Sensors and Materials, Vol. 33, No. 11 (2021) 4057–4068
MYU Tokyo

S & M 2745

Interactive Sound Generation to Aid Visually Impaired People via Object Detection Using Touch Screen Sensor

Tias Kurniati,[1] Chuan-Kai Yang,[2] Tzer-Shyong Chen,[3*]
Yu-Fang Chung,[4] Yu-Min Huang,[1] and Chih-Cheng Chen[5**]

[1]Department of Statistics, Tunghai University,
No. 1727, Section 4, Taiwan Blvd., Xitun District, Taichung City 407224, Taiwan
[2]Department of Information Management, National Taiwan University of Science and Technology,
No. 43, Section 4, Keelung Rd., Da'an District, Taipei City 106, Taiwan
[3]Department of Information Management, Tunghai University,
No. 1727, Section 4, Taiwan Blvd., Xitun District, Taichung City 407224, Taiwan
[4]Department of Electrical Engineering, Tunghai University,
No. 1727, Section 4, Taiwan Blvd., Xitun District, Taichung City 407224, Taiwan
[5]Department of Automatic Control Engineering, Feng Chia University,
No. 100, Wenhua Rd, Xitun District, Taichung City 407, Taiwan

(Received July 1, 2021; accepted October 25, 2021)

Keywords: sound generation, touch sensor, object detection, image processing

Along with the invention of portable devices, such as smartphones and tablets, computer-based touch screen sensor-assistive technologies have become significantly more affordable than traditional tactile graphics. The sensor panel in these technologies allows users to receive visual and auditorial responses via interaction with the device. However, visually impaired individuals (with a lack or loss of ability to see) will not find visual responses useful when using tablets or smartphones. Therefore, in this paper we propose a system that helps visually impaired people comprehend information on electronic devices with the help of auditory action feedback. We develop a multimedia system for sound production from a given image via object detection. In this study, YOLO (You Only Look Once) is used in object detection for sonification. A pre-trained model is used; thus, a wider range of object classification can be identified. The system generates the corresponding sound when an object on the sensor screen is touched. The purpose of our research is to aid visually impaired people to perceive information of a picture shown on the device by touching the detected object. The device was tested by simulating visually impaired people by blindfolding people with normal vision, who filled out questionnaires on its performance. The results indicate that most of the users found that the sound presented by the device was helpful for telling them what the shown image was.

1. Introduction

All our understanding of the world is built upon our ability to process information through the five senses. All of these senses are important; however, hearing is one of the most essential

*Corresponding author: e-mail: arden@thu.edu.tw
**Corresponding author: e-mail: chenccheng@fcu.edu.tw
https://doi.org/10.18494/SAM.2021.3528

ISSN 0914-4935
https://myukk.org/

and fundamental senses in obtaining information about the world. Sound is well known for its important role in the way we perceive and interact with the environment. Regarding vision, the World Health Organization (WHO) estimates that there are 285 million visually impaired people in the world, out of which 39 million are blind.[1] Two common ways of presenting information to help visually impaired people are tactile graphics and sonification. However, these days people use electronic devices and screens instead of printed braille to access text and pictures, especially the younger generation.

In this paper, we propose a multimedia system for sound production from a given image based on object detection utilizing the You Only Look Once (YOLO) deep learning method. Firstly, a random image is uploaded into the system. Immediately after it appears on the touch screen sensor, users can hear a sound associated with the image by touching the object. A pre-trained model is used in the system, so a more extensive range of objects can be recognized. Once an image is uploaded to the system, it is run through a single convolutional network to detect multiple bounding boxes and class probabilities, then weights are used to optimize the predicted bounding boxes and send back the picture with a bounding box to the system. This technology helps visually impaired people perceive information of a picture shown on the sensor device with the use of touch. For example, Fig. 1 illustrates the difference in how the system handles pictures of an object with and without an associated sound. The sound of a cat (meow) is generated when the user touches the cat shown on the sensor screen. Meanwhile, for the chair, a silent object, the system pronounces the word of the object automatically and the word "chair" is spoken by the system. A user study showed that more than 90% of users can recognize objects correctly.

This paper is organized as follows. In Sect. 2, we explain some of the previous related works, while in Sect. 3, we discuss the system methodology. In Sec. 4, we present the experimental results of the system. Finally, in Sect. 5, we conclude the paper and discuss some potential future works.

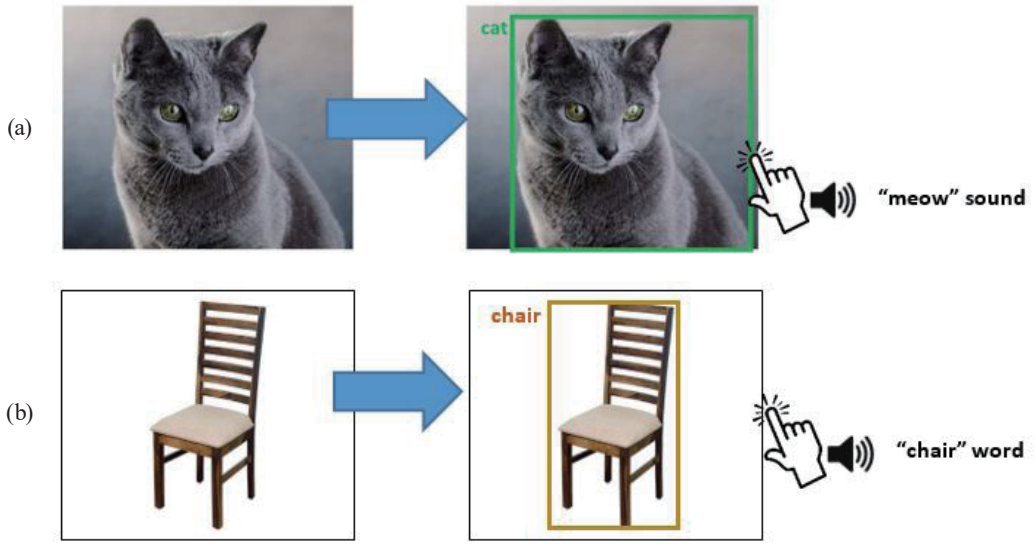

Fig. 1. (Color online) (a) Picture of an object with a sound. (b) Picture of an object without a sound.

2. Related Works

Object detection has been a popular area of research over the past few decades, as indicated by the large number of new applications related to identifying objects based on visual detection, such as facial expression recognition,[2] navigation assistants,[3] autonomous robot navigation,[4] self-driving systems,[5] image recognition,[6] and pedestrian detection.[7]

Many approaches have been used to assist visually impaired people in obtaining information on a digital platform. A software package called PLUMB for a tablet PC was developed by Cohen *et al.*[8] They used audio feedback and a pen-based interface to relay information on graphs from the start vertex to the finish vertex. The authors used the variations of the loudness and vibration intensity based on the HSB color model in the elements of the graphs. Hence, the system was unable to guarantee real-time audio feedback, which caused discomfiture in users when the graph information was conveyed.

Wörtwein *et al.*[9] proposed image sonification through a mobile, interactive, and web-based approach. To evaluate the approach, visually impaired users were given three tasks to complete: mathematical graph identification, proportion estimation in bar charts, and pathfinding on floor maps. However, there were some drawbacks related to the system. The implemented system could not guarantee real-time reactions, which sometimes irritated users trying to identify the shapes and details of an image through mouse/finger movement.

Cavaco *et al.*[10] developed a software tool that assisted visually impaired people in identifying the color and luminosity of an image through image sonification. The software tool extracted the color information of an image or video by extracting HSV (hue, saturation, value) information, which was converted into the audio attributes of pitch, timbre, and loudness, respectively. Nonetheless, the audio generated only seven colors within one musical octave, which was insufficient for users to differentiate the sounds for different colors. This resulted in an overall correct answer rate of only 60% for two subjects who were born blind and 48.33% for the remaining subjects.

More sonification-based studies were presented by Yoshida *et al.*[11] and Krishnan *et al.* [6] They proposed a framework to assist visually impaired users in recognizing an object in an image according to image edge features and distance-to-edge maps by transforming basic object shapes into sounds. The system was implemented on the touch screen of a mobile device, allowing users to explore the image content by moving their fingers over the screen.

A multilevel approach to the sonification of images was also developed in 2013 by Banf and Blamz.[12] They presented a system to help visually impaired people obtain direct perceptual access to images via acoustic signals. Users explored an image actively on a touch screen then received auditory feedback about the image content at the current position. In the system, low-level information (color, edges, and roughness) was combined with mid-level and high-level information from machine learning algorithms. For object recognition, the OpenCV library was employed, which enabled us to implement a Bag of Visual Words as well as perform support vector machine (SVM) detection and localization algorithms. Then, both algorithms were trained on the 20 object classes provided by the Visual Object Classes Challenge 2008 (VOC2008). The experimental results indicated that the system was useful and could help

visually impaired people access the content of an image. However, the auditory feedback was limited to an acoustic rhythm such as a drum, bass, or noise, whereas a wider range of acoustic feedback is desirable.

The most recent research related to the use of deep learning was presented by Saldana and Mendizabal-Ruiz.[13] They synthesized sound from a geometrical image using a method based on a convolutional neural network (CNN). The process started from a deep learning network that learned how to associate a pattern in an image with a sound, then some parameters were set to generate the sound. For each modification of the image pattern, the system produced a different sound.

Cardillo *et al.* proposed an electromagnetic sensor to assist visually impaired people become aware of obstacles surrounding them in a range wider than that provided by a traditional cane by mounting a microwave radar on a traditional cane.[14] Patil *et al.* presented a NavGuide system, which was implemented in shoes, to help visually impaired people navigate outdoors by informing the user of the surrounding environment.[15] The feedback utilized vibration and audio feedback mechanisms. These two works mainly focused on helping visually impaired people to navigate.

In many of these previous studies, systems were developed to assist visually impaired people in different ways. In contrast, the purpose of our system is to help visually impaired people identify objects, especially when they are presented as pictures on smartphones or tablets. Therefore, an image sonification platform that can produce the corresponding sound feedback of an object detected is needed to assist visually impaired people in retrieving information.

3. System Overview

The system architecture is illustrated in Fig. 2. Object detection in our system is based on the YOLO framework. We developed the system under the webserver node.js associated with YOLO. As shown in Fig. 2, before the image is processed by YOLO, it is uploaded and sent to node.js to be stored permanently. Then, the image is sent to the YOLO system to run bounding box prediction, class prediction, prediction across scales, and feature extraction, resulting in multiple bounding boxes and class probabilities. Then, weights are used to maximize the predicted bounding boxes and confidence, then the final detection of the region and class is performed. The resulting image with bounding boxes is then sent back to the interface. The

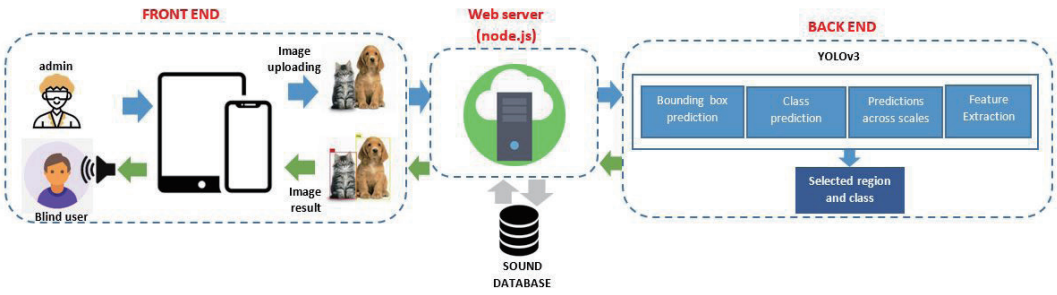

Fig. 2. (Color online) System architecture of object detection sonification.

sound corresponding to the object is played when the object on the screen sensor panel is touched.

3.1 Object detection

The image classification procedures in the sonification approach are next explained in Sect. 3.1, then sound generation is discussed in Sect. 3.2.

3.1.1 Bounding box prediction

As shown in Fig. 3, an image is first divided into $A \times A$ grids, where each grid predicts the number of bounding boxes, B, and confidence scores for the boxes and C class probabilities. These confidence scores reflect how confident the model is of predicting the content of a single box and also how accurate it is in predicting an object. Here, we define the confidence as $c(object) \times IOU^{truth}_{pred}$, where IOU refers to intersection over union. If no object exists in the cell, the confidence score is zero.

Each bounding box involves five predictions, namely, x, y, w, h, and confidence, where (x, y) represents the coordinates of the box related to the grid cell and (w, h) represents the width and height of the whole image. Finally, the confidence score in the predictions is encoded as an $A \times A \times (B \times 5 + C)$ tensor.[16]

3.1.2 Class prediction

Each grid cell predicts C class probabilities, $c(Class_i|Object)$, given by Eq. (1). These probabilities are applied to each grid cell containing an object. Each set of class probabilities per grid cell is predicted, regardless of B.

$$c(Class_i|Object)c(Object) * IOU^{truth}_{pred} = c(Class_i) * IOU^{truth}_{pred} \quad (1)$$

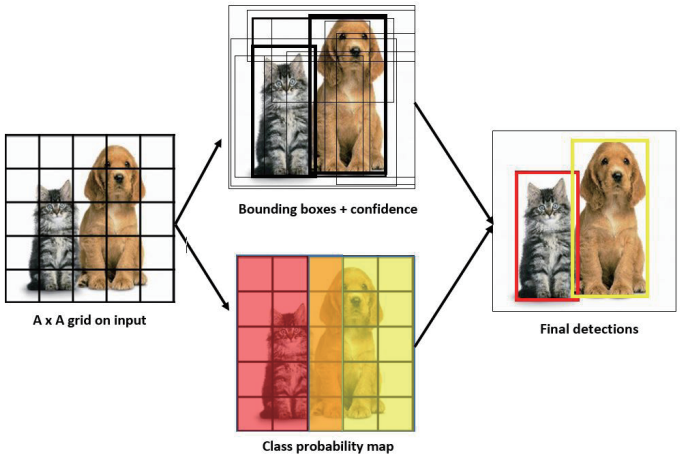

Fig. 3. (Color online) Model of image classification.

3.1.3 Prediction across scales

YOLO predicts boxes at three different scales: bounding box, object identification, and class prediction. The system extracts features from those scales using a similar concept to feature pyramid networks (FPNs) as shown in Fig. 4.[17] Compared with other architectural systems, the FPN shows a significantly higher level of performance for testing images. The FPN comprises bottom-up and top-down pathways, where the bottom-up pathway employs a CNN for feature extraction, while the top-down pathway decreases the spatial resolution. When the higher-level structures of the image are detected, the semantic value (where the object is assigned to class prediction) for each layer is increased. Then, the same feature is performed one more time to obtain the final scale of the predicted boxes.

3.1.4 Feature extractor

Darknet-53, the backbone of YOLO, is used to extract features. Darknet-53 is a CNN with 53 layers, which allows the system to load a pre-trained version of the network with millions of image datasets. The pre-trained network classifies thousands of objects. As a result, objects can be detected from a wide range of images. This will allow the system to help visually impaired people detect a wide range of objects in pictures. The network uses successive 3 × 3 and 1 × 1 convolutional layers with an image input size of 256 × 256.[18] In addition, the network also adds a residual structure, which sets up a shortcut link between several layers so that it can increase the depth of the network without reducing its accuracy and solve the problem of gradient explosion or disappearance that can easily occur owing to the excessive depth of the network. [19] A list of the network framework parameters of Darknet-53 is shown in Table 1.

3.2 Sound generation

The system produces sounds in two ways: by producing a suitable sound for an object and by pronouncing the word of the object. The sounds used for object detection are downloaded from https://www.freesound.org and stored in a local database until called by the system. Users hear the corresponding sound when they touch the bounding box of the detected object shown. The interface of the system is shown in Fig. 5.

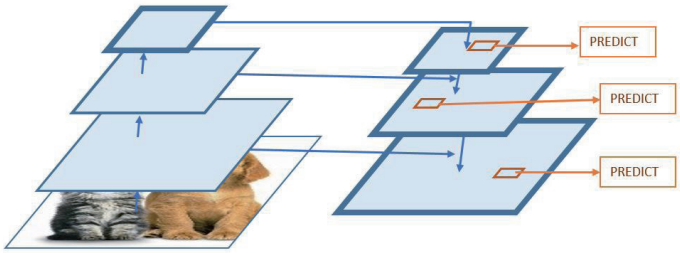

Fig. 4. (Color online) Feature pyramid network.

Table 1
Network parameters of Darknet-53.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3/2	128 × 128
	Convolutional	32	1 × 1	
1×	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3/2	64 × 64
	Convolutional	64	1 × 1	
2×	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3/2	32 × 32
	Convolutional	128	1 × 1	
8×	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3/2	16 × 16
	Convolutional	256	1 × 1	
8×	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3/2	8 × 8
	Convolutional	512	1 × 1	
4×	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

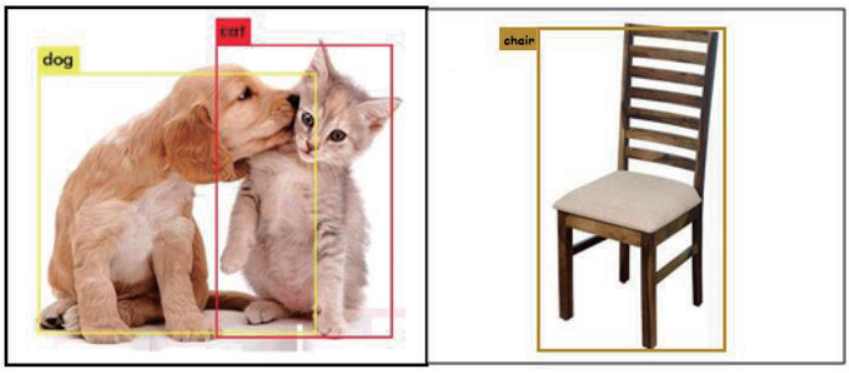

Fig. 5. (Color online) Interface of object detection through sonification.

As depicted in Fig. 5, each object has its own bounding box. Once the user touches any of the bounding boxes, the system calls the sound in accordance with the label name of each bounding box. For example, when users touch the bounding box of the cat or dog, the meow of a cat or the bark of a dog is emitted by the system, and if an object without a sound is touched, the word of the object is spoken.

4. Experimental Results and Discussion

4.1 Experiments

The system was developed using JavaScript and C++, both of which are compatible with node.js. We run the system on an ASUS machine powered by an Intel Core i7-6700 Processor with 3.40 GHz, 32 GB memory and an NVIDIA GeForce GTX-1080 Ti 11 GB graphics card. We use Ubuntu as the operating system as it is an open source. The specifications of the PC used are given in Table 2. The interface is viewed using Google Chrome, whose local host system is tunneled into a temporary URL, so that all units of the sensor panel are accessible in real time.

4.1.1 Participants

A total of 20 blindfolded volunteers aged between 20 and 40 participated in the evaluation and were recruited through invitation. Most of the participants were graduate students from various university departments. They consisted of nine females and eleven males. All participants had normal hearing.

4.1.2 Environment

During the user study, the participants entered the room equipped with a computer, a sensor screen unit, and a headband to cover their eyes. Once the eyes of the participants were covered, the researcher uploaded a random image to the system and asked the participants to hold the sensor screen unit and touch the sensor screen with their fingers. Once their fingers hovered on the image, they heard the corresponding sound or spoken word of the image, then guessed the name of the detected object. Five images, including images containing objects with and without a sound, were presented for each participant.

4.1.3 Questionnaire

Immediately after the experiment, participants were asked to check the correct answers. Then, they filled in a questionnaire to evaluate the system regarding how well it detected objects. The Likert scale was adopted to classify each evaluation as a score from 1 to 10, where 1 indicated "very inappropriate" and 10 indicated "very appropriate". The questionnaire was divided into two sections. The questions in the first section asked about participants' details

Table 2
Specifications of PC used.

Device type	PC desktop (server system)	Touch sensor panel (interface system)
CPU	Intel Core i7-6700, 3.40 GHz	2 × 2.15 GHz + 2 × 1.6 GHz
Memory	32 GB	4 GB
GPU	NVIDIA GTX 1080 Ti 11 GB	Qualcomm Adreno 530
Operating system	Ubuntu 16.04 LTS 64-bit	Android 7.0 Nougat

including gender, age, and hearing condition, and those in the second section asked them to evaluate the system of object detection through sonification.

4.2 Results

The overall percentage of correct answers was 98% (Fig. 6). Incorrect answers were reported to be due to the network error and the very high resolution of some images, which resulted in a long loading time and bounding boxes not appearing during the test. Users reported that when the network was stable and the uploaded image had a normal resolution, they were able to give the correct answer.

As we can see in Fig. 7, most of the users agreed that the sound was appropriate by giving a score of 8 out of 10 or higher, and none of them gave a score of less than 5. This indicates that most participants were satisfied with the performance of the system for object detection through sonification.

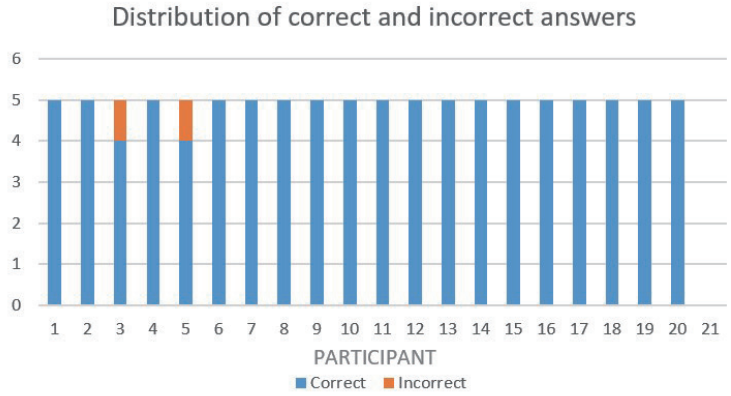

Fig. 6. (Color online) Distribution of correct and incorrect answers.

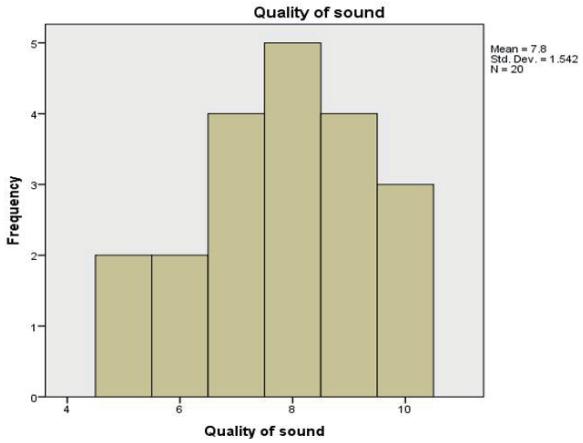

Fig. 7. (Color online) Users' perceptions of object detection appropriateness.

Most users agreed that the presented sound or the spoken word matched the object detected. Only two users thought that the sound did not match the image. This was because during the testing, sometimes bounding boxes were not shown, so the sound could not be heard, which occurred when a large image with high resolution was uploaded and a longer time was required to process the image.

4.3 Discussion

Many systems have been developed to help visually impaired people perceive information of images on a screen.[6,9–13] Some previous systems also implemented a deep learning methodology;[4,13,14,19] however, none of them used deep learning to detect objects in images. Therefore, the novelty of our system is that a deep learning method is implemented for object detection, then the sound or word of the object is presented to the user. In the future, we will improve the system so that it can process images with both low and high resolutions. It may also be possible to incorporate this technology into wearable IoT devices to be implemented for the recognition of a wider range of images to help visually impaired people live independently.

5. Conclusions

In this study, we have presented an interactive sonification platform using a touch screen sensor to help visually impaired people access digital content on electronic devices. To evaluate the system, we performed a user study involving blindfolded participants. Results showed that 98% of the answers given by participants were correct and matched the object detected by the system. Among the participants, 90% of them also agreed that our system produced the associated sound or spoken word associated with the detected object. The proposed system is a new medium for helping visually impaired people distinguish objects shown on the touch screen of electronic devices, such as smartphones and tablets. The system has the potential to provide more features to assist visually impaired people. We also hope that in the future, this technology will be incorporated into wearable devices and used to recognize a wider range of images. For instance, the utilization of wearable devices with a small camera to detect objects could help visually impaired people more easily live independently.

Acknowledgments

This work was partially supported by the Ministry of Science and Technology of Taiwan, R.O.C., through Grant No. 109-2221-E-029-019.
Funding: This research received no external funding.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available because further study will be carried out using the same data.
Conflicts of Interest: The authors declare no conflicts of interest.

References

1 World Health Organization : https://www.who.int/blindness/publications/globaldata/en/ (accessed May 2021).
2 H. Kusuma, M. Attamimi, and H. Fahrudin: Bull. Electr. Eng. Inf. **9** (2020) 1208. https://doi.org/10.11591/eei.v9i3.2030
3 Y.-Z. Hsieh, S.-S. Lin, and F.-X. Xu: Multimed. Tools Appl. **79** (2020) 29473. https://doi.org/10.1007/s11042-020-09464-7
4 M. Luong and C. Pham: J. Int. Rob. Syst. **101** (2021) 1. https://doi.org/10.1007/s10846-020-01262-5
5 D.-H. Wang, J. Li, and S. Zhu: Neurocomputing **428** (2021) 361. https://doi.org/10.1016/j.neucom.2020.02.128
6 K. G. Krishnan, C. M. Porkodi, and K. Kanimozhi: Proc. 2013 Int. Conf. Communications and Signal Processing (IEEE, 2013) 943–946. https://doi.org/10.1109/iccsp.2013.6577195
7 P. Dollar, C. Wojek, B. Schiele, and P. Peron: IEEE Trans. Pattern Anal. Mach. Intell. **34** (2011) 743. https://doi.org/10.1109/TPAMI.2011.155
8 R. F. Cohen, V. Haven, J. A. Lanzoni, A. Meacham, J. Skaff, and M. Wissell: Proc. 2006 ACM SIGACCESS Conf. Computers and Accessibility (IEEE, 2006) 119–124. https://doi.org/10.1145/1168987.1169008
9 T. Wörtwein, B. Schauerte, K. Müller, and R. Stiefelhagen: Proc. 2016 Int. Conf. Computers Helping People with Special Needs (ICCHP, 2016) 212–219. https://doi.org/10.1145/2818346.2823298
10 S. Cavaco, J. T. Henriques, M. Mengucci, N. Correia, and F. Medeiros: Proc. 2013 Int. Conf. Health and Social Care Information Systems and Technologies (HCIST, 2013) 1048–1057. https://doi.org/10.1016/j.protcy.2013.12.117
11 T. Yoshida, K. M. Kitani, H. Koike, S. Belongie, and K. Sclei: Proc. 2011 Augmented Human Int. Conf. (ACM, 2011) 1–4. https://doi.org/10.1145/1959826.1959837
12 M. Banf and V. Blanz: Proc. 2013 Augmented Human Int. Conf. (ACM, 2013) 162–169. https://doi.org/10.1145/2459236.2459264
13 R. Saldana and G. Mendizabal-Ruiz: Proc. 2020 17th Int. Conf. Electrical Engineering, Computing Science and Automatic Control (CCE, 2020) 1–5. https://doi.org/10.1109/CCE50788.2020.9299207
14 E. Cardillo, V. D. Mattia, G. Manfredi, P. Russo, A. D. Leo, A. Caddemi, and G. Cerri: IEEE Sens. J. **18** (2018) 2568. https://doi.org/10.1109/JSEN.2018.2795046
15 K. Patil, Q. Jawadwala, and F. C. Shu: IEEE Trans. Hum.-Mach. Syst. **48** (2018) 172. https://doi.org/10.1109/THMS.2018.2799588
16 J. Redmon, S. Divvala, R. Girshick, and A. Farhadi: Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2016) 779–788. https://doi.org/10.1109/CVPR.2016.91
17 T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie: Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2017) 936–944. https://doi.org/10.1109/CVPR.2017.106
18 J. Redmon and. A. Farhadi: Technical Report YOLOv3: An Incremental Improvement (2018). https://arxiv.org/abs/1804.02767
19 J. Cao, C. Song, S. Peng, S. Song, X. Zhang, Y. Shao, and F. Xiao: J. Sens. **20** (2020) 3646. https://doi.org/10.3390/s20133646
20 YOLO: Real-Time Object Detection: https://pjreddie.com/darknet/yolo (accessed May 2021).

About the Authors

Tias Kurniati received her Master's degree in information management from National Taiwan University of Science and Technology in 2018. She is currently a Ph.D. student at the Department of Statistics, Tunghai University, focusing on information management. Her main research interests are in deep learning, image and text recognition, computer graphics, and information security.

Chuan-Kai Yang received his Ph.D. degree in computer science from Stony Brook University, USA, in 2002 and his M.S. and B.S. degrees in computer science and mathematics from National Taiwan University in 1993 and 1991, respectively. He is currently a professor of the Information Management Department, National Taiwan University of Science and Technology. His research interests include computer graphics, scientific visualization, multimedia systems, and computational geometry.

Tzer-Shyong Chen received his Ph.D. from the Department of Electrical Engineering (Computer Science), National Taiwan University, Taiwan, in 1996. He is currently the chair of the Department of Information Management, Tunghai University, Taiwan. He has served in an evaluation committee of the Institute of Electrical Engineering Taiwan and is a member of IEEE. He has authored/co-authored over 80 refereed publications. His main research interests are in information security, cryptography, and network security.

Yu-Fang Chung received her B.A. degree in English language, literature, and linguistics from Providence University in 1994, her M.S. degree in computer science from Dayeh University in 2003, and her Ph.D. degree in computer science from National Taiwan University, Taiwan, in 2007. She is currently a professor of the Department of Electronic Engineering and Information Management, Tunghai University, where she is involved in research on information security and cryptography.

Yu-Min Huang received her Ph.D. in statistics from the University of Minnesota-Twin Cities, USA. Currently, she is an assistant professor in the Department of Statistics, Tunghai University. Her research interests include time series, multivariate data analysis, network data analysis, quantitative modeling, and high-dimensional data.

Chih-Cheng Chen is a professor at Jimei University, China. He is a senior member of IEEE. He earned his Ph.D. degree in mechatronics engineering from National Changhua University of Education. He has published 43 articles and owns three patents. He is currently researching RFID application systems, AIoT, machine learning, and information security.