position in the image, a sonified signal is generated using musical notes and meaningful visual information within the active region like the color and the luminance, then the gradient and the texture. Secondly, we integrated the semantic understanding of the image into our model using DeepLab semantic segmentation of the image and created a dataset of audio and images aligned on the 20 classes of the PASCAL VOC 2012 dataset. The dataset of images are organized based on color, gradient, texture for low-level sonification and on semantic content with sounds for mid-level sonification. Thirdly, in order to provide both types of information in a complementary way, the slide, tap and hold actions of a touch screen are incorporated in the model. The semantic audio providing a brief description of the visual object is played on slide action, the generated signal with color details of the object on the tap action, gradient and texture of the object on hold action. Finally, we validated our sonification model on the provided dataset during a pilot study and the subjects were generally able to identify the objects in the image, the color of the objects and even provide a general description of the scene of the image. Our system could be useful to visually impaired persons in a photo sharing application using a smartphone or for painting art description in a digital museum.

Keywords Sonification · Visually impaired · Touch screen · Image accessibility · Auditory feedback

O. K. Toffa and M. Mignotte contributed equally to this work.

✉ O. K. Toffa
ohini.kafui.toffa@umontreal.ca

Extended author information available on the last page of the article.

1 Introduction

Image sonification is the translation of image data, which describe shape, color and texture (sometimes depth information), objects, into sounds. It applies the sonification, the use of non-speech audio to convey information or perceptualize data, to image accessibility domain (see [25] for an interesting review of image accessibility). With an increased number of peoples who suffer from visual impairment (blind or low vision) due to the aging and growth of the world population and the increased presence of images on screens in daily activities through social media, image accessibility using sonification has gained much more attention. Sonification is already used in emergency services, aircraft cockpits, assistive technologies (Microsoft text to speech or iOS VoiceOver), climate sciences [13], elite sports [30], multimodal interactive environments [32], engineering analyses and simulations and interpretations based on the sonification of physical quantities [10]. More than ever, sonification has become more interactive [9] with the evolution of the technology of touch and tactile screen on top of smartphones, tablets and wearable devices, the increased computing capacity of CPU/GPU and advanced techniques of machine learning. Sonification applied to image is a recent field that has naturally emerged after the development of image and sound processing techniques and is used in assistive technologies for blind people through navigation [1, 21], art sonification in digital museum [4, 16], music composition [35], photo sharing [39], social media [34].

2 Background

Research in image sonification is usually classified in two categories: high-level sonification where a natural language is used to describe the visual content of the image and low level sonification where the same content is translated to a non-speech audio. High-level sonification aims to describesthe semantic content of an image to a visually impaired people. To detect that semantic content some will use an object recognition engine like Sudol et al. [31] with LookTel a system that remotely captures the stream video from the camera of a smartphone using a 3G signal, process it with object recognition engine and returns in real time the name of the object using text to speech engine. In [2], the authors used SVM detection and Bag of visual world available in OpenCV to detect the presence of objects in the scene and inform the user via a speech output. Though recent breakthroughs in machine learning using deep learning [17] have pushed the boundaries of semantic image analysis, it stays very complicated, to fully understand the semantic content of an image. That is why some researchers [23] prefer to use real-time crowdsourcing and image annotation technique to speak the alt text aloud to people who are visually impaired. Some methods will manually parse the content of the image, especially in the art domain, in order to create an exploration map or a 3D printed version of the original image, then guide the user using voice control and haptic feedback [4, 16, 26, 27]. Such methods have the advantage to describe not only the objects in the image but also the color, the luminosity and texture since it is manually made. Definitely, most of automated methods in high-level sonification will only describe the objects identified in the image without being able to provide full information related to the edges, the shape, the color variation, and texture. Such sonification is not able to interpret abstract drawings and painting.

Low-level image sonification aims to translate image features into a non-speech audio by mapping data between the visual and the audio domains, between the 2 dimensions time-independent of the image and 1 dimension time-dependent of the audio. This is not trivial

and requires heuristic design to create a sonification system that is intuitive and easy to interpret by a listener who must hear a sound to visualize a content. In the domain of assistive technologies for blind people, low-level sonifications were initially used to develop navigation systems to improve visually impaired people's mobility using cameras [1, 5, 7, 21]. In terms of direct experience with a natural or synthetic image, peoples prefer to use the same approach as Braille alphabet with 3D printed tactile graphics to cover a tablet or haptic touchscreen which transform virtual image in a physical one or a tactile feedback for touch screens in the form of physical guides that are overlaid on the screen and recognized by the underlying application [12, 15]. Few works have used sonification to help visually impaired people to recognize objects, feel colors, gradient and texture, perceive edges and shapes in a synthetic, natural image or painting, in order to draw some conclusions about the content of the image (or video frames). Simple conversion methods were dedicated to map different characteristics of the image to sound. The brightness of the pixel is converted to the volume of generated sound [37] or musical notes [20], texture pattern into a periodic signal [19], colors to the wave envelope, waveform and frequency of a sound of an oscillator [14], edge to frequency of a sound of an oscillator [38]. More sophisticated methods were developed like the one in [29] which exploit the richness of the color by mapping HSV color space to different characteristics of the sound: Hue (H) to the fundamental frequency, Saturation (S) to signal's spectral envelope, Value (V) to the loudness of the synthesized sound. The authors in [3], used the concept of color, color mixture with the combination of acoustical entities and the grade of roughness on pre-classified natural regions and edges with color distribution for each region of the image) features. The proposed system mainly uses musical notes at several octaves, the notion of timbre, and loudness but also uses pitch, drum rhythms in their sonification system. A more recent approach developed in our previous paper [33] captures the most useful and discriminant local information about the image content at different levels of abstraction, ranging from low-level (at the pixel level) to high-level (segmentation) and combining low-level(color edges and texture), mid-level and high-level (gradient or and the distortion effect in an intuitive way to sonify the image content both locally and globally. Though a low-level sonification can interpret the abstract content of an image it can not identify and name the objects present in that image.

In this work, we present a new image sonification system, called mid-level sonification, because it exploits low-level and high-level sonification techniques in a complementary way and use a semantic non-speech audio instead of a voice description of the objects. Such system offers to the end user a global experience that covers most of features available in an image: objects and semantic content, the edges, the shape, the color variation, the luminosity and texture. We segmented the original image using a label fusion model [22] and used a low-level hierarchical-based sonification approach [33] to generate a sonified signal that exploits musical notes and the position of the user in the screen in order to convey meaningful visual information within the active region like the color and the luminance at first, then the gradient and the texture finally. In order to understand the semantic context of the image for high-level sonification, we integrated DeepLab [6, 28] semantic segmentation of the image that permits, not only to identify up to 20 classes of the PASCAL visual object class 2012 [11], but also their edge and shape instead of just their presence using a bounding box as in some previous papers [2]. We created a dataset of non-speech audio aligned with the 20 classes because sounds provide better description to a blind person than a vocal description of something they can never see but are used to hear. Since there is not any images dataset dedicated to image sonification and peoples usually struggle to find data on which testing their model we provided a dataset of images that contain different color

variation, gradient, texture for low-level sonification and 20 different classes of objects for high-level sonification. To provide both types of information in a complementary way without any additional haptic or tactile accessory device, we incorporated the slide, tap and hold actions of a touch screen into the model. A semantic non-speech audio providing a brief description of the visual object is played on slide action, a generated signal with more detailed on the color on tap, gradient and texture of the object on hold action. Contrary to methods [4, 16, 26, 27] that offer an equivalent complete experience, our approach is automated, uses non-speech descriptive sound instead of a vocal description of content that a blind person has never seen and does not involve additional haptic or tactile material except a touch screen. Our system could be useful to visually impaired persons in a photo sharing application using a smartphone or for painting art description in a digital museum. The dataset and source code of the application is available at https://github.com/ohinitoffa/ImgSonficiation2.

3 Sonification model

To have a complete and detailed description of an image, a sonification model must convey as much as discriminant local information available at the position of the pointer in the image. Different levels of abstraction must be involved from pixel level to object level passing by a segmentation level. High-level sonification permits to tackle object and semantic content while low-level handle all other abstract information like the luminance, gradient, color, texture, edge and shape. The challenging part on touch screen is to combine both levels of sonification without any additional hap-tic or tactile device.

3.1 High-level sonification

Understanding the semantic content of an image is one of the important points in high-level image sonification system. There are usually three ways to handle such concern: image description, object detection with bounding box and semantic segmentation. Image description permits to understand the global context of the image without going into details. The image description of Fig. 1a would be for eg. *A person on bicycle in front of cars*. Such global description does not permit to the user to experience the details of the image as an object detection technique would do by indicating an approximative position and size of the objects identified using bounding boxes (Fig. 1b). A semantic segmentation (Fig. 1c)) goes beyond the two previous techniques by labelling all the pixel of the image, then permits to detect the shape and contours of each object.

For that reason, we exploit one of the advanced semantic segmentation algorithms available in the domain and developed by DeepLab [6] using deep convolutional networks. More specifically, we integrate its mobile version based on MobileNet [28] into our sonification mobile application called TalkingImage 2 that we will use for experimentation in Section 5. See Fig. 1c for example of segmentation result provided by DeepLab algorithm.

Most of methods [4, 16] translate the context information extracted from the image to voice using a Text to Speech engine but we think that in a pure sonification experience, where eyes are replaced by ears, hearing a cow moo is better than hearing the

(a) Original (b) Objects detection (c) Semantic segmentation

Fig. 1 Context of an image

voice of person indicating the presence of a cow. Based on this assumption, we downloaded from FreeSound[1] a list of audio under a creative commons license, representing the cry, the interaction or the manifestation of each of the 20 classes of object of PASCAL Visual Object Class (VOC) 2012 dataset. PASCAL VOC challenge is a benchmark in visual object category recognition and detection, which provides the research community with a standard dataset of images and annotation, and standard evaluation procedures [18]. The train/validation data of 2012 has 11,530 images containing 27,450 regions of interest annotated objects and 6,929 segmentations. To cover as much as possible cries for some animals, we concatenated or mixed multiples sounds. For eg. we will hear the dog bark then pant while the cat will purr and meow. The length of each audio clip is 4s, sufficient for a human to identify environmental sounds with accuracy [8]

3.2 Low-level sonification

Low-level sonification strive to describe all other abstract visual content not covered by the high level sonification like the color, the gradient, the luminosity, the texture of an object. It is complementary to the high-level sonification since after the identification of an object, it permits to sonify the color and all other characteristic of the object. For this purpose, we will use the hierarchical feature-based approach developed in [33] which translate using musical notes, most of the properties of an image in the audio domain, in a very predictable way.

The hierarchical feature-based approach supposes that the original image is preliminary segmented into regions and since we do not have ground-truths for the new dataset we are proposing in Section 4, we use an automatic segmentation based on label fusion [22] which obtains a segmentation score, in terms of the Rand Index equals to 0.81 on BSD300 [18] dataset. BSD300 is a segmentation dataset of 300 color images of size 481×321 provided by the university of Berkeley and divided into a training set of 200 images, and a test set of 100 images. A set of benchmark segmentation results provided by human observers are available for each image and used as ground truth to quantify the reliability of the proposed segmentation algorithm. See the result of such segmentation on Fig. 2.

The second phase is to use the HSL color model for the mapping to audio domain because it easily describes the human perception of color with words or simple concepts such as Hue, Saturation and Luminance. [24]. Considering the HSL color space, the Hue of each

[1] https://freesound.org/

segmented region is mapped to the pitch of the generated sound by a system of 7 bins quantization mapped to the 7 musical notes of a piano at the octave C4 Middle C. The first, second, ..., seventh values of the histogram are converted to an impulse function centered on the frequency corresponding to the different notes of the musical game, *i.e.*, 262Hz (Do), 294Hz (Re), 330Hz (Mi), 349Hz (Fa), 392Hz (Sol), 440Hz (La), and 494Hz (Si), respectively (represented by the white keys of a piano keyboard for the octave C_4 Middle C and determine the *pitch* of the generated sound.

The luminance or bright depends on the amount of energy that is being radiated thus was translated to the level of octave or vibration of a piano. The luminance value, initially in the interval $[0 - 255]$ is divided into 8 equal intervals, and each interval is assigned the name of an octave scale C_n.

The saturation which represents the amount of white a color contains and generally used to describe the purity of a color is converted to the purity of sound. More or fewer harmonics, depending on the saturation value are added to the octave of the original sound thus make the sound more or less pure.

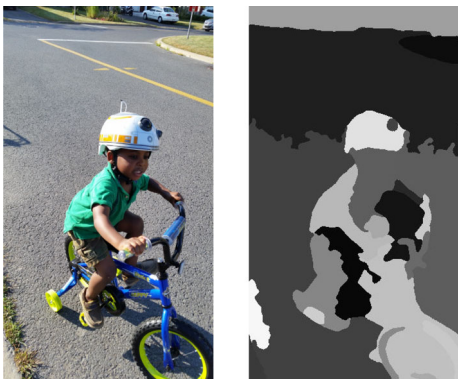
The roughness of the texture is translated to the rhythm and the loudness of the sound by adding distortion in phase and in magnitude to the original signal. See the visual to audio mapping and calibration in Fig. 3 extracted from [33] where a complete description of the algorithm is available.

In this paper we will generate two types of audio: one that contains only the color information translated to pitch, octave and purity then the second one that contains the full visual features.

3.3 Touch screen interaction

Conveying the information produced by the low-level and high-level sonification using only touch screen action without any additional haptic or tactic device is one of the tricky parts of our sonification system. In order to develop a system that can be easily used on a touch screen mobile device as on a touch screen desktop and also on a desktop with a classic mouse, we find it simple to limit triggering actions to slide (mouse move), tap (mouse left-click), hold (mouse right-click).

In the final model, the semantic segmentation (from high-level) and segmentation (from low-level) are superimposed. On the slide action within a semantic-segmented zone tagged



(a) Original (b) Segmentation

Fig. 2 Image segmentation

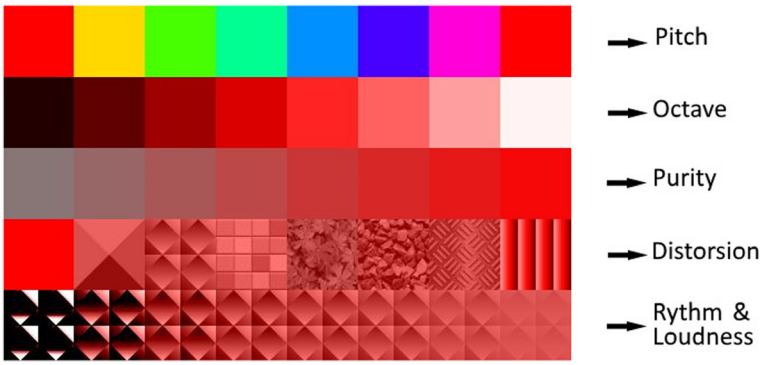


Fig. 3 Visual to audio mapping calibration and training image

with an object label, the non-speech audio of the class of object is played. If labelled as background, the generated audio containing full visual features of the segmented region is played. On a hold action within a segmentation zone, a generated audio conveying the color information is played then on a tap, the audio conveying full visual features information is played. The system interaction flow is represented in Fig. 4. With this interaction flow, the user can identify an object, gets the color information of different parts of the object then going beyond with gradient and texture information.

4 Dataset

One of the first obstacles encountered during our research on image sonification is the availability of a dataset on which to evaluate our model. Most of research in the domain used set of images based on the specific feature they want to sonify. It is then difficult to

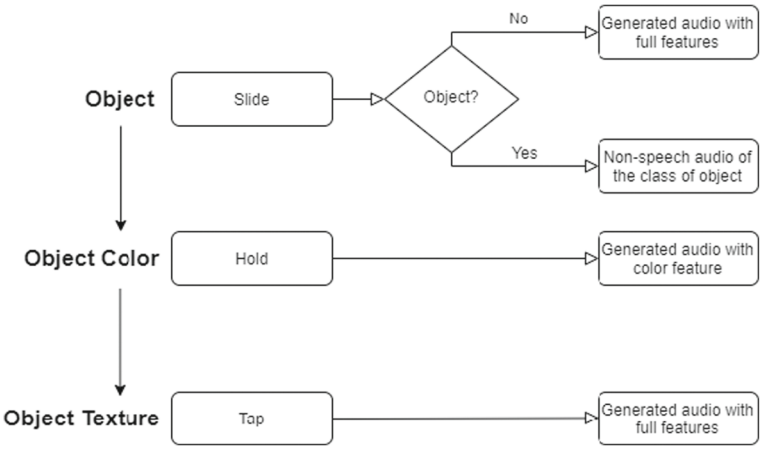


Fig. 4 User interaction flow

have a dataset that cover multiple features from low-level to high-level sonification. In [33], a subset of the BSD300 [18], a dataset normally used for image segmentation and boundary detection study, is used but it is not suitable for semantic segmentation.

In this paper, we come up with a dataset[2] of images that covers different features used in an image sonification system like color, gradient, texture, segmentation, and semantic segmentation. Some images come from our previous research on low-level sonification [33] while others come from personal library and Flick data under creative common licence for a total of 122 images. The height and width of the images were reduced to a maximum of 320 pixels because of the requirements of the segmentation algorithm used [22].

Primary colors and multiple colors variation in terms of hue, saturation and value are present in the dataset. Figure 5 presents the mosaic of images of abstract content of the dataset in terms of colors variation but also luminosity, gradient, texture and roughness of the texture. In Fig. 6, a minimum of three images per class of object are presented with a variation of colors, texture and number of instances for a total of 20 classes of objects aligned on PASCAL VOC 2012 [18]. Such diversity in the datasets offers the possibility to evaluate the detection of different objects, the evaluation of the color, luminosity, and texture of different objects of the same class, number of instance of objects...

5 Experimental results

We implemented the described model into an Android mobile application called TalkingImage 2 and made it tested by a group of persons who were part of the previous pilot study, then more experimented with the low-sonification part of our model [33]. All of them were volunteers composed of students and nonstudents and only one of them had a *musical ear*. Due to the Covid19 pandemic impact and the lack of availability of some persons, we were able to mobilize only six of the eleven peoples. Unlike the first phase of the study which took place in a [33] office, this second phase was carried out remotely. Subjects were instructed on how to use the application and were recommended to use a headset Three types of tests were proposed: object Identification, object's color Identification and scene exploration. In all the experiments the images were hidden to the user with a white or a blue screen and the images were displayed in the same order.

5.1 Experiment I : Object identification

The goal of this experiment was to test the semantic part of the sonification model by identifying different classes of objects present in the image. User was given 5 minutes to familiarize himself with the application by playing with the sonified version of the image of Fig. 1. Then he was given 5 minutes per image to explore the four images of Fig. 7 by sliding his finger and use the sound heard to name all the objects, part of the 20 classes, which are present. The results displayed in Table 1 shows that the users are able in most of the cases to identify the presence of the appropriate objects. However, the small resolution of the images (maximum of 320 pixels in width or height) negatively impacted the detection of a person's class because its instances cover a reduced number of pixels in the sample of images tested (Fig. 7b and d). Some peoples confused the sound of sheep to cow in the

[2] https://github.com/ohinitoffa/ImgSonficiation2

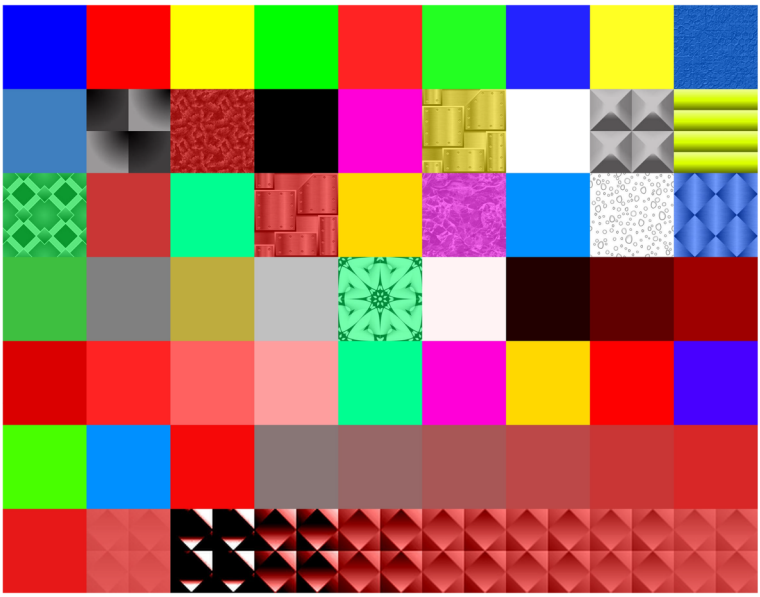


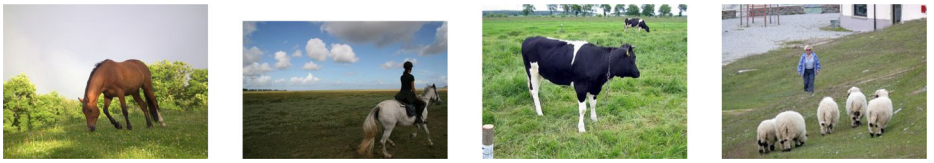
Fig. 5 Mosaic of colors and textures of the dataset

image Fig. 7d or the whistle of the person inside image Fig. 7b to a bird song. This means that we need to improve the audio of some objects to make them more discriminative.

5.2 Experiment II : Object's color identification

The users were given a few minutes to refresh their memory with the calibration image. They were then given four images containing object with different colors (see Fig. 8). They must use the slide action to identify the object, then hold their finger to identify the color

Fig. 6 Mosaic of objects of the dataset



(a) (b) (c) (d)

Fig. 7 Experiment I: Object identification

Table 1 Results of experiment I

Image	Ground truth	Answers
(a)	horse	horse (83.33%)
(b)	horse, person	horse (83.33%), person (50%)
(c)	cow	cow(100%)
(d)	sheep, person	sheep (60%), person (0%)

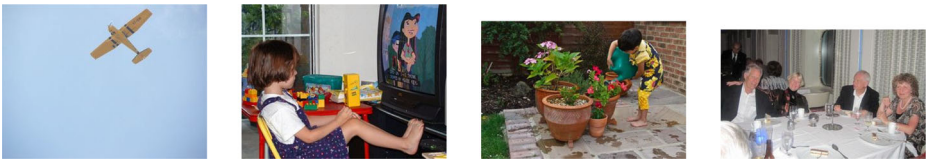
(a) (b) (c) (d)

Fig. 8 Experiment II: Object's Color Identification

Table 2 Results of experiment II

Image	Ground truth	Answers
(a)	bottle red/white	83.33%
(b)	bird blue/orange	60%
(c)	bird green/dark	50%
(d)	bird red	50%

of the object. The results of the experience reported in Table 2 shows that users are able to easily detect the color of a big object Table 2a with homogeneous pixels and low variation than one which is small Table 2d or contain too much color variation Table 2c.



(a) (b) (c) (d)

Fig. 9 Experiment III: Scene description

5.3 Experiment III : Scene description

The users were given 5 minutes per image to explore four images shown in Fig. 9, without further information. They had then to provide a description of the image based on the audio feedback. The goal is to see if their description fits the content of the images. Five of the six subjects were able to complete this test and their results are displayed in Table 3. As we can observe, a simple image with an aeroplane in the air (Fig. 9a) was easy to describe. The girl watching a television (Fig. 9b) was ambiguous since the sound of the TV could be interpreted as a musical concert. Once again, the decision to use a person whistling sound to represent the non-speech sound of a person's class was a bad idea since some subjects will continue to confuse it with the song of a bird. While the scene of a person watering a potted plant (Fig. 9c) was globally understood by most of the subjects, the image of persons around a dinning table (Fig. 9b) was the most difficult to describe. Subjects claimed that

Table 3 Results of experiment III

Image	Subject	Description
(a)	1	Airplane
	2	Sound of airplane flying in the air.
	3	Airplane dark yellow.
	4	An airplane in the air.
	5	An airplane in the air.
(b)	1	Television, I hear the music playing while exploring.
	2	A person playing music.
	3	Bird and musical instrument.
	4	A music concert with spectators.
	5	Bird in the water.
(c)	1	Sound of pouring water (probably someone watering): potted plant.
	2	A person pouring water.
	3	Bird near a water
	4	Liquid that is poured into a glass.
	5	Water pouring.
(d)	1	Unable to understand the scene
	2	Group of persons, probably in meeting.
	3	Birds and plates on the table.
	4	Unable to understand the scene.
	5	Whistle of bird.

the sound used to describe the dining table (a sound of dish placed on a table) was not clear enough to help them identifying the context of diner.

6 Conclusion

In this paper, we have presented a mid-level image sonification system that uses a non-speech audio dataset to describe the semantic content (20 classes of object) and a generated signal based on musical notes to describe the abstract content. We implemented our system in an Android application called Talking Image 2 and proposed an image dataset for evaluation.

The validation results showed that the subjects were generally able to identify the objects in the image, the color of the objects and even provide a general description of the scene of the image. However, the non-speech sound used for some classes was confusing and need to be improved. We learned that the choice of sound that represents each class is highly important in a system where a vocal description is not used. Our system is a prototype that can be greatly improved using a hybrid deep-learning model on the image instead of convolutional neural network (CNN) model. Such method achieved higher detection accuracy in [36] where the bearing vibration signals were converted into time-frequency images using the continuous wavelet transform (CWT), then a CNN was used to extract intrinsic fault features from the images and feed them into a gcForest classifier.

Declarations

Conflict of Interests

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.
- The authors obtained a certificate of ethics from the Université de Montréal to perform the pilot study.

References

1. Balakrishnan G, Sainarayanan G, Nagarajan R, Yaacob S (2008) A stereo image processing system for visually impaired. International Journal of Information, Control and Computer Sciences 2(9):1–10
2. Banf M, Blanz V (2013) Sonification of images for the visually impaired using a multi-level approach. In: Proceedings of the 4th augmented human international conference (AH '13), pp 162–169
3. Banf M, Mikalay R, Watzke B, Blanz V (2016) Picturesensation - a mobile application to help the blind explore the visual world through touch and sound. Journal of Rehabilitation and Assistive Technologies Engineering 3
4. Bartolome JI, Quero LC, Sunhee K, Um MY, Cho J (2019) Exploring art with a voice controlled multimodal guide for blind people. In: Proceedings of the Thirteenth international conference on tangible, embedded, and embodied interaction. TEI '19. Association for Computing Machinery, New York, NY, USA, pp 383–390. https://doi.org/10.1145/3294109.3300994
5. Capp M, Picton P (2000) The optophone: An electronic blind aid. Engineering Science and Education Journal 9(3):137–143
6. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European conference on computer vision, pp 833–851

7. Chidester B, Do M (2013) Assisting the visually impaired using depth inference on mobile devices via stereo matching. In: 2013 IEEE International conference on multimedia and expo workshops (ICMEW), pp 1–6. https://doi.org/10.1109/ICMEW.2013.6618381
8. Chu S, Narayanan S, Kuo C-CJ (2009) Environmental sound recognition with time–frequency audio features. ieee Transactions on Audio, Speech and Language Processing 17
9. Degara N, Hunt A, Hermann T (2015) Interactive sonification [guest editors' introduction]. IEEE MultiMedia 22(1):20–23. https://doi.org/10.1109/MMUL.2015.8
10. Dubus G, Bresin R (2013) A systematic review of mapping strategies for the sonification of physical quantities. PLoS ONE 8(12):82491
11. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A (2015) The pascal visual object classes challenge: A retrospective. Int J Comput Vis 111(1):98–136
12. Gotzelmann T (2018) Visually augmented audio-tactile graphics for visually impaired people. ACM Trans Access Comput 11(2). https://doi.org/10.1145/3186894
13. Goudarzi V (2015) Designing an interactive audio interface for climate science. IEEE MultiMedia 22(1):41–47. https://doi.org/10.1109/MMUL.2015.4
14. Ivan K, Radek O (2008) Hybrid approach to sonification of color images. In: Proceedings of the international conference on convergence and hybrid information technology
15. Kane SK, Morris MR, Wobbrock JO (2013) Touchplates: Low-cost tactile overlays for visually impaired touch screen users. In: Proceedings of the 15th International ACM SIGACCESS conference on computers and accessibility. ASSETS '13. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2513383.2513442
16. Kwon N, Koh Y, Oh U (2019) Supporting object-level exploration of artworks by touch for people with visual impairments. In: The 21st international ACM SIGACCESS conference on computers and accessibility. ASSETS '19. Association for Computing Machinery, New York, NY, USA, pp 600–602. https://doi.org/10.1145/3308561.3354620
17. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539
18. Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc of the 8th international conference on computer vision (ICCV), vol 2, vancouver, British Columbia, Canada, pp 416–423
19. Martins ACG, Rangayyan RM, Ruschioni RA (2001) Audification and sonification of texture in images. J Electronic Imaging 10(3):690–705
20. Matta S, Kumar DK, Yu X, Burry M (2004) An approach for image sonification. In: First international symposium on control, communications and signal processing, 2004, pp 431–434
21. Meijer PBL (1992) An experimental system for auditory image representations. IEEE Trans Biomed Eng 39(2):112–121
22. Mignotte M (2014) A label field fusion model with a variation of information estimator for image segmentation. Inform Fusion 20:7–20
23. Morris MR, Johnson J, Bennett CL, Cutrell E (2018) Rich representations of visual content for screen reader users. In: Proceedings of the 2018 CHI conference on human factors in computing systems. CHI '18. Association for Computing Machinery, New York, NY, USA, pp 1–11. https://doi.org/10.1145/3173574.3173633
24. Munsell AH (1912) A pigment color system and notation. J Psychol 23(2):236–244. https://doi.org/10.2307/1412843
25. Oh U, Joh H, Lee Y (2021) Image accessibility for screen reader users: A systematic review and a road map. Electronics 10(8). https://doi.org/10.3390/electronics10080953
26. Quero LC, Bartolome JI, Lee S, Han E, Kim S, Cho J (2018) An interactive multimodal guide to improve art accessibility for blind people. In: Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility. ASSETS '18. Association for Computing Machinery, New York, NY, USA, pp 346–348. https://doi.org/10.1145/3234695.3241033
27. Rodrigues JB, Ferreira AVM, Maia IMO, Junior GB, de Almeida JDS, de Paiva AC (2019) Image processing of artworks for construction of 3d models accessible to the visually impaired. In: Advances in manufacturing, production management and process control. Springer, Cham, pp 243–253
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, pp 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
29. Scavaco S, Henriques JT, Mengucci M, Correia N, Medeiros F (2013) Color sonification for the visually impaired. In: Cruz-Cunha MM, Varajão HKJ, Martinho R (eds) Proceedings of international conference

on health and social care information systems and technologies (HCist). Procedia Technology, Elsevier, ???, pp 1048–1057
30. Schaffert N, Mattes K (2015) Interactive sonification in rowing: Acoustic feedback for on-water training. IEEE MultiMedia 22(1):58–67. https://doi.org/10.1109/MMUL.2015.9
31. Sudol J, Dialameh O, Blanchard C, Dorcey T (2010) Looktel, a comprehensive platform for computer-aided visual assistance. In: 2010 IEEE computer society conference on computer vision and pattern recognition - workshops, pp 73–80. https://doi.org/10.1109/CVPRW.2010.5543725
32. Tajadura-Jiménez A, Bianchi-Berthouze N, Furfaro E, Bevilacqua F (2015) Sonification of surface tapping changes behavior, surface perception, and emotion. IEEE MultiMedia 22(1):48–57. https://doi.org/10.1109/MMUL.2015.14
33. Toffa OK, Mignotte M (2020) A hierarchical visual feature-based approach for image sonification. IEEE Transactions on Multimedia 23:706–715. https://doi.org/10.1109/TMM.2020.2987710
34. Winters RM, Joshi N, Cutrell E, Morris MR (2019) Strategies for auditory display of social media. Ergon Des 27:11–15
35. Wu X, Li Z-N (2008) A study of image-based music composition. In: 2008 IEEE International conference on multimedia and expo, pp 1345–1348. https://doi.org/10.1109/ICME.2008.4607692
36. Xu Y, Li Z, Wang S, Li W, Sarkodie-Gyan T, Feng S (2021) A hybrid deep-learning model for fault diagnosis of rolling bearings. Measurements 169:108502. https://doi.org/10.1016/j.measurement.2020.108502
37. Yeo WS, Berger J (2006) Application of raster scanning method to image sonification, sound visualization, sound analysis and synthesis. In: Proceedings of the Int Conf on digital audio effects (DAFx-06), Montreal, Quebec, Canada, pp 309–314
38. Yoshida T, Kitani KM, Koike H, Belongie S, Schlei K (2011) Edgesonic: Image feature sonification for the visually impaired. In: Proceedings of the 2Nd augmented human international conference. AH '11. ACM, New York, NY, USA, pp 11–1114
39. Zhao Y, Wu S, Reynolds L, Azenkot S (2017) The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. In: Proc ACM Hum-Comput Interact 1(CSCW). https://doi.org/10.1145/3134756

Affiliations

O. K. Toffa[1,2] · M. Mignotte[1,2]

M. Mignotte
mignotte@iro.umontreal.ca

[1] Vision lab. of the Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal Montréal, H3C 3J7, QC, Canada

[2] Faculté des Arts et des Sciences, Université de Montréal, Montréal, H3C 3J7, QC, Canada