Speaker Tracking Method Based on Audio-Visual Bimodal Fusion and Its Application to Mobile Robot

Hao-Yan Zhang[*1*2*3], Qi-Feng Shi[*1*2*3], Long-Bo Zhang[*1*2*3], and Zhen-Tao Liu[*1*2*3†]

[*1]School of Automation, China University of Geosciences, Wuhan, Hubei 430074, China
E-mail: liuzhentao@cug.edu.cn
[*2]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems
Wuhan, Hubei 430074, China
[*3]Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education
Wuhan, Hubei 430074, China
†Corresponding author

Initiative service is one of the key research directions of the new generation of service robots. It is important to track human automatically for initiative service in human-robot interaction. To solve the problems of low precision and poor anti-interference capability of only using single-modal (audio or visual information), a speaker localization and tracking method based on audio-visual bimodal fusion is proposed. Firstly, the azimuth of speaker is obtained based on time difference of arrival (TDOA) using microphone array and face detection based on AdaBoost algorithm is carried out using camera. Frame rate tracker and spatial coordinate fusion are used to obtain the location of speaker. Secondly, path planning according to the preset tracking mode is designed, by which the azimuth and distance between the robot and the speaker is maintained in a limited range. To simulate different situations, different angle and distance for speaker tracking were set. Finally, a mobile robot is driven to follow the planned path by the real-time control system using STM32, and information from the microphone array and the camera are collected by Raspberry Pi. The proposed method was tested by setting 16 different sets of azimuth and distance data (the speaker moves slowly in the experimental place). Experimental results demonstrate that the running trajectory of the mobile robot is close to the ideal preset trajectory, which ensures effective tracking of the target speaker.

Keywords: Speaker Tracking, Audio-Visual Bimodal, Information Fusion, Mobile Robot.

1. INTRODUCTION

Intelligent robot is a system that simulates human beings [1] in various ways such as environment perception, autonomous planning and information interaction. Service robot is the representative of future science and technology, the core technology of which is to locate [2] and autonomously track the target person. It can obtain the information of human's location, state and demand, and communicate with humans initiatively. The development of service robot has an increasing demand for the rapidity, accuracy, and universality of human-robot interaction. However, it is difficult for robots to initiatively interact with humans through current human-robot interaction technology, which greatly limits the flexibility and usability [3]. Therefore, to further improve the efficiency of human-robot interaction, an automatic speaker locating and tracking method is introduced.

With the maturity of computer vision, it has been used in the research of target localization by many researchers [4]. In the field of sound source localization, the time difference of arrival (TDOA) [5] method is widely used because of its advantages of low computational complexity and good real-time performance [6]. However, most of the current methods only use single-modal (audio or visual information), which cannot fully utilize environmental information, resulting in low sensitivity, poor stability, and blind spots [7]. It is more difficult to achieve the desired effect by single-modal based method when tracking humans. Therefore, we propose a speaker recognition and tracking method based on bimodal fusion, which utilizes information redundancy to eliminate the blind spots and improve accuracy and stability. A microphone array is used to obtain the speaker's azimuth information in the range of 0~360° based on the TDOA method, and a camera is used to obtain both the azimuth and distance information of the face based on face detection. When the speaker is out of the field of vision, the advantage of wide coverage of microphone array [8] is utilized to steer the robot to find the speaker. When the speaker is in the vision of the camera, the robot is steered to track the speaker according to the preset tracking mode. In addition, face feature extraction are used to realize the multi-target screening strategy. In the field test, the speaker tracking robot based on audio-visual bimodal fusion can effectively locate and track the speaker.

The remainder of this paper is organized as follows. In Section 2, the research status of speaker localization and target tracking are introduced. In Section 3, a speaker tracking method based on bimodal fusion is proposed and applied to a mobile robot. Section 4 presents the experimental setting and result analysis. Section 5 summarizes the research and prospects the future research direction.

2. METHODOLOGY

The design of speaker tracking method based on bimodal fusion is shown in **Fig. 1**. The method mainly consists of microphone array audio recognition and localization, monocular face detection and localization, and autonomous path planning.

2.1. Sound Source Localization

The speech signal is acquired by the six-element annular microphone array and processed by TDOA method. The schematic diagram of Iflytek's annular six-element microphones is shown in **Fig. 2**. The microphones are arranged in an annulus in a centrosymmetric way, which can identify sound signals within a range of 360° and effectively eliminate echo and reverberation.

In sound source localization based on microphone array, an azimuth calculation model based on TDOA is shown in **Fig. 3**. After getting the difference of the time through GCC time delay estimation, the azimuth of the sound source relative to the microphone array is calculated.

Generally, the distance between each microphone unit is far less than the distance between the sound source and the microphone array, so the distance difference between the sound signal and the two microphone units is approximated as $dsin\theta$, where d is the distance between two microphone units and θ is the azimuth of the sound source relative to the microphone array.

The azimuth of the sound source relative to the microphone array is calculated as

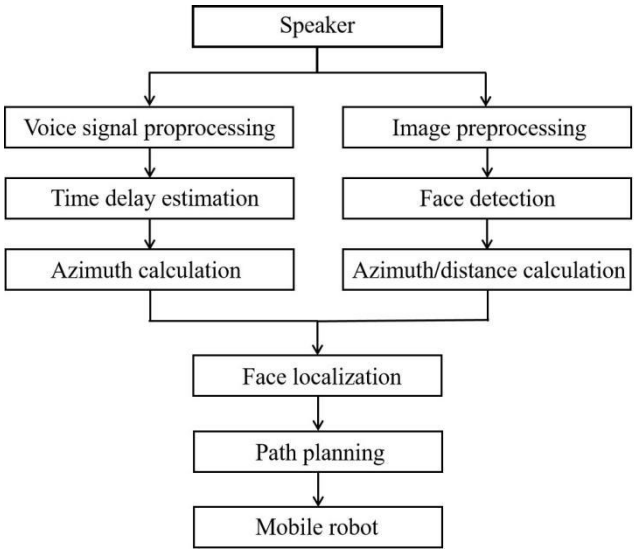


Fig. 1 Procedure of the speaker tracking method

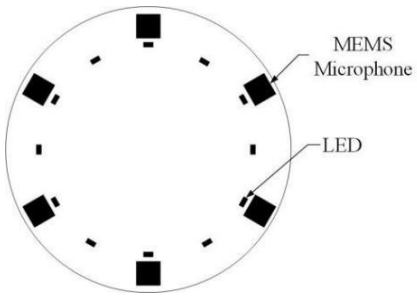


Fig. 2 Annular six-element microphone array

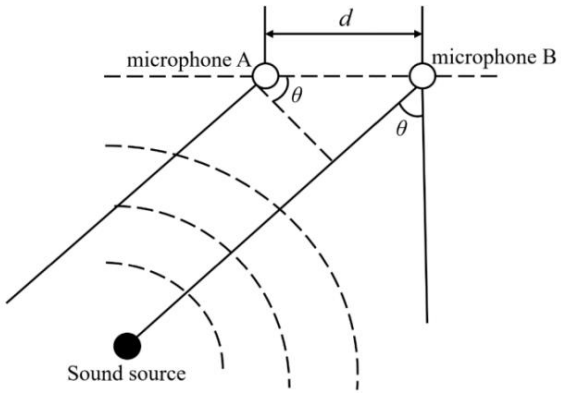


Fig. 3 Azimuth calculation based on delay estimation

$$\theta = \arcsin\frac{c \times \Delta t}{d} \tag{1}$$

where Δt is the time difference of arrival and c is sound velocity.

Since the complete array of microphones is arranged in an annulus, the azimuth obtained by any set of microphones is converted by 0~2π addition and subtraction to obtain the available azimuth. Finally, take the mean value to obtain accurate measurement results as

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} f_i(\theta_i)}{n} \tag{2}$$

2.2. Monocular Visual Face Localization

Face detection and localization with camera is mainly divided into two steps. Firstly, face detection and feature extraction are carried out by using Haar feature classifier and Adaboost algorithm in OpenCV. Secondly, face localization is carried out by calculating face orientation and size in the image.

2.2.1. Face Detection Based on Adaboost

Because the original image collected by the camera has strong noise interference, image graying and histogram equalization are conducted in OpenCV for preprocessing.

Then, Haar feature extraction is performed on the preprocessed images and Adaboost algorithm is used to train the cascade classifier. Firstly, the difference between pixels of different depth in the image is calculated to construct the Haar feature, and then the weight distribution of the training data is initialized. Suppose that there are N training sample data, then each sample is evenly divided with the same weight at the beginning.

$$D = (w_1, w_2, ..., w_N), \quad w_i = \frac{1}{N}, i = 1,2,...,N \tag{3}$$

where D is training sample set, and w_i is sample point. Then, the weak classifier is train. Different sample points are assigned different weights according to the classification results. Weights are updated to form new sample sets, which are used to train the next classifier, and the whole training process goes on iteratively. Finally, the trained weak classifiers are combined into a strong classifier, where classifiers with lower error rates will be

given higher weights, so that they can play a leading role in the final classification function [9].

Fig. 4 shows the image collected by the camera and the face detection result obtained by using the strong classifier. It can not only accurately locate the position of the face, but also detect the opening and closing states of the mouth, so as to assist in determining whether it is a speaker.

2.2.2. Face Distance Calculation

Based on the result of face detection, the coordinate of a face is converted from image coordinate system to space coordinate system, and the azimuth angle and distance of face relative to the camera are calculated.

Due to the hardware limitation of the camera, the image is distorted, which seriously affects the measurement accuracy. A checkerboard calibration method [10] is used to calibrate the image.

The process of the camera photographing an object is regarded as pinhole imaging. The distance between the face and the optical center of camera are measured according to the physical object and object image size conversion relationship. As is shown in **Fig. 5**, L_1 and L_2 are the length of the object and the length of the image, O is the optical center of camera, α is the plane passing through the optical center and perpendicular to the optical axis. d is the distance between the object and the optical center. f is the focal length.

According to the properties of similar triangles, we have

$$d = f \times \frac{L_1}{L_2} \tag{4}$$

It can be seen that in the case where f and L_1 are fixed values, d and L_2 are inversely proportional. f is obtained according to the camera parameters, L_1 is the actual width of a face, which is regarded as a constant value.

Let x be the number of pixels taken by face width in the screen, which is obtained by extracting the coordinates of human face feature points in the pixel coordinate system and calculating the difference value. Replace L_2 with x, then Eq. (4) is simplified as

$$d = \frac{k}{x} \tag{5}$$

Firstly, different target points are taken for measurement, and the value of d and x are recorded each time. Secondly, the d-x curve is fitted based on inverse relationship to get the value of k. Thirdly, the distance between the face and the camera's optical center is calculated by the number of pixels occupied by the face width in the image.



Fig. 4 An example of face detection

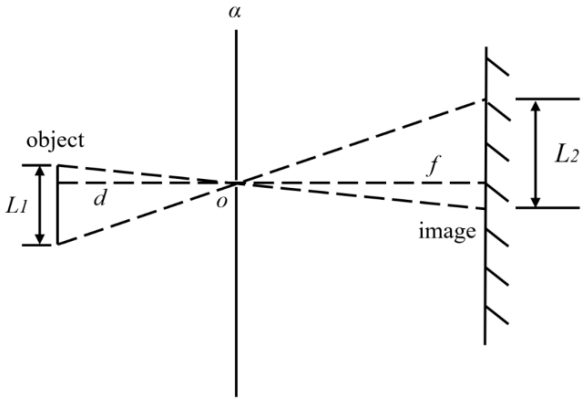


Fig. 5 Distance calculation model

2.2.3. Face Azimuth Calculation

When measuring the angle between the straight line connecting the face and the camera optical center and the camera optical axis, the face is regarded as a particle, and the center of the face detection frame is regarded as the position of its image. As **Fig. 6** shows, β is the field angle of camera, b/a is the ratio of the distance between the center of the frame and the edges of the frame in the pixel coordinate system.

The law of sine are applied in the isosceles triangle:

$$\frac{b}{\sin(\beta - \lambda)} = \frac{a}{\sin \lambda} \tag{6}$$

The value of λ is

$$\lambda = \arctan \frac{\sin \beta}{b/a + \cos \beta} \tag{7}$$

The azimuth is

$$\chi = \frac{\beta}{2} - \lambda \tag{8}$$

Substituting (7) into (8), we have

$$\chi = \frac{\beta}{2} - \arctan \frac{\sin \beta}{b/a + \cos \beta} \tag{9}$$

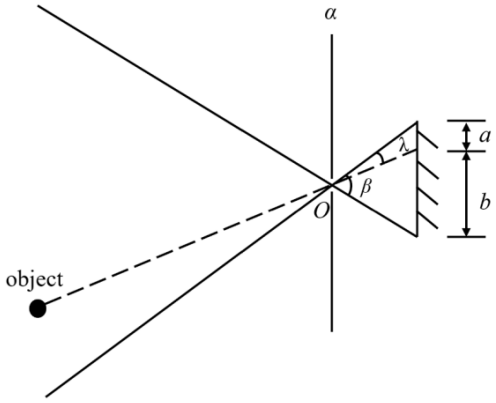


Fig. 6 Azimuth calculation model

2.3. Autonomous Path Planning

The autonomous path planning strategy based on bimodal localization is shown in **Fig. 7**. There are two cases for the face capture, i.e., the face is captured by camera or not.

When the camera does not capture any face, the microphone array is used to detect the ambient sound. To simulate the scene of user speaking and effectively eliminate the interference of environmental noise, specific recognition words are set. Once the array captures a sound signal, Iflytek's voice engine are used to match the feature. If they are matched up, then the azimuth of the sound source is obtained, and the motion control system will steer according to the azimuth to let the camera find the face. When the camera captures a face, the system will follow it according to the distance and azimuth information. When the camera captures multiple faces, the system chooses the speaker based on whether there is a mouth opening and closing motion, otherwise it chooses the nearest face.

When the camera captures the speaker for tracking, to simulate different application scenarios and make the system reach the different orientation of the target speaker, the tracking distance d and tracking azimuth χ are preset. When the system captures the face, b/a , the ratio of the image width to the number of pixels in the image and the distance from the center of the detection frame to both sides of the image is calculated in real time. The ideal value of x and b/a are obtained from Eq. (5) and Eq. (9) as

$$x_{id} = \frac{k}{d} \tag{10}$$

$$\left(\frac{b}{a}\right)_{id} = \frac{\sin\beta}{\tan(\beta/2-\chi)} - \cos\beta \tag{11}$$

Since the actual values of x and b/a always deviate from the ideal value, two open intervals centered on x_{id} and $(b/a)_{id}$ are set. When the measured value falls into the interval, it is regarded as reaching the ideal value.

We choose an Ackermann four-wheel car as the motion system of the mobile robot. The two rear wheels are driving wheels, and the two front wheels are steering wheels controlled by a group of steering gear.

The motion state of the system is regarded as the superposition of the states s_1 (forward and backward controlled by the driving wheel, s_1+ forward and s_1- backward) and s_2 (left and right controlled by the steering wheel, s_2+ left and s_2- right).

If the value of x falls on the left side of the interval (x_1, x_2), then the state of s_1 is positive. If x falls on the right side of the interval (x_1, x_2), then the state of s_1 is negative. Otherwise stop.

If the value of b/a falls on the left side of the interval (y_1, y_2), then the state of s_2 is positive. If b/a falls on the right side of the interval (y_1, y_2), then the state of s_2 is negative. Otherwise stop.

In particular, if x falls within the interval (x_1, x_2) but b/a does not fall within the interval (y_1, y_2), then the state of s_1 is negative, so that the robot can relocate the target within the safe range.

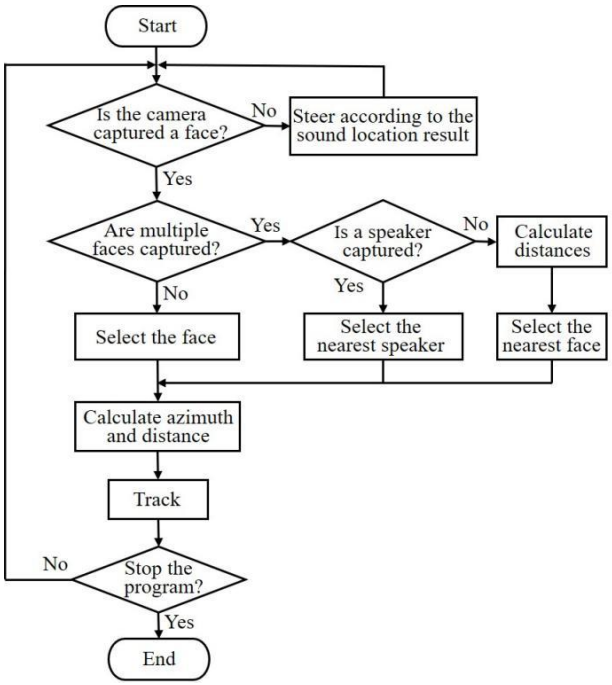


Fig. 7 Flowchart of path planning

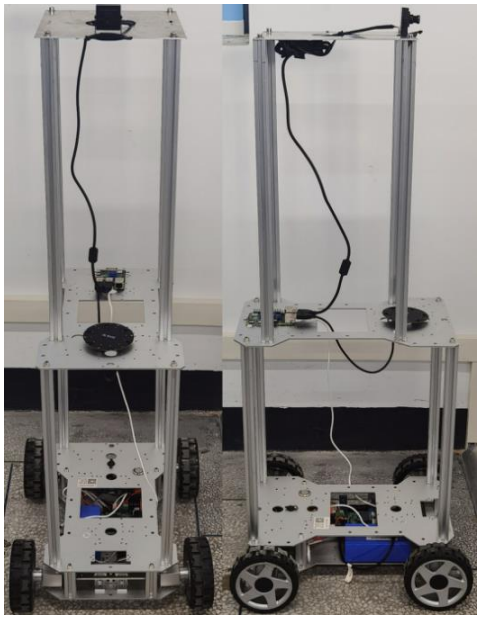


Fig. 8 Physical diagram of the robot

3. EXPERIMENTS

3.1. Experimental Setting

In this system, Iflytek's annular six-element microphone array and a USB camera were used as external sensors, STM32 microcontroller was used to control the robot's

movement, Raspberry Pi received the information transmitted by microphone array and camera for processing, and then sent motion control instructions to the slave computer STM32 to realize the control of the robot.

The physical diagram of the mobile robot is shown in **Fig. 8**. The robot is 0.442 m in length, 0.354 m in width and 1.053 m in height. An Ackermann four-wheel car was used as robot's chassis, and two platforms were supported by metal beams to place sensors and Raspberry Pi. Importantly, the optical center of the camera and the geometric center of the microphone array were on the same vertical line.

C++ and Python were used as programming languages. Linux (Ubuntu) operating system and ROS (Robot operating system) were installed for Raspberry Pi, which were used as software working environment. The official SDK of Iflytek was used to preprocess the voice signal and then sent information to Raspberry Pi, which calculated the azimuth of the sound source. The Raspberry Pi directly read the image information collected by the camera, determined whether there was a face in the vision of the camera and calculated the location of the face.

The experiment was carried out indoors in a space of 8 m $\times$ 4 m $\times$ 3.5 m, with a lot of noise, echo and light interference. To command Raspberry Pi to start or stop the tracking program, use SSH to remotely log in to Raspberry Pi.

Firstly, the tracking accuracy of the system was tested in the case of single speaker. To simulate a variety of tracking modes, 16 different target points were set, and each corresponding coordinate is represented by (d, χ) shown in **Table 1**. The face was treated as a primary point. The coordinates represented the positions that the speaker's face should reach after initiative tracking, which took the camera optical center of the robot as the reference point. d is the distance between the face and the optical center and χ is the azimuth between the camera's optical axis and the straight line connecting the face and the camera's optical center. While tracking, the speaker was still and kept his face at the same height as the camera. The robot started to track from a random position. After the tracking system was stabilized, the tracking program was stopped and the relative position between the robot and the speaker was measured. Then, let the speaker change his position and conducted the same procedure. The same procedure were repeated for 10 times and the result at each time was recorded. The value of d and χ were averaged and the average results were used as the test result of this point. The same procedure was repeated until all 16 points were tested. Throughout the process, the speaker was always in the vision of the camera. Secondly, the tracking effect in the case of multiple faces was tested. There were several people in the field, but only one speaker, and the tracking was tested when the target was not in the vision of the camera at the beginning. Record the response time for the robot to find the speaker. Moreover, for the situation where the camera captured multiple faces at the same time, we randomly let some of them to open and close their mouths to simulate speaking, and observed the system's reaction and recorded the response time.

Table 1. Target point set

A	B	C	D
(0.3, 0.0°)	(0.5, 0.0°)	(1.0, 0.0°)	(1.5, 0.0°)
E	**F**	**G**	**H**
(2.0, 0.0°)	(2.3, 0.0°)	(3.0, 0.0°)	(0.8, 5.0°)
I	**J**	**K**	**L**
(1.6, -5.0°)	(0.7, 8.0°)	(1.7, -10.0°)	(2.4, 15.0°)
M	**N**	**O**	**P**
(1.5, -20.0°)	(1.3, 25.0°)	(2.0, -30.0°)	(2.0, -18.0°)

3.2. Experimental Results and Analysis

The experimental results of tracking accuracy were recorded in **Table 2**. d and χ are the preset tracking distance and azimuth. For each target point, 10 tests were conducted randomly. The actual distance and azimuth between the robot and the speaker were measured and the mean value is obtained as $\hat{d}$ and $\hat{\chi}$. If the target point was beyond the effective tracking range, then it was marked as Error.

According to **Table 2**, the system can reliably track the speaker when the value of d is between 0.4 m and 2.3 m and the value of χ is between -25° and +25°. But for the target points beyond the range, the tracking accuracy is obviously reduced or even invalid.

When the camera does not capture a face, the system can calculate the source direction of the voice signal and turn a specific angle to find the speaker. In this case, the time for the mobile robot to find the speaker is determined by the initial azimuth. The closer the azimuth is to 180°, the longer the time. The response time is 1~12 s. When the camera captures multiple faces, the robot tracks the nearest face by default. However, once the system judges that there exist speakers in the field through mouth's opening and closing detection, the nearest speaker will be tracked. From

Table 2. Experimental results of tracking accuracy

Target points	Ideal value		Experimental results	
	d/m	χ	$\hat{d}$/m	$\hat{\chi}$
A	0.3	0.0°	0.248	2.31°
B	0.5	0.0°	0.500	3.67°
C	1.0	0.0°	0.938	-1.98°
D	1.5	0.0°	1.565	-1.83°
E	2.0	0.0°	2.014	0.17°
F	2.3	0.0°	2.458	0.37°
G	3.0	0.0°	Error	Error
H	0.8	5.0°	0.775	5.33°
I	1.6	-5.0°	1.536	-1.90°
J	0.7	8.0°	0.680	6.41°
K	1.7	-10.0°	1.629	-7.90°
L	2.4	15.0°	Error	Error
M	1.5	-20.0°	1.435	-15.70°
N	1.3	25.0°	1.181	21.90°
O	2.0	-30.0°	Error	Error
P	2.0	-18.0°	2.007	-12.91°

the experimental results, we find that the mobile robot has a response time of within 2 s when someone inside the vision of the camera starts speaking or stops speaking.

The experimental results show that the system has a reliable general tracking effect in various situations. When the camera captures a face, the system has good real-time tracking performance and fast response speed.

The error sources and reliability analysis of the tracking system are shown as follow.

(1) When the distance between the speaker and the robot is less than 0.4 m or greater than 2.3 m, the program cannot frame a complete face or the face is smaller than the threshold of detection. Similarly, when the actual azimuth is larger than half of the camera field angle, the camera cannot capture the complete face. Therefore, when the tracking distance d is less than 0.4 m or greater than 2.3 m, or the tracking azimuth χ falls out of $(-25°, +25°)$, the actual distance and azimuth will converge and cannot continue to approach the ideal value.

(2) Due to the robot's inertia, there is a time difference between the robot's executing the stop instruction and its really stopping, which produces a slight impact on the tracking accuracy.

(3) Although the camera has been calibrated, the edge of the image is still slightly distorted, resulting in localization errors when the face is located at the edge of the image.

4. CONCLUSION

A speaker tracking method based on audio-visual bimodal fusion was proposed to improve efficiency of human-robot interaction. To simulate different situations, our proposed method was tested by setting 16 different sets of tracking azimuth and distance data. The experiment was conducted on an indoor field by building a mobile robot. The experimental results show that our method would obtain a high accuracy of speaker tracking if the preset target distance between speaker and the tracking system ranges from 0.4 m to 2.3 m and the preset azimuth range is from −25° to +25°. The mobile robot has the ability to autonomously track the speaker and is able to response quickly when the speaker changes. The method of fusing voice and image bimodal information guarantees the tracking effect of the system under various situations, and obtains a high accuracy of speaker tracking for human-robot interaction.

To widen the interaction range, a way of combining multiple cameras will be developed so that the robot has a 360° range of visual perception. The current motion system will also be replaced by an omnidirectional wheel chassis to improve the robot's flexibility while steering. With these two improvements, the robot's efficiency of speaker localization and tracking will be greatly improved. In addition, human pose detection will also be studied to help determine if a person is speaking. We will focus on the study of speaker tracking of mobile robots to promote the development of initiative service [11].

Acknowledgments

This work was supported in part by the College Students' Innovative Entrepreneurial Training Plan Program under Grant 202110491066, and in part by the Fundamental Research Funds for National University, China University of Geosciences (Wuhan) under Grant CUGDCJJ202246. In addition, we would like to thank Mr. Ying-Zuo Long, Mr. Yue-Wei Jian, Mr. Hao-Ran Jin for their suggestions on this paper.

REFERENCES:

[1] Y. E. Zhu, Y. Yang, "Design of a new robot intelligent control system," China Southern Agricultural Machinery, vol. 50, pp. 147-148+165, September 2019.

[2] Y. E. Zhu, Y. Yang, "Design of a new robot intelligent control system," China Southern Agricultural Machinery, vol. 50, pp. 147-148+165, September 2019.

[3] J. X. Gao, X. G. Ruan, S. C. Ma, "Mobile robot active routing navigation based on fuzzy behavior decision-making," vol. 40, pp. 1308-1314, September 2014.

[4] S. Savio, R. Carvalho, M. Xavier,M. Marciano and A. Ripardo, "New trends on computer vision applied to mobile robot localization," Internet of Things and Cyber-Physical Systems, vol. 2, pp. 63-69, 2022.

[5] Y. T. Li et al., "Research on sound source location system based on microphone array,"Internet of Things technology, vol. 11, pp. 26-28, 2021.

[6] R. W. Zeng, "Disturbance suppression of ecological monitoring data based on microphone array," China Science and Technology Information, vol. 34, pp. 118-120, May 2022.

[7] Y. X. Zhu and H. R. Jin, "Speaker Localization Based on Audio-Visual Bimodal Fusion," Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 25, pp. 375-382, 2021.

[8] R. Wang, "Research on Calibration Methods for Distributed Microphone Arrays," Doctoral Dissertation, Dalian University of Technology, doi:10.26991/d.cnki.gdllu.2021.003823, 2021.

[9] Y. B. Ding, "Face detection Based on Adaboost algorithm and OpenCV," Computer Knowlegde and Technology, vol. 14, pp. 167-169, September 2018.

[10] Z. Y. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 1330-1334, 2000.

[11] M. Hao, W. H. Cao, M. Wu, Z. T. Liu, J. H. She, L. F. Chen, R. Zhang, "Proposal of initiative service model for service robot," CAAI Transactions on Intelligence Technology, vol. 2, pp. 263-272, December 2017.