IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 17, 2020, accepted May 25, 2020, date of publication June 1, 2020, date of current version June 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998776

Audio-Processing-Based Human Detection at Disaster Sites With Unmanned Aerial Vehicle

YUKI YAMAZAKI[1], CHINTHAKA PREMACHANDRA[1], (Member, IEEE), AND CHAMIKA JANITH PEREA[2], (Member, IEEE)
[1]Department of Electronic Engineering, Shibaura Institute of Technology, Tokyo, Japan
[2]Center for Advance Robotics, University of Moratuwa, Moratuwa, Sri Lanka

Corresponding author: Chinthaka Premachandra (chintaka@sic.shibaura-it.ac.jp)

This work was supported in part by the Branding Research Fund of Shibaura Institute of Technology.

ABSTRACT This paper describes a human search system that uses an unmanned aerial vehicle (UAV). The use of robots to search for people is expected to become an auxiliary tool for saving lives during a disaster. In particular, because UAVs can collect information from the air, there has been much research into human search using UAVs equipped with cameras. However, the disadvantage of cameras is that they struggle to detect people who are hidden in shadows. To solve this problem, we mounted an array microphone on a UAV and to detect the human voice as a means of finding people that cameras cannot. Also a search method is proposed that combines voice and camera human detection to compensate for their respective shortcomings. The rate and accuracy of human detection by the proposed method are assessed experimentally.

INDEX TERMS Victim detection, rescue support, UAV application, sound source separation, voice recognition.

I. INTRODUCTION

In the event of a large-scale disaster such as an earthquake, it is expected that many people will go missing. In such situations, the survival rate is related directly to how long it takes to rescue the victims. However, because the number of people who can engage in rescue operations is limited, it is important to have an efficient means of obtaining information about victims. Against this background, in recent years, human search systems using robots have been actively investigated in order to improve the efficiency of rescue activities [1]–[3]. In particular, unmanned aerial vehicle (UAVs) that can search for people aerially have been developed for situations in which rescuers cannot access damaged locations directly [4]–[10]. Such detection helps rescuers to understand the situation at the disaster site, thereby facilitating rescue operations. Although such robots now make it possible to detect people visually, a drawback of this search method is that it remains difficult to detect people who are either in camera blind spots or hidden in shadows.

Therefore, to detect people more reliably, we have addressed these problems by using a UAV equipped with not only a camera for visual information but also a microphone for audio information. Fig. 1 shows the system schematically. A UAV equipped with a loudspeaker and a microphone hovers over the disaster site and broadcasts an audio request for a response from anyone below. The microphone detects the voice of anyone who responds, thereby determining whether there is anyone there who requires rescuing.

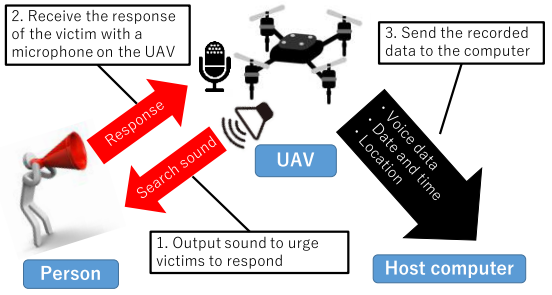

FIGURE 1. System overview.

However, a hovering UAV and a microphone are highly incompatible. First, the microphone picks up the sound of the proximate UAV propellers, thereby obscuring the person's voice. Second, the farther away the person, the fainter their voice. In this paper, to solve these two problems, we detect only the human voice by applying sound-source separation processing to the mixed sound of the recorded propellers and human voices.

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

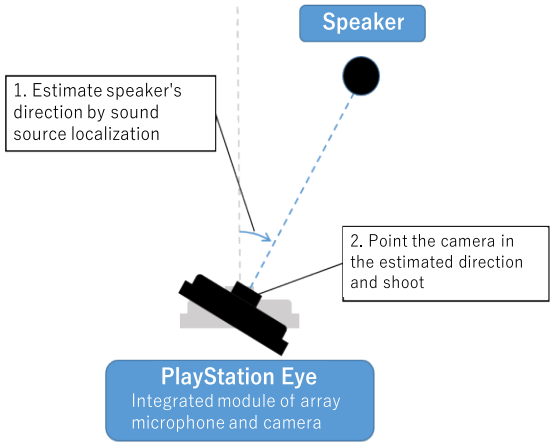

FIGURE 2. Relationship between array microphone and camera.

In sound-source separation by an array microphone, sound-source localization for estimating the direction of a human voice is performed. We use this localization result to point the UAV-mounted camera toward the sound source to photograph the person whose voice was detected. Fig. 2 shows the relationship between the camera and the microphone. This is an operational image of a PlayStation Eye, in which the array microphone and camera used in this study are integrated. By simultaneously performing human detection using a microphone and a camera on a single UAV, we constructed a system with high detection accuracy that takes advantage of voice recognition and image processing, as shown in Fig. 3. Here, in the proposed system, when the human voice is only acquired, the system detects availability of a human. On the other hand, the on-board camera can also be used to detect humans same as UAV-mounted camera based human detection systems in the literature.

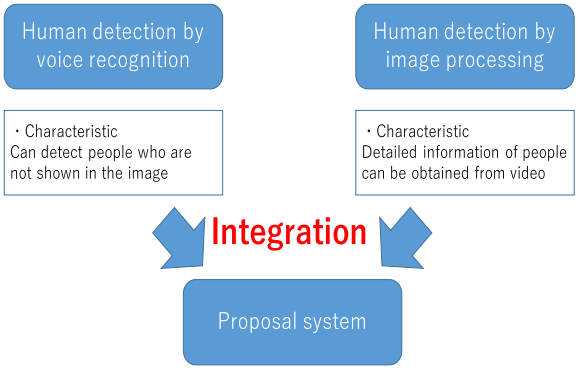

FIGURE 3. Relationship between previous human detection methods and the proposed method.

II. RELATED WORK

In this section, we describe previous research into the two main human search technologies used in the present research, namely, voice processing and image processing.

For human search by voice processing, we previously studied sound recording using a unidirectional microphone mounted on a UAV and detecting people from voice information [11]. We tried using digital-filter voice processing to remove only noise from the mixed sound of voices and propellers, but it was difficult to remove only the propeller sound without losing the voice sound. Another disadvantage was that the detection accuracy dropped sharply with distance from the microphone.

In recent years, products such as smart speakers, cars, and robots that all use voice recognition technology have been developed as systems for detecting human voices using microphones. These products also have the problem of ambient sound being picked up by the microphone and interfering with speech recognition. Therefore, such products use sound-source separation technology in the form of an array microphone equipped with multiple microphones. Sound-source separation technology estimates the direction of the sound source based on the sound pressure and time difference of the sound picked up by each microphone, and the mixed sound containing the surrounding noise pick up by the array microphones is separated. Nakadai *et al.* [12] researched and developed the HARK system that realizes sound-source localization and separation and voice recognition [13].

Techniques for using image processing to search for people or objects in captured images have been actively studied. In particular, object detection methods based on deep-learning networks such as Faster R-CNN [14] and SSD [15] have been rapidly developed in recent years. One such object detection method is the algorithm YOLO v3 developed by Redmon and Farhadi, [16], [17]. A problem with object detection based on a deep-learning network has been that the recognition accuracy for small objects in an image is poor. YOLO v3 uses a concept similar to pyramid network features [18] to extract features from three different scales and predict objects, thereby improving the recognition accuracy of small objects. This feature of YOLO v3 works very well with UAVs, thereby increasing the target distance, and is used in research to detect objects from UAV-mounted cameras [19].

Based on these previous studies, we have built a UAV equipped with an array microphone and a human detection system based on sound-source separation, and we have integrated camera-based human detection as a search aid to build a system with a high detection rate. However, in this application, the distance from the UAV to the search target must be considered. The shorter the distance, the easier it is to pick up sound but the narrower the range that the camera can capture. Therefore, we use the direction information obtained at the time of sound-source localization so that when a person is detected by their voice, we can aim the camera toward them to achieve compatibility between the two systems.

As a summary, that human search can be done with UAV-mounted image and voice processing individual systems. Both approaches are weak in searching humans when UAV flies at higher altitudes. The target humans appear in images from UAV-mounted camera becoming smaller at

higher altitudes while the voice of the target human cannot be captured by UAV-mounted microphones at higher altitudes. These reasons sharply drop the detection accuracy of both approaches. However, the both approaches show comparatively better detection rate at the lower UAV altitudes like 3m. However, the image processing based systems can only search humans when target humans appear in UAV-mounted camera images. As a solution for this issue, we propose an UAV-mounted voice processing based human search systems. Finally, by combining the two type of human search systems, we expect to improve the human detection performance.

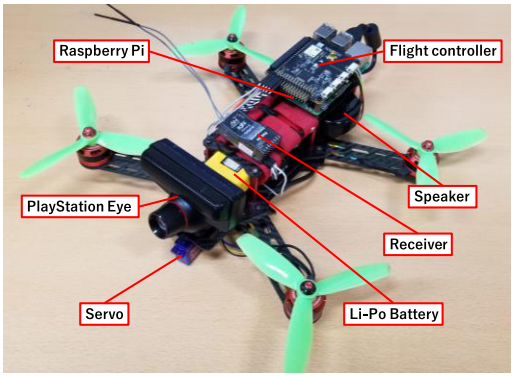

FIGURE 4. Overall view of produced unmanned aerial vehicle (UAV).

III. SYSTEM OVERWIEW

We have been developing different type of UAV systems for making applications in indoor mapping [20], and autonomous flight [21]–[28]. Fig. 4 shows the UAV produced in this study. It is equipped with a loudspeaker to request responses from people, an array microphone to acquire sound, a servomotor to move the camera, and a small lightweight computer to control the aforementioned items. The UAV is also equipped with a Raspberry Pi single-board computer, but the voice and image processing, which is the core of this research, was performed on a separate host computer because the Raspberry Pi had insufficient processing capability. The PlayStation Eye shown in Fig. 5 served as both the array microphone and camera mounted on the UAV. A PlayStation Eye is an integrated camera and array microphone, the latter having four microphones arranged horizontally to collect sound in four channels. Meanwhile, the camera has a maximum resolution of 640 × 480 pixels and an angle of view of 56°. Because the camera must point toward the sound source, the former is attached to a servomotor as shown in Fig. 5.

Fig. 6 shows an overview of the constructed human detection system. First, sound data recorded by the array microphone are sent to the host computer, and sound-source separation processing is performed through sound-source localization. Next, voice recognition is applied to the separated human-voice data. In this way, any words spoken by a person are detected, thereby detecting the presence or absence of a person. The results of the aforementioned sound-source localization, voice recognition, and presence–absence

FIGURE 5. PlayStation Eye and servomotor.

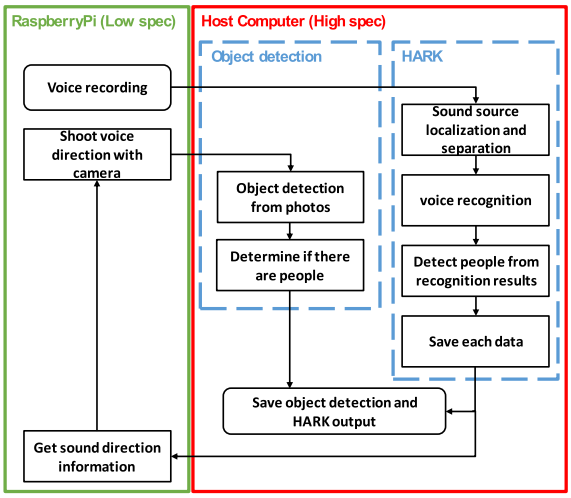

FIGURE 6. Overview of the human detection system.

detection are saved, and the sound-source direction information obtained by the sound-source localization is sent to the Raspberry Pi on the UAV. Using the sound-source direction information, the Raspberry Pi rotates the camera via the servomotor to point in the sound-source direction. The resulting photographic data are also sent to the host computer, which looks for a person in the image. The above sequence is the process of human detection. As for the system execution time, the host computer receives an audio file recorded for 10 s and takes 20 s to process it; as such, the detection system completes one cycle every 30 s.

IV. AUDIO PROCESSING BASED HUMAN DETECTION

A. SOUND-SOURCE SEPERATION AND LOCALIZATION BY HARK

Fig. 7 shows a network diagram of the audio processing flow in this experiment. Sound-source separation is a technique for performing separation based on the input direction of a target sound source. Therefore, it is necessary to localize the direction of the target sound source as preprocessing. In Fig. 7, the sound-source localization node is entitled LocalizeMUSIC and uses what is known as the MUSIC method, in which a transfer function from a sound source to each microphone is measured in advance and is used as prior information. If $h_M(\theta, \omega)$ is the transfer function in the frequency

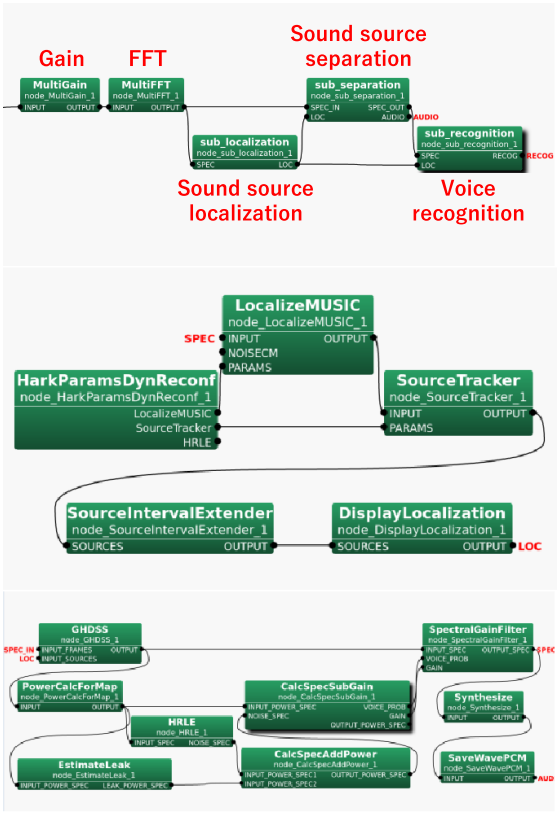

FIGURE 7. Network diagram of audio processing (top: entire system; center: sound-source localization node; bottom: sound-source separation node).

domain from the array microphone to microphone M in the θ direction, then the transfer function vector $\mathbf{H}(\theta, \omega)$ can be expressed as

$$\mathbf{H}(\theta, \omega) = [\mathrm{h}_1(\theta, \omega), \cdots, \mathrm{h}_\mathrm{M}(\theta, \omega)]. \tag{1}$$

Because this transfer function changes greatly depending on the environment, it is necessary to measure it for each experimental environment. In this study, we experimentally determined the transfer function. The transfer function for sound source separation method introduced by HARK [12] is applied. Here, HARK separates inputted sound mixtures into a set of separated sounds and this process needs a set of transfer functions to estimate a separation matrix. These transfer functions are determined by real world measurements between a microphone array and sound sources. Generally, time stretched pulse (TSP) responses or impulse responses recording, according to sound sources are used to determine the transfer functions when the real measurements are used. In this work, we use TSP responses. They can be received in two different categories, as synchronized recording and unsynchronized recording. The unsynchronized recording can be conducted with most microphone array devices. Therefore, in this study unsynchronized recording is conducted. TSP responses are recorded by moving a sound source in a circle at the 5° intervals while keeping the sound source in a fix point. The radius of the circle is set as 3m. After collecting the TSP responses, the transfer functions are determined from the TSP response wav files with HARK-TOOL5 [12].

Next, an inter-channel correlation matrix of the input signal is calculated. First, an M-channel input signal is subjected to a fast Fourier transform (FFT) at the MultiFFT node in Fig. 7 to obtain a frequency-domain signal vector $\mathbf{X}(\omega, f)$ as

$$\mathbf{X}(\omega, f) = [\mathrm{X}_1(\omega, f), \mathrm{X}_2(\omega, f), \cdots, \mathrm{X}_\mathrm{M}(\omega, f)]^\mathrm{T}, \tag{2}$$

where ω is a frequency and f is a frame. The inter-channel correlation matrix $\mathbf{R}$ of the input signal $\mathbf{X}$ is

$$\mathbf{R} = \mathbf{X}\mathbf{X}^*, \tag{3}$$

where $\mathbf{X}^*$ is the complex-conjugate transpose of $\mathbf{X}$. However, in this system, to obtain a stable correlation matrix, an average of the correlation matrix in the time direction is used. Next, by performing eigenvalue decomposition on $\mathbf{R}$, the M-dimensional space is decomposed into a signal subspace and other subspaces. In this paper, SEVD (Standard EigenValue Decomposition) is specified as the algorithm of MUSIC, so eigenvalue expansion is performed as

$$\mathbf{R} = \mathbf{E}\Lambda\mathbf{E}^{-1}. \tag{4}$$

Here, the matrix $\mathbf{E} = [\mathrm{e}_1, \mathrm{e}_2, \cdots, \mathrm{e}_\mathrm{M}]$ consists of mutually orthogonal eigenvectors, and $\boldsymbol{\Lambda}(\boldsymbol{\omega})$ is a diagonal matrix whose eigenvalues correspond to each eigenvector as diagonal components. The eigenvalues corresponding to the eigenvector space $\mathbf{E}$ obtained by the eigenvalue decomposition are correlated with the power of the sound source. Therefore, by taking the eigenvector corresponding to the largest eigenvalue, only the subspace of the target sound with high power is selected. That is, if the number of sound sources to be considered is N, then $[\mathrm{e}_1, \cdots, \mathrm{e}_\mathrm{N}]$ is an eigenvector corresponding to the sound source and $[\mathrm{e}_{\mathrm{N}+1}, \cdots, \mathrm{e}_\mathrm{M}]$ is an eigenvector corresponding to noise.

Based on the above, the MUSIC spectrum for sound-source localization is calculated as

$$\mathrm{P}(\theta, \omega, f) = \frac{\left|\mathbf{H}^*(\theta, \omega)\,\mathbf{H}(\theta, \omega)\right|}{\sum_{i=N+1}^{M} \left|\mathbf{H}^*(\theta, \omega)\,\mathbf{e}_\mathrm{i}(\theta, \omega)\right|}, \tag{5}$$

where the denominator on the right-hand side is the inner product of the transfer function and the eigenvector of the noise component in the input signal. If the transfer function is a vector corresponding to the direction of the target sound source, then the denominator is zero because it is orthogonal to all the eigenvectors corresponding to noise in the input signal. Therefore, in theory, $\mathrm{P}(\theta, \omega, f)$ becomes infinite in the direction of the sound source. In practice, however, it remains finite because of the influence of noise and the like, but the sound-source direction can be obtained nevertheless because a peak is observed.

The sound-source separation process can be formulated as follows. If the transfer vector between the sound source and the microphone is $\mathbf{H}(\omega)$ and the spectrum vector for multiple

sound sources is $\mathbf{s}(\omega)$, then the microphone input $\mathbf{x}(\omega)$ is represented as

$$\mathbf{x}(\omega) = \mathbf{H}(\omega)\,\mathbf{s}(\omega). \tag{6}$$

Using the separation matrix $\mathbf{W}(\omega)$, the sound-source separation result $\mathbf{y}(\omega)$ is expressed as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\,\mathbf{x}(\omega). \tag{7}$$

Here, $\mathbf{W}(\omega)$ where $\mathbf{y}(\omega) = \mathbf{s}(\omega)$ is an ideal separation matrix. Because the environment in this study is assumed to include noise with high directivity, such as the UAV propeller sound, we used the GHDSS (Geometric High-order Dicorrelation-based Source Separation) algorithm to obtain $\mathbf{W}(\omega)$ [31]. This algorithm performs decorrelation among sound-source signals, forms directivity in the direction of the sound source, and is effective at suppressing a noise source that has high directivity. The GHDSS algorithm receives the multi-channel complex spectrum output from the MultiFFT node and the sound-source direction output from the LocalizeMUSIC node and then separates the mixed sounds for each input direction to the microphone.

B. HUMAN VOICE RECOGNITION

Next, the speech obtained after sound-source separation is sent to the speech-recognition node. In this research, we constructed a system that uses speech-recognition technology to distinguish between the human voice and other sounds such as noise [29]. Specifically, speech recognition is performed on the separated speech, and the result is output. In addition, a system is provided in which a word list is created in advance, and a match is made with a speech-recognition result to determine that there is a person when a match is found. This is because the propeller sound is input even when a person is not responding, so that the propeller sound is processed as a human voice and an incorrect speech recognition result is output. Because the erroneous recognition result has a grammatical error, this method can exclude this type of erroneous recognition from the recognition result. In addition, speech recognition generally has the problem that the first part of an utterance section cannot be detected well, and thus a recognition trigger called a magic word is required.

However, the purpose of this research is to search for people, and it is not possible to have the search target issue a magic word. Therefore, we set up a system that can recognize speech without a magic word by notifying the search target that the utterance section has started. Specifically, the loudspeaker mounted on the UAV informs the person below that the recording interval has started. After that, recording is performed, and voice recognition is performed on the recorded voice data. Although this is a simple method, it reduces the rate of false recognition without impairing the usability of the dialogue between the person and the UAV.

C. MERGING WITH CAMERA BASED HUMAN DETECTION

This paper aims at detecting a person on the ground from a UAV in flight, and it is assumed that the UAV hovers at an altitude of least 3 m above the ground. Because the scale of the detected target person in the captured image decreases with distance, it is necessary to consider the detection of a small object. Therefore, we attempted object detection using YOLO v3 as a method of human detection. As described above, YOLO v3 is an object detection method that can handle the recognition of objects of different scales, which is difficult in object-recognition processing.

YOLO v3 involves object detection based on a deep-learning network, and its detection accuracy depends on the quality of the input image and the training dataset. Therefore, we chose to use the COCO dataset [30], which is a large-scale object detection, segmentation, and captioned dataset that features rich annotations. As described above, in this study we detected a small photographed person by performing object detection using YOLO v3 trained using the COCO dataset.

V. EXPERIMENTAL EVALUATION

A. EXPERIMENTAL ENVIRONMENT

We performed three experiments on the proposed system and evaluated its performance. In the first experiment, we mounted it on a UAV and measured the rate of human detection when talking to the array microphone. The measurement was performed with and without the UAV propellers turning, and the detection rates of each case were compared. In addition, the same experiment was performed in an environment in which a single microphone was mounted as the voice input device [11], and the results when voice recognition was performed without performing sound-source separation processing were also measured. For the measurement, we attempted speech recognition 20 times at each distance and checked whether it was possible to detect people. Note that the sound pressure of the voice is not always constant and changes because of sound diffusion and reverberation. Therefore, when the distance is large, the sound pressure input to the array microphone becomes small, and localization may not be performed. When this phenomenon occurred, we responded by lowering the threshold of sound pressure necessary for localization. In the second experiment, we used the sound-source localization results to calculate the accuracy of the sound-source direction. Finally, in the third experiment, we analyze the performance of human detection combining the proposed voice based human detection method with a human detection method by UAV mounted camera.

B. HUMAN DETECTION EVALUATIN

Fig. 8 shows the measured data for the rate of human detection. For data comparison with a conventional method [11], we also measured the results of human search using sound data recorded with a one-channel microphone with neither sound-source localization nor sound-source separation.

In the Fig. 8, the orange line and blue line indicate the human detection performance of proposed method and conventional method respectively, when propeller sound does not exist in the environment. On the other hand, the yellow

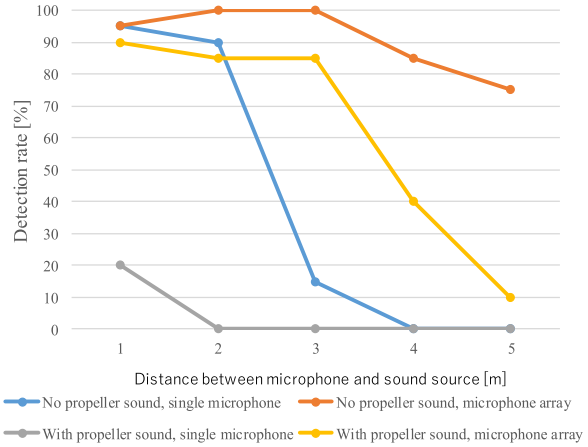

FIGURE 8. Human detection comparison.

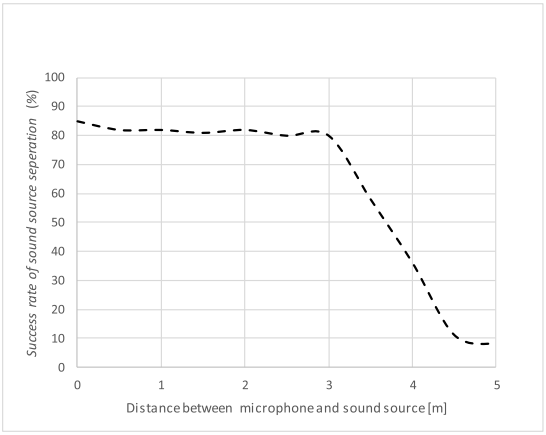

FIGURE 9. Sound source separation results of the proposal.

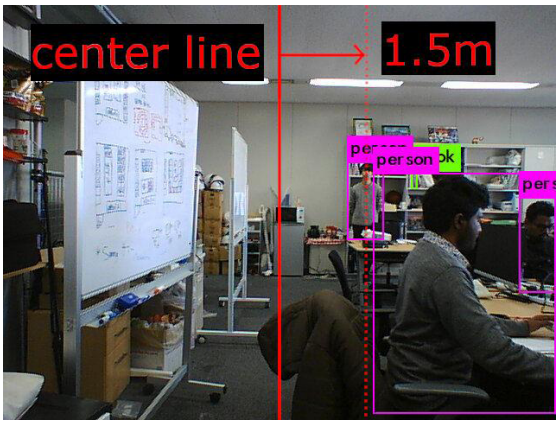

FIGURE 10. Object recognition performed on an image.

line and grey line indicate the human detection results of the proposed method and conventional method respectively, when both human sound and propeller sound exist in the environment. The latter experiments with availability of UAV sound, were conducted with the real UAV at both indoor and outdoor environments. For these experiments, we used our developed UAV shown in Fig. 4. Here, 20 experiments were conducted in each indoor and outdoor environments by changing the distance between microphone and sound source (human voice) approximately between 1m to 5m as shown Fig. 8. During the outdoor experiments, the outdoor environment was noisy for a certain level with natural sounds such as wind. On the other hand, during the indoor experiments, the environment did not include such natural sounds, only included UAV sound and human voice.

We did further analysis to confirm the effectiveness of the sound source separation part and voice recognition part in our proposal. For this analysis, we used experimental data acquired from both indoor and outdoor. According to analysis, the sound source separation performance varies sharply according to the distance between microphone and human sound, as shown in Fig. 9. On the other hand, the average voice recognition rate after the sound source separation was 92.8% and this value was almost constant and did not vary much according to the distance changes between microphone and human sound.

The results of conventional method shows that when the sound is collected using only one channel, the distance of around 3 m is the limit of speech recognition even without propeller sound. This is because the sound diffuses with distance and becomes difficult to acquire. By contrast, although the accuracy decreased with distance when the array microphone was used, human detection was confirmed possible even in a noisy environment.

C. SOUND-SOURCE LOCALIZATION EVALUATION

Fig. 10 shows an image taken after rotating the camera based on the direction information obtained by sound-source localization. When source localization was performed for a voice uttered by a person 5 m away from the UAV's position, the direction of the input sound source was determined to be 20.023989° with respect to the array microphone. However, when the camera was moved to that angle, the person was not in the center of the image, as shown in Fig. 9. There was a deviation of 1.5 m from the center of the image, corresponding to an angular deviation of approximately 17°. This was attributed to the resolution of the array microphone. In this study, we used an array microphone comprising four microphones arranged horizontally. Because the accuracy of sound-source localization is correlated with the number of microphones mounted, we expect improved accuracy by mounting more microphones. However, because the purpose of this study was to show the person in the image, sufficient localization accuracy was confirmed. In addition, on a host computer with a Core i5-7500 processor and 8 GB of memory, the average detection time for one image file using YOLO v3 was around 6.

D. EVALUATION OF COMBINED VOICE BASED AND IMAGE BASED HUMAN DETECTION

In the previous section, we discussed the sound source localization by using the human detection with an UAV-mounted camera. In this section we discuss human detection combining both voice based method (proposed method) and an image

based method with UAV-mounted camera. In these experiments, UAV altitudes were approximately 2m-5m. We used YOLO v3 to detect humans from the UAV-mounted camera [17]. According to the results, the YOLO v3 itself human detection rate by using UAV-mounted camera images was 84.5%. The reason for most missed-detections was that the humans did not appear in camera images even though humans are in the environment. However, the human detection rate by combining these two systems was 93.3%. Thus, according to this combination, we could confirm a clear human detection improvement. Furthermore, human detection using images from UAV-mounted camera is difficult at higher altitudes. However, this problem would be solved for a certain level by using a wide-angle high resolution camera.

VI. CONCLUSION

In this paper, we proposed a human search system using a UAV with an array microphone. Specifically, we have constructed a system that detects the human voice by applying sound-source separation processing to the mixed sound of voice and propellers collected by the array microphone on the UAV. In addition to voice-based human detection, we used YOLO v3—a deep-learning-based object detection method—for human detection from images. In doing so, we obtained more information by integrating the two detection methods, thereby realizing a possible human detection system.

In experiments to assess the performance of the proposed system, we focused on the accuracy of human detection and sound-source localization and the processing speed. We measured the accuracy of human detection with and without sound-source separation processing and obtained higher accuracy with separation. The localized position of the sound source deviated from the actual position in some cases, but the accuracy required for human detection using images was maintained.

REFERENCES

[1] G.-J.-M. Kruijff, F. Pirri, M. Gianni, P. Papadakis, M. Pizzoli, A. Sinha, V. Tretyakov, T. Linder, E. Pianese, S. Corrao, F. Priori, S. Febrini, and S. Angeletti, "Rescue robots at earthquake-hit Mirandola, Italy: A field report," in *Proc. IEEE Int. Symp. Saf., Secur., Rescue Robot. (SSRR)*, Nov. 2012, pp. 1–8.

[2] D. Chen, Z. Liu, L. Wang, M. Dou, J. Chen, and H. Li, "Natural disaster monitoring with wireless sensor networks: A case study of data-intensive applications upon low-cost scalable systems," *Mobile Netw. Appl.*, vol. 18, no. 5, pp. 651–663, 2013.

[3] M. Erdelj and E. Natalizio, "UAV-assisted disaster management: Applications and open issues," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2016, pp. 1–5.

[4] Y. Ham, K. K. Han, J. J. Lin, and M. Golparvar-Fard, "Visual monitoring of civil infrastructure systems via camera-equipped unmanned aerial vehicles (UAVs): A review of related works," *Vis. Eng.*, vol. 4, no. 1, p. 1, Dec. 2016.

[5] S. Lee, D. Har, and D. Kum, "Drone-assisted disaster management: Finding victims via infrared camera and Lidar sensor fusion," in *Proc. 3rd Asia–Pacific World Congr. Comput. Sci. Eng. (APWC CSE)*, Dec. 2016, pp. 84–89.

[6] A. Gaszczak, T. P. Breckon, and J. Han, "Real-time people and vehicle detection from UAV imagery," *Proc. SPIE*, vol. 7878, Jan. 2011, Art. no. 78780B.

[7] S. Yang, X. Yang, and J. Mo, "The application of unmanned aircraft systems to plant protection in China," *Precis. Agricult.*, vol. 19, no. 2, pp. 278–292, Apr. 2018.

[8] X. Li, Y. Zhao, J. Zhang, and Y. Dong, "A hybrid PSO algorithm based flight path optimization for multiple agricultural UAVs," in *Proc. IEEE 28th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2016, pp. 691–697.

[9] D. Mellinger and V. Kumar, "Minimum snap trajectory generation and control for quadrotors," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2520–2525.

[10] J. Yoon, I. Kim, W. Chung, and D. Kim, "Fast and accurate car detection in drone-view," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2016, pp. 1–3.

[11] Y. Yamazaki, M. Tamaki, C. Premachandra, C. J. Perera, S. Sumathipala, and B. H. Sudantha, "Victim detection using UAV with on-board voice recognition system," in *Proc. 3rd IEEE Int. Conf. Robotic Comput. (IRC)*, Feb. 2019, pp. 555–559.

[12] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition HARK and its evaluation," in *Proc. 8th IEEE-RAS Int. Conf. Humanoid Robots*, Dec. 2008, pp. 561–566.

[13] K. Nakadai, M. Kumon, H. G. Okuno, K. Hoshiba, M. Wakabayashi, K. Washizaki, T. Ishiki, D. Gabriel, Y. Bando, T. Morito, R. Kojima, and O. Sugiyama, "Development of microphone-array-embedded UAV for search and rescue task," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5985–5990.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 2016, *arXiv:1612.08242*. [Online]. Available: http://arxiv.org/abs/1612.08242

[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[18] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[19] J. Wang, S. Jiang, W. Song, and Y. Yang, "A comparative study of small object detection algorithms," in *Proc. Chin. Control Conf. (CCC)*, Jul. 2019, pp. 8507–8512.

[20] K. Nakajima, C. Premachandra, and K. Kato, "3D environment mapping and self-position estimation by a small flying robot mounted with a movable ultrasonic range sensor," *J. Electr. Syst. Inf. Technol.*, vol. 4, no. 2, pp. 289–298, Sep. 2017.

[21] C. Premachandra, M. Otsuka, R. Gohara, T. Ninomiya, and K. Kato, "A study on development of a hybrid aerial terrestrial robot system for avoiding ground obstacles by flight," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 1, pp. 327–336, Jan. 2019.

[22] C. Premachandra, D. Ueda, and K. Kato, "Speed-up automatic quadcopter position detection by sensing propeller rotation," *IEEE Sensors J.*, vol. 19, no. 7, pp. 2758–2766, Apr. 2019.

[23] C. Premachandra and M. Otuka, "A basic automation approach for hybrid aerial/terrestrial robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2017.

[24] C. Premachandra, S. Takagi, and K. Kato, "Flying control of small-type helicopter by detecting its in-air natural features," *J. Electr. Syst. Inf. Technol.*, vol. 2, no. 1, pp. 58–74, May 2015.

[25] F. Sato, Y. Motomura, C. Premachandra, and K. Kato, "Absolute positioning control of indoor flying robot using ultrasonic waves and verification system," in *Proc. 16th Int. Conf. Control, Automat. Syst.*, Oct. 2016, pp. 1600–1605.

[26] C. Premachandra, T. Yoshida, and K. Kato, "A basic study of landing system for multicopters using raspberry pi," in *Proc. Int. Symp. Consum. Electron. (ISCE)*, Jun. 2015, pp. 1–2.

[27] S. Sato, C. Premacandra, and K. Kato, "Position and attitude estimation using ultrasonic waves for autonomous flying robots and system construction," in *Proc. 15th Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2015, pp. 243–248.

[28] C. Premachandra, T. Ninomiya, R. Gohara, and K. Kato, "Improvement of multicopter detection using an infrastructure camera," in *Proc. IEEE 6th Int. Conf. Consum. Electron. Berlin (ICCE-Berlin)*, Sep. 2016, pp. 1–2.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. Workshop Autom. Speech Recognit. Understand.*, 2011.
[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[31] K. Nakadai, G. Ince, K. Nakamura, and H. Nakajima, "Robot audition for dynamic environments," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2012, pp. 125–130.

YUKI YAMAZAKI received the B.S. degree in electronic engineering from the Shibaura Institute of Technology, Tokyo, Japan, in 2017. He is currently pursuing the M.S. degree with the Graduate School of Engineering and Science, Shibaura Institute of Technology, Tokyo, Japan.

His research interests include audio processing, pattern recognition, and human support systems.

CHINTHAKA PREMACHANDRA (Member, IEEE) was born in Sri Lanka. He received his B.Sc. and M.Sc. degrees from Mie University, Tsu, Japan, in 2006 and 2008, respectively, and the Ph.D. degree from Nagoya University, Nagoya, Japan, in 2011.

From 2012 to 2015, he was an Assistant Professor with the Department of Electrical Engineering, Faculty of Engineering, Tokyo University of Science, Tokyo, Japan. From 2016 to 2017, he was an Assistant with the Department of Electronic Engineering, School of Engineering, Shibaura Institute of Technology, Tokyo. In 2018, he was promoted to Associate Professor with the Department of Electronic Engineering, Graduate School of Engineering, Shibaura Institute of Technology. In addition, he is the Manager of the Image Processing and Robotic Lab, Department of Electronic Engineering. His lab conducts research in two main fields: image processing and robotics. Former research includes computer vision, pattern recognition, speed up image processing, and camera-based intelligent transportation systems, while latter fields include terrestrial robotic systems, flying robotic systems, and integration of terrestrial robot and flying robot.

Dr. Premachandra is the member of IEICE, Japan, SICE, Japan, and SOFT, Japan. He received the FIT Best Paper Award and FIT Young Researchers Award from IEICE & IPSJ, Japan, in 2009 and 2010, respectively. He has served many international conferences and journals as a steering committee member and an editor respectively. He is the Founding Chair of International Conference on Image Processing and Robotics (ICIPRoB).

CHAMIKA JANITH PEREA (Member, IEEE) received the B.Sc. degree in mechanical engineering and the M.Sc. degree in robotics from the University of Moratuwa, Sri Lanka, in 2016 and 2018, respectively. His past research experiences include short term research internship with the Shibaura Institute of Technology, Tokyo, Japan, and research assistantship with the Lady Ridgeway Hospital, Sri Lanka. He is currently working with the Center for Advance Robotics, University of Moratuwa. His research interests include unmanned aerial vehicles, assistive robotics, and bio signal processing.