A Speaker Localization Method Based on Voice and Image Multimodal Fusion

Hao-Ran Jin[1, 2], Chen-Ning Lu[1, 2], Ying-Zuo Long[1, 2], Bao-Han Wu[1, 2], Zhen-Tao Liu[1, 2, †]

[1]School of Automation, China University of Geosciences, Wuhan 430074, China
[2] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China
[†]Corresponding author, E-mail: liuzhentao@cug.edu.cn

Abstract: Single-modal localization technology based on computer hearing or computer vision has the deficiencies of accuracy and stability. Considering that the comprehensive utilization of multimodal information can effectively improve the accuracy and anti-interference of the positioning system, a speaker localization method based on voice and image multimodal fusion is proposed. Firstly, we use the method based on TDOA using microphone array for voice localization and AdaBoost algorithm for face detection separately. Secondly, a multimodal fusion method based on temporal and spatial fusion between voice and image is proposed. After developing the frame rate tracker for temporal fusion, the pixel coordinate system and the world coordinate system are fused for the fusion of the features from face image and voice localization. The proposed method was tested by positioning the speaker stand at 15 different points, and each point was tested for 50 times, from which the experimental results show that there is a high accuracy when speaker standing in front of the positioning system with a distance between 0.5 m to 3.5 m, and an azimuth between -45° to 45°.

Keywords: Speaker positioning, microphone array, image, information fusion

1. INTRODUCTION

For the human-robot interaction system, target positioning is a crucial problem to address. Speaker localization is a comprehensive research, which involves signal acquisition, information processing, modality fusion and so on. Over the past few years, there has been increasing interest in speaker localization, including voice localization [1] and face recognition [2].

Microphone array can be used for voice localization. Salvati et al. [3] proposed a DU transformation method for beam-forming, which achieves robust positioning with low computational complexity. Sun et al. [4] proposed an array signal processing system, which utilizes the low-cost microphone array and works together with general data acquisition card and personal computer. Kim et al. [5] proposed a way of improved sound source localization, which is based on the generalized cross-correlation method.

Besides, face features can also be used for speaker detection. Melek et al. [6] proposed a face recognition method based on sparse representation, which effectively improves the accuracy of face recognition. Liu et al. [7] proposed a method composed of face alignment and 3D reconstruction, which improves the accuracy of face recognition.

The speaker localization based on microphone array, however, would be more affected by various noises, reverberation, other speakers and some special factors in actual environment. Identification of human facial expressions through camera in real-time conditions contains many limitations, such as lighting condition problem, head pose, and occlusion [8].

Compared with the single-modal localization system, the comprehensive utilization of multimodal information can effectively improve the anti-interference ability and accuracy of the positioning system, thus a localization method based on the multimodal fusion of voice and image is proposed to improve the positioning accuracy and stability. The location information of the speaker obtained by the TDOA [9] method and the face localization information obtained from the images are fused from both spatial and temporal aspects. Spatial fusion uses the transformation relationship among physical coordinates, image coordinates and pixel coordinates to determine the speaker's direction. The transformation between the physical coordinate system and the image coordinate system is introduced to fuse the position information of the two modalities, while the conversion between the image coordinate system and the pixel coordinate system is introduced to reduce the imaging error. The frame rate tracker is employed to align the sampled signal in time fusion. What's more, an experimental system which composed by a Kinect and a software is built. Fifteen different points in front of the camera are selected to verify the accuracy of speaker localization. We did 50 tests at each point and calculated the accuracy using the test data. The experimental results

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61976197, No. 61403422 and No. 61273102, in part by the Hubei Provincial Natural Science Foundation of China under Grant No. 2018CFB447 and No. 2015CFA010, in part by the Wuhan Science and Technology Project under Grant 2017010201010133, in part by the 111 project under Grant B17040, and in part by the Fundamental Research Funds for National University, China University of Geosciences (Wuhan) under Grant 1910491T01.

show that there is a high accuracy when speaker standing in front of the positioning system with a distance between 0.5 m to 3.5 m, and an azimuth between -45° to 45°.

The rest of this paper is organized as follows. In Section 2, a speaker localization method based on voice and image multimodal fusion is proposed. Experimental results and analysis are given in Section 3.

2. THE PROCEDURE OF SPEAKER LOCALIZATION METHOD

The procedure of the proposed method is shown in Fig. 1. It's mainly composed of four processes, including signal acquisition, data processing, preliminary positioning and the fusion of voice and image.

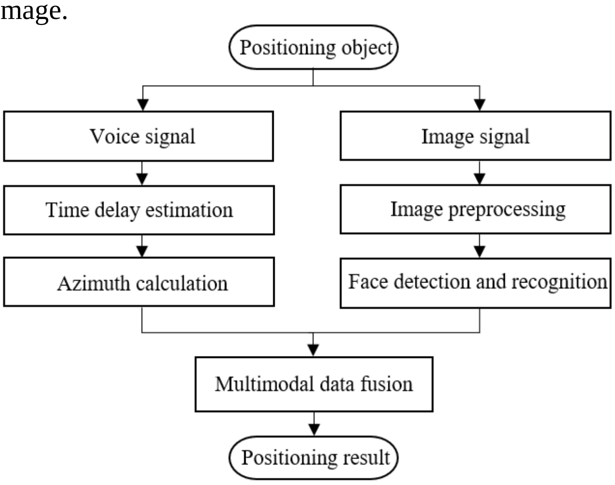


Fig. 1 Procedure of the localization method

2.1 Voice Localization Based on TDOA

The linear microphone array in Kinect and the method of time differences of arrival (TDOA) [9] is used to acquire the voice signal.

The distribution of Kinect's linear microphone array is shown in Fig. 2. The microphone array can achieve advanced sound effects, including noise suppression, echo cancellation and automatic gain control.

The method of voice localization based on TDOA is divided into time delay estimation and position estimation.

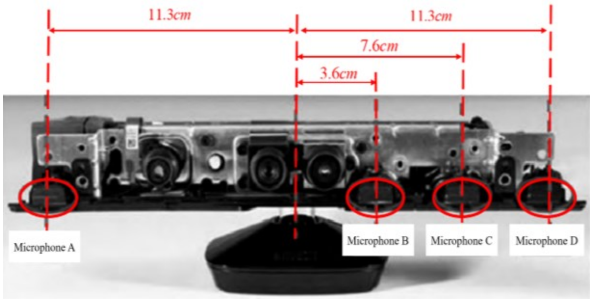


Fig. 2 Distribution of Kinect's linear microphone array

2.1.1 Time Delay Estimation

The main goal of time delay estimation is to calculate the time difference between the audio signals received by two microphones accurately and quickly.

Suppose that the audio signals received by the different microphone M_1 and M_2 are

$$x_1(t) = A_1 s(t-\tau_1) + n_1(t) \tag{1}$$

$$x_2(t) = A_2 s(t-\tau_2) + n_2(t) \tag{2}$$

where $s(t)$ is the voice signal, τ_i is the propagation time from the voice signal to the microphone M_i, A_i is the attenuation factor, and $n_i(t)$ is the noise.

Suppose the correlation function of the two signals $x_1(t)$ and $x_2(t)$ received by the microphone is

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t-\tau)] \tag{3}$$

On the premise that the source signal and noise are independent, the two noises are uncorrelated to each other, and signal $s(t)$ is a stationary random signal, we can substitute (1) and (2) into (3) and get

$$R_{x_1x_2}(\tau) = A_1A_2R_{ss}(\tau-(\tau_1-\tau_2)) \tag{4}$$

when $\tau = \tau_1 - \tau_2$, $R_{x_1x_2}(\tau)$ gets the maximum value, and $\tau_1 - \tau_2$ represents the time difference between the audio signals received by the two microphones. The result of time delay estimation is

$$\hat{\tau} = \arg\max_{\tau} R_{x_1x_2}(\tau) \tag{5}$$

2.1.2 Location Estimation

The location estimation is to calculate the azimuth angle in the microphone array sound field model according to the time difference obtained by the time delay estimation.

In general, when the distance between the speaker and microphone array is larger than the wavelength of signal, it can be considered as far-field case. In this case, the amplitude difference of sound wave can be ignored. If the acoustic wave is simplified as plane wave, it can be approximately considered that there is only a simple delay difference between the ground signals received by the microphone array elements. As shown in Fig. 3, taking the signal received by microphone A as the reference signal, the delay of the signal received by microphone *i* can be calculated as

$$\tau_i = \frac{d\sin\theta}{c} \tag{6}$$

where *d* is the distance between microphone A and microphone *i*, and *c* is the propagation speed of sound wave in the air.

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

$$\theta = \arcsin\frac{L}{d} = \arcsin\frac{\tau_i \times c}{d} \tag{7}$$

Substituting $\hat{\tau}$ into Eq. (7), the expression of θ is

$$\theta = \arcsin\frac{\hat{\tau}\times c}{d} \tag{8}$$

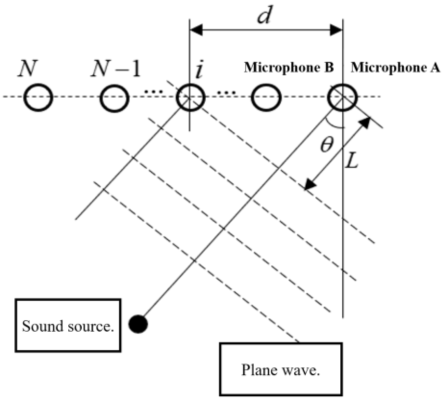


Fig. 3 Far field model of microphone array

2.2 Speaker Localization Based on Computer Vision

Computer vision can also be applied to speaker positioning. The relative processing is as follows.

2.2.1 Image Preprocessing

The input image needs to be preprocessed to eliminate the impact before face detection and localization, because image collection can be affected by noise, light, posture and other factors. The image graying and histogram equalization are used.

1) Image Graying

Image graying is to convert color to gray, remove color changes, and only reflect the brightness changes of the image [10]. The gray image is the basic of subsequent image processing operation [11]. Gray image will not be different due to the change of external light intensity. This feature can effectively remove the invalid information caused by the change of external light.

The conversion formula of grayscale is

$$G = \sqrt[2.2]{\frac{R^{2.2} + (1.5G)^{2.2} + (0.6B)^{2.2}}{1 + 1.5^{2.2} + 0.6^{2.2}}} \tag{9}$$

where R, G, B represent the RGB color value in the pixel of the picture.

2) Histogram Equalization

Histogram of the image is the graphical representation of the probability of occurrences of the intensities versus Intensity values in the given image [12]. Histogram equalization is to transform the histogram of face image into a state of uniform distribution, to increase the image contrast and make the image clearer. The processed image is more suitable for application than the original image.

2.2.2 Face Detection Based on AdaBoost Algorithm

Haar feature and AdaBoost algorithm can be used to train face detection cascade classifier. The algorithm is that different training samples are used to train the same weak classifier, and the best combination of the trained weak classifiers is cascaded to form a strong classifier.

The training of classifier is generally divided into the following three steps：

Step 1: Extraction of Haar features and calculation of eigenvalues

Fig. 4 is a common Haar feature template diagram. Place and slide the feature template over the images so that the images can be entirely covered by it. Calculate each Haar feature in turn to get a set of data, which is the Haar eigenvalue of the image.

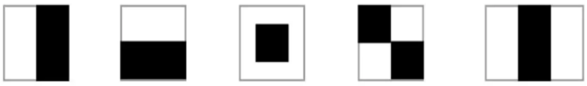


Fig. 4 Commonly used Haar feature template

Step 2: Initializing and training weak classifiers

Each weak classification contains four parameters, i.e., unequal sign direction p, threshold θ, feature f and detection window x. The classifier parameters are obtained from the position information of the corresponding Haar features. According to the Haar eigenvalues calculated in the previous step, the initialization process is to generate several weak classifiers and assign initial values to the classifier, thus it can obtain the initial classifier $H(x, f, \theta, p)$ as

$$H(x, f, \theta, p) = \begin{cases} 1 & pf(x) < p\theta \\ 0 & \text{other} \end{cases} \tag{10}$$

The optimal threshold θ is obtained by training the weak classifier, which can minimize the classification error of all samples passing through the classifier.

Step 3: Calling the AdaBoost algorithm to train a strong classifier

A series of weak classifiers $H_1, H_2, \cdots, H_n$ with the best performance are obtained by iterating and updating the weights of weak classifiers. In the process of filtering, the number of iterations is equal to the number of the best classifiers. Finally, the weighted voting mechanism is used to construct a strong classifier. The weight of the weak classifier with good classification performance is larger than that of the weak classifier with poor classification performance.

In the actual system, the image collected by Kinect's color camera and the face detection classifier are shown in Fig. 5.

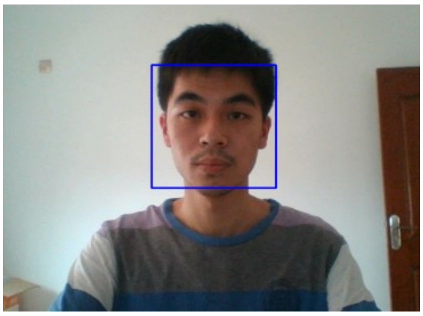


Fig. 5 Effect picture of face detection

2.3 The Fusion Method of Voice and Image

Data fusion can increase the redundancy of positioning information, so as to improve the stability and accuracy of the system [13]. The data fusion model of microphone array and camera mainly consists of two parts, spatial fusion and time fusion. Spatial fusion is designed to improve positioning accuracy, and the core idea of time fusion is to align data.

2.3.1 Spatial Fusion of Microphone Array and Camera Data

Voice localization is indicated in the physical coordinate system, while the face detection is indicated in the image coordinate system. Since image information can be converted from 2D pixel coordinates to 3D physical coordinates [14], a sensitive region can be set as image coordinate system. The conversion relationship is shown in Fig. 6.

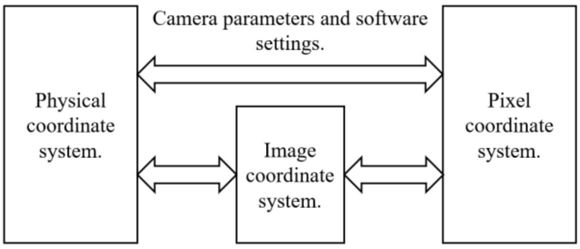

Camera parameters and software settings.
Physical coordinate system.
Image coordinate system.
Pixel coordinate system.

Fig. 6 Coordinate transformation diagram

A Kinect sensor, which combines a microphone array with a camera, is used for positioning speakers. Therefore, the projection of the physical coordinate system's origin can be approximated as the origin of the image coordinate system. The image model of the camera is shown in Fig. 7, where $O_w - X_w Y_w Z_w$ represents the world coordinate system, xO_1y represents the image coordinate system, O_1 is the optical center of the camera lens, Z_w is the optical axis, and uO_0v represents the pixel coordinate system.

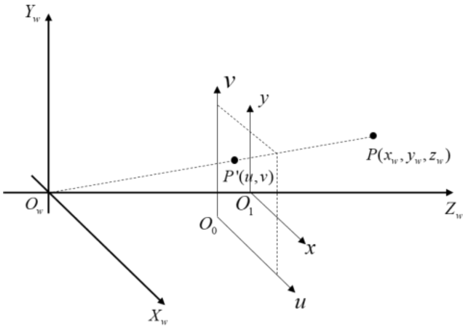


Fig. 7 Camera imaging model

Supposing a point $P(x_w, y_w, z_w)$ in the physical coordinate system, the steps to convert the point to the corresponding point $P'(u,v)$ in the pixel coordinate system are as follows.

Step 1: Convert physical coordinate system to image coordinate system

Point *P* is imaged by the lens at point (x, y) of a 2D planar image. According to the geometric relationship of optical imaging, the results are

$$x = \frac{X_w \times f}{Z_w}, \quad y = \frac{Y_w \times f}{Z_w} \tag{11}$$

The formula is transformed into matrix form as

$$Z_w \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{12}$$

where f is the focal length.

Step 2: Convert image coordinate system to pixel coordinate system

In practice, the origin of the image coordinate system often deviates from the center of the pixel coordinate system due to errors. Assuming that the actual coordinate of origin in the pixel coordinate system is (u_0, v_0) and the offset distance is d_u , d_v . the conversion formula between image coordinate (x, y) and pixel coordinate (u, v) is

$$u = \frac{x}{d_u} + u_0 \ , \quad v = \frac{y}{d_v} + v_0 \tag{13}$$

The formula is transformed into matrix form as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{d_u} & 0 & u_0 \\ 0 & \frac{1}{d_v} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{14}$$

Substitute Eq. (14) into Eq. (12) we get

$$Z_w \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f}{d_u} & 0 & u_0 & 0 \\ 0 & \frac{f}{d_v} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{15}$$

where f, d_u, d_v, u_0, v_0 are determined by the camera internal parameter. The transformation from physical coordinate system to pixel coordinate system can be realized by the formula above. And through the coordinate transformation in the experimental system, the coordinates of the voice localization origin based on microphone array are adjusted.

2.3.2 Time Fusion of Microphone Array and Camera Data

The temporal fusion of microphone array and camera's data can be interpreted as the synchronous transmission and processing of both data in time. The time fusion part of this system is implemented by the frame rate tracker.

The sampling frequency of Kinect microphone array is 16 KHz, and the frame rate of the image captured by the camera is 30 fps. The frame rate tracker can keep the speed of image's acquisition at 30 fps, continuously detect and process the voice signals, thus the data of the two sensors can be transmitted and processed simultaneously.

3. EXPERIMENTS

3.1 Experimental Setting

The hardware used in this system is the Kinect developed by Microsoft, which is a peripheral device equipped with camera and microphone array, and can enable users to control the computer with voice commands.

C++ is used as programming language, and Microsoft official SDK as medium. The whole system that controls and reads the input and output of the Kinect runs on a PC with Windows 10. The data flow between software and hardware is shown in Fig. 8.

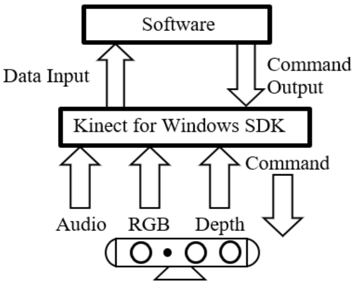
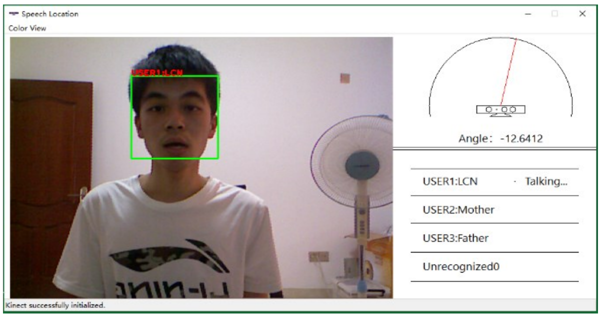


Fig. 8 Data flow between software and hardware

The real-time display interface of the system is shown in Fig. 9. The image captured by the camera of Kinect is displayed on the left. The voice localization and face recognition of the speaker is displayed in right side. The orientation relationship between the speaker and the hardware's center is expressed as the form of a pointer rotation.



Kinect successfully initialized.

Fig. 9 Interface Display

The experiment is carried out indoor with the space of $9\ \mathrm{m}\times 4.5\ \mathrm{m}\times 3\ \mathrm{m}$. There is only one speaker that is used as the target for system positioning but there are many noise sources and echoes during the experiments. The effective interaction range and the reliability of the positioning system are tested by multi-points and multi-times positioning.

The remaining test points' coordinate are set as shown in Table 1.

Table 1 The test points' coordinate

A	B	C	D	E
(0,0.3)	(0,1.0)	(0,1.5)	(0,2.0)	(0,6.0)
F	**G**	**H**	**I**	**J**
(0.5,1.8)	(1.5,1.0)	(1.5,3.5)	(2.0,4.0)	(2.0,4.0)
K	**L**	**M**	**N**	**O**
(-0.8,0.5)	(-0.8,1.0)	(-1.8,0.6)	(-1.8,2.5)	(-2,6)

a. The unit of coordinates is the meter

The location information of 15 test points is shown in Fig. 10. The central point of Kinect is located at (0,0).

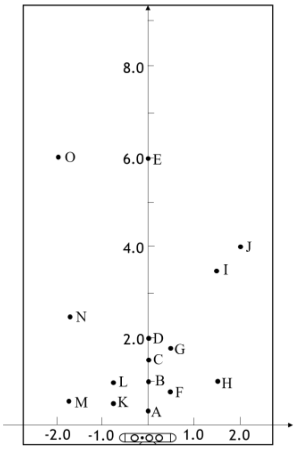


Fig. 10 The location information of 15 test points

3.2 Experimental Results and Analysis

The experimental results are shown in Table 2.

Table 2 Experimental results

Test points	Actual localization		Texted results		
	$(x/\mathrm{m}, y/\mathrm{m})$	& θ	$\hat{\theta}$	α	β
A	(0.0, 0.3)	0.0°	Error	0%	0%
B	(0.0, 1.0)	0.0°	4.51°	97.5%	100%
C	(0.0, 1.5)	0.0°	2.37°	100%	100%
D	(0.0, 2.0)	0.0°	4.81°	100%	48%
E	(0.0, 6.0)	0.0°	Error	0%	0%
F	(0.5, 0.8)	32.0°	30.08°	100%	100%
G	(0.5, 1.8)	15.5°	17.56°	100%	90%
H	(1.5, 1.0)	56.3°	52.00°	20%	100%
I	(1.5, 3.5)	23.2°	22.87°	100%	30%
J	(2.0, 4.0)	26.6°	23.92°	46%	0%
K	(-0.8, 0.5)	−57.99°	Error	0%	100%
L	(-0.8, 1.0)	−38.66°	−39.74°	100%	100%
M	(-1.8, 0.6)	−71.57°	Error	0%	100%
N	(-1.8, 2.5)	−35.75°	−31.16°	100%	18%
O	(-2.0, 6.0)	−18.44°	Error	0%	0%

The data in Table 2 show the actual localization and the results measured by the speaker localization system at 15 test points. In Table 2 x, y is the real coordinate, and θ is the real azimuth. After an average of 50 tests, the test azimuth is represented by $\hat{\theta}$, the success rate of positioning recognition is represented by α and β, and the test azimuth with A deviation of 5° from the true value is regarded as an error.

According to the data in Table 2, the system would get a high reliability when the speaker's position had a distance between 0.5 m to 3.5 m from the central point, and an azimuth between -45° to 45°. But for the voice signal beyond this range, the positioning accuracy is obviously reduced or even invalid.

The analysis of invalid recognition and errors of the positioning system are as follows.

(1) When the source signal is located directly in front of the Kinect's central line, the actual time-delay of the signal received between each element is small. Due to the limitation of Kinect's sampling frequency, the error will be occurred when the minimum time difference of arrival is less than 62.5 μs.

(2) Although the built-in hardware and software of Kinect have excellent performance in noise elimination and reverberation suppression, the processed audio still has some small interference.

(3) When the speaker gets too close or too far from the Kinect, images captured by the camera would not include the face or be too small. That's why the highest recognition rate is obtained when the distance between speaker and Kinect is in the range of 0.5 m to 3.5 m.

4. CONCLUSION

A speaker localization method based on voice and image multimodal fusion was proposed to achieve higher speaker localization accuracy during human-robot interaction. The proposed was tested in an indoor spacer by building the localization system. The experimental results show that our method would get a high accuracy of speaker localization for human-robot interaction if the distance between speaker and localization system is ranged from 0.5 m to 3.5 m and the azimuth range is ranged from -45° to 45°. The positioning accuracy was improved by fusing the voice and image modal information.

In addition, a rotating motor platform can be developed to widen the range of the effective azimuth in this system, and a multi-sound source acquisition and processing can be developed to increase the systems' ability of locating multiple speakers.

REFERENCES

[1] A. Sepas-Moghaddam, F. M. Pereira and P. L. Correia, "Face recognition: a novel multi-level taxonomy based survey," in IET Biometrics, vol. 9, no. 2, pp. 58-67, Mar. 2020.

[2] J. M Fresno, G. Robles, J. M. Martínez-Tarifa, and B. G. Stewart, "Survey on the Performance of Source Localization Algorithms," Sensors, vol. 17, no. 11, pp. 2666, 2017.

[3] D. Salvati, C. Drioli and G. L. Foresti, "A Low-Complexity Robust Beamforming Using Diagonal Unloading for Acoustic Source Localization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 3, pp. 609-622, Mar. 2018.

[4] Y. Sun, X. Yang, L. Guo and T. Long, "Experimental array signal processing demonstration system by utilizing microphone array," 2016 CIE International Conference on Radar (RADAR), pp. 1-5, 2016.

[5] U H. Kim, K. Nakadai, H G. Okuno. "Improved Sound Source Localization and Front-Back Disambiguation for Humanoid Robots with Two Ears". in Proceedings of the 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 282-291.2013

[6] M. Melek, A. Khattab and M. F. Abu-Elyazeed, "Fast matching pursuit for sparse representation-based face recognition," in IET Image Processing, vol. 12, no. 10, pp. 1807-1814, October. 2018.

[7] F. Liu, Q. Zhao, X. Liu and D. Zeng, "Joint Face Alignment and 3D Face Reconstruction with Application to Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 3, pp. 664-678, 1 Mar. 2020.

[8] K. M. Kudiri, A. M. Said and M. Y. Nayan, "Human emotion detection through speech and facial expressions," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), pp. 351-356, 2016.

[9] J. Qu, H. Shi, N. Qiao, C. Wu, C. Su, A. Razi, "New three-dimensional positioning algorithm through integrating TDOA and Newton's method," EURASIP Journal on Wireless Communications and Networking, vol 77, pp 1-8, 2020.

[10] Y. He, B. Jin, Q. Lv and S. Yang, "Improving BP Neural Network for the Recognition of Face Direction," 2011 International Symposium on Computer Science and Society, pp. 79-82,2011.

[11] X. Zhang and X. Wang, "Novel Survey on the Color-Image Graying Algorithm," 2016 IEEE International Conference on Computer and Information Technology (CIT), pp. 750-753, 2016.

[12] S. Patel and M. Goswami, "Comparative analysis of Histogram Equalization techniques," 2014 International Conference on Contemporary Computing and Informatics (IC3I) pp. 167-168, 2014

[13] E. D'Arca, N. M. Robertson and J. Hopgood, "Person tracking via audio and video fusion," in 9th IET Data Fusion & Target Tracking Conference (DF&TT 2012): Algorithms and Applications, pp. 1-6, 2012.

[14] Z. Zhang, "A flexible new technique for camera calibration," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330-1334, Nov. 2000.