IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 5, 2020, accepted November 23, 2020, date of publication December 1, 2020,
date of current version December 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041445

Acoustic Enhanced Camera Tracking System Based on Small-Aperture MEMS Microphone Array

LEI LI, KECHAO LIAN, JINTAO FU, PENGFEI ZHU, ZHIYONG HU, AND CE GUO
School of Physics and Microelectronics, Zhengzhou University, Zhengzhou 450001, China

Corresponding author: Lei Li (lilei@zzu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61601322, and in part by the Henan Provincial Key Science and Technology Research Project under Grant 162102210018.

ABSTRACT The camera tracking systems based on visual image processing face a problem that they are completely ineffective in their blind zones. To address this problem, a design of acoustic enhanced tracking system combining visual and auditory target tracking methods is reported in this article. The system holds the abilities of performing sound direction estimation and target tracking in real-time. Estimating direction of arrival of the sound accompanied with the target helps the camera turn towards the target outside the field of view. This sound-triggered mode of camera operation makes a significant supplement to conventional cameras' working state. Considering the embedded system is necessary in consideration of the cost and size of the system in practical application, we designed a small aperture array with 7 digital omnidirectional MEMS microphones and built the overall system based on FPGA and ARM. The experiments were carried out in a normal indoor environment and the results confirmed that the system can perform auditory and visual tracking in real-time.

INDEX TERMS Visual target tracking, MEMS microphone array, acoustic localization, computer perception.

I. INTRODUCTION

Real-time target tracking, as a basic core technology in the field of computer perception, is widely employed in many applications, such as intelligent robot [1], security monitoring [2], and drone detection [3]. Among these applications, the detection and tracking of moving targets have been attracting great research interest. In the past few decades, relevant researchers have made great progress in moving targets tracking based on video sequence [4]–[9]. Actually, the development of the visual tracking algorithm has not dealt with the problem existing in conventional camera monitoring. These cameras are completely ineffective for visual tracking while the object is out of their field of view (FOV). As shown in FIGURE 1, except for vision, hearing accounts for the most proportion of all human perception, about 13% [10]. A person who hears what is happening out of sight will turn his neck to see the object. Imitating these joint actions from hearing to vision, we developed a tracking system with auditory and visual methods to address the blind angle problem. The functions of ears, eyes, neck and brain can be imitated by microphones, a camera, a pan-tilt actuator and CPU, respectively. However, both the integration of such embedded system and the implementation of effective algorithms bring great challenges for this design in practical applications.

In the field of acoustic source localization, the methods based on microphone array (MA) are very popular, which estimate the sound direction by processing the spatial information [11]–[21]. These methods can be generally divided into three categories: beamforming based methods [11], [12], [16], subspace based methods [17]–[19], and parametric methods [20], [21]. Parametric methods feature high computational cost so that they are not suitable for real-time processing. The subspace based methods are characterized by high resolution, but these methods only care about the direction result, without a synthesized output. That makes them not a better choice than beamforming in terms of

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

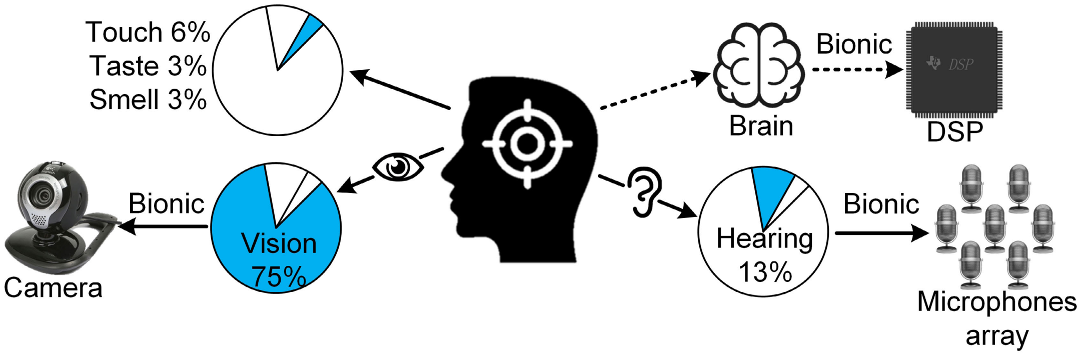


FIGURE 1. The human perception system.

functionality. Additionally, the beamforming based methods are small in computation, which gives beamforming an advantage in real-time processing. The direction information is mapped to the phase delay of signals received by each array element. The detailed explanations of phased array beamforming technologies are well documented in [22]–[24]. Adaptive beam-forming is a significant task in array signal processing, of which the Capon beamformer is a representative example [25]–[27]. The signal of interest (SOI) is allowed to pass through without distortion while the interference signals and noise are suppressed as much as possible. However, it has been found that the Capon beamformer is sensitive to modelling mismatches [28]–[33], especially in small aperture arrays. To address this problem, a class of robust adaptive beamformers (RABF) are designed to offer acceptable array output performance [34]–[41]. It was found that the influence of array elements mismatch on locating accuracy in small aperture arrays can be reduced by adjusting the weighting vector of beamformer. Thus, RABF algorithms provide a solution for the acoustic localization of this design.

It is necessary to integrate the system in practical application. But the array of traditional electret condenser microphone (ECM) is often featured by larger size and power consumption. Furthermore, additional amplification circuits and AD acquisition modules also complicate the system. All these factors above bring great difficulties to the integration of the system. Fortunately, the emergence of the Micro-Electro-Mechanical Systems (MEMS) microphone makes it possible to miniaturize the acoustic sensor array [42], [43]. The MEMS microphones have an acoustic transducer, an amplifier, and even an analog-to-digital converter (ADC) integrated in the chip [44], that contributes to the MA's small aperture. They can be directly connected to the FPGA through the I^{2}S interface, without using an audio decoder, enabling the further reducing of system complexity. Additionally, compared to ECM, MEMS microphones have less sensitivity to temperature, vibrations or mechanical shocks [45], [46]. These advantages including high quality and small package, make MEMS MA more portable and suitable for our design, that brings solutions for scheme of small-aperture MA.

The combination of visual and auditory information processing has attracted the interests of many researchers [47]–[53]. A perception sensor net-work [47] capable of detecting emergency situations was presented for school safety, using a Kinect with four microphones to acquire audio signals. D'Arca *et al.* [48] used distributed directional microphones to recognize speakers, combined with video information captured by the camera. Viciana-Abad *et al.* [49] proposed an audio-visual perception system to direct the behavioral responses of the robot with two microphones and two cameras attached to the head. Wilson *et al.* [50] combined a video camera array and a MA to locate the speaker in a conference room, with 32 omnidirectional microphones spread across the ceiling and 2 cameras on adjacent walls. Despite the great effort in studying the audio-visual information processing, it has rarely been reported that using such a small aperture MEMS MA to solve the camera blind spot problem. As described before, in practical applications, embedded implementation is necessary to meet the cost and size constraints.

In this article, we have reported an acoustic enhanced camera tracking system based on a small-aperture MEMS MA, in order to extend the detection angle of the camera tracking system to all directions, imitating the hearing-vision interactions of human. To address the integration problem, a circular small-aperture MEMS MA is designed with a 4.5 cm radius. The embedded platform is constructed based on FPGA and ARM for data parallel acquisition and system control, respectively. The estimation results confirmed that the whole system can perform all the localization and tracking functions reliably in real-time.

This article is organized as follows: Section II describes the related algorithms implemented. Section III shows the system architecture. The experiments and discussions to evaluate the performance of the system are explained in Section IV.

II. METHODS

In this section, the algorithms used in the system are introduced, including the voice activity detection (VAD), the sound source localization, visual detection and target tracking.

A. VOICE ACTIVITY DETECTION AND SOUND SOURCE LOCALIZATION

1) VOICE ACTIVITY DETECTION

In order to reduce the computing burden as well as provide trigger signals for the tracking system, the VAD processing is implemented. The VAD is based on the estimation of the short-term energy and short-term zero-crossing rate (ZCR). The short-term energy of an n-th frame audio signal can be given by:

$$E_n = \sum_{m=0}^{N-1} [\omega(m)x(n+m)]^2 \tag{1}$$

where N is the window length; $x(n)$ is the audio signal; $\omega(m)$ is given by the following equation:

$$\omega(m) = \begin{cases} 1 & m = 0, 1, \cdots, (N-1) \\ 0 & \text{others} \end{cases} \tag{2}$$

And ZCR is expressed as:

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1} |sgn[x_n(m)] - sgn[x_n(m-1)]| \tag{3}$$

where

$$sgn[x] = \begin{cases} 1 & (x \geqslant 0) \\ -1 & (x < 0) \end{cases} \tag{4}$$

2) SOUND SOURCE LOCALIZATION

In this work, the beamforming algorithm is used to estimate the direction of arrival (DOA) of the sound. This method is based on the time difference of arrival signals of the microphone array elements. The time difference in time domain is reflected as phase shift in frequency domain. The accuracy of delay in the time domain is limited by the sampling rate, while the accuracy of phase shifting in the frequency domain can be higher. At the same time, most of the natural sound are broadband signals, which need to be analyzed after being decomposed. Therefore, the fast Fourier transform (FFT) processing is necessary.

If the wavelength of the signal is known, the phase has a corresponding relationship with the angle of arrival. The output power of the SOI is maximized when the most suitable compensation phase making the signals of all channels become coherent signals is found. Thus, we establish the relationship between output power and the direction of arrival.

Supposing that $\boldsymbol{a}(\theta, \varphi)$ is the direction vector of a plane wave propagation in space:

$$\boldsymbol{a}(\theta, \varphi) = -[\sin\varphi\cos\theta \quad \sin\varphi\sin\theta \quad \cos\varphi]^T \tag{5}$$

where θ and φ are azimuth and elevation angle in spherical coordinates, respectively. In this design, we only care about the azimuth angle θ. So, the value of φ is treated the same, and $\boldsymbol{a}(\theta, \varphi)$ is simplified to $\boldsymbol{a}(\theta)$ to facilitate the expression.

Supposing that the array elements number is M and the position of each is $\boldsymbol{P}_m(m = 0, 1, \cdots, M-1)$, then the time difference between m-th array element and reference point is given by

$$\tau_m = \frac{\boldsymbol{a}^T \boldsymbol{P}_m}{c} \tag{6}$$

where c represents the velocity of the plane wave and $\boldsymbol{a}^T$ is the transpose of vector $\boldsymbol{a}$.

The signal received by each array element is given by $S_m(t) = S(t - \tau_m)$, where $S(t)$ is the signal received at the reference point. So its frequency spectrum is

$$S_m(\omega) = \int_{-\infty}^{\infty} S(t-\tau_m)e^{-j\omega\tau}d\tau = S(\omega)e^{-j\omega\tau_m} \tag{7}$$

where j and ω are imaginary unit and frequency, respectively.

Define the wave number $\boldsymbol{k} = \omega\boldsymbol{a}/c$, then there is $\omega\tau_m = \boldsymbol{k}^T\boldsymbol{P}_m$. The receive signal matrix can be expressed as

$$\boldsymbol{X}(\omega) = \begin{bmatrix} S_0(\omega) \\ S_1(\omega) \\ \vdots \\ S_{M-1}(\omega) \end{bmatrix} = S(\omega)\begin{bmatrix} e^{-j\boldsymbol{k}^T\boldsymbol{P}_0} \\ e^{-j\boldsymbol{k}^T\boldsymbol{P}_1} \\ \vdots \\ e^{-j\boldsymbol{k}^T\boldsymbol{P}_{M-1}} \end{bmatrix} = S(\omega)\boldsymbol{v} \tag{8}$$

where $\boldsymbol{v}$ is array steering vector, that is the function of $\boldsymbol{k}$, describing the response of the array to the signal in spatial domain.

Actually, $\boldsymbol{X}(\omega)$ is the output matrix of the array. The conventional beamformer $\boldsymbol{H}^T(\omega)$ is used to compensate for the phase difference. It is given by

$$\boldsymbol{H}^T(\omega) = \frac{1}{M}\boldsymbol{v}_s^H \tag{9}$$

where $\boldsymbol{k}_s$ is the wave number of the plane wave we are interested in. The output is

$$\boldsymbol{Y}(\omega) = \boldsymbol{H}^T(\omega)\boldsymbol{X}(\omega) \tag{10}$$

Supposing that the signal is a unit of power, i.e., $S(\omega) = 1$. The Eq. (10) is defined as the beam pattern to describe the array corresponds to a unit power plane wave signal in space.

In the narrow band snapshot model processing at ω_c, the output of the beamformer is given by

$$\boldsymbol{y}(n) = \boldsymbol{\omega}^H\boldsymbol{x}(n) \tag{11}$$

where complex weight vector $\boldsymbol{\omega}^H = \boldsymbol{H}^T(\omega_c)$ and ω_c represents the central frequency of the narrow band.

Beam output power can be calculated as follows:

$$\begin{aligned} P &= E[\,\boldsymbol{y}(n)\boldsymbol{y}^H(n)] \\ &= E[\,\boldsymbol{\omega}^H \cdot \boldsymbol{x}(n) \cdot (\boldsymbol{\omega}^H \cdot \boldsymbol{x}(n))^H] \\ &= \boldsymbol{\omega}^H \cdot \boldsymbol{R}_x \cdot \boldsymbol{\omega} \end{aligned} \tag{12}$$

where $\boldsymbol{R_x}$ is the matrix of the input signal and it can be expressed as:

$$\boldsymbol{R_x} = E[\boldsymbol{x} \cdot \boldsymbol{x}^H] \tag{13}$$

Also, $\boldsymbol{\omega}$ is a function of θ and φ. In this application, only the azimuth angle θ is concerned. So the Eq. (12) becomes

$$P(\theta) = \boldsymbol{\omega}^H(\theta) \cdot \boldsymbol{R_x} \cdot \boldsymbol{\omega}(\theta) \tag{14}$$

The angular distribution of power can be represented as power azimuth spectrum (PAS) by beam scanning in azimuth. The azimuth of maximum power is the direction of the sound.

The standard Capon beamforming (SCB) or minimun variance distortionless response (MVDR) beamforming algorithm can be summarized as that the SOI is allowed to pass through without distortion while the interference signals and noise are suppressed as much as possible. It is given by the following constraints:

$$\min_{\boldsymbol{\omega}} \boldsymbol{\omega}^H \boldsymbol{R} \boldsymbol{\omega} \quad \text{subject to} \quad \boldsymbol{\omega}^H \boldsymbol{v}_s = 1 \tag{15}$$

The Lagrange multiplier methodology is used to solve formula above, and the weight vector of MVDR is obtained as follows:

$$\boldsymbol{\omega}_{MVDR} = \frac{\boldsymbol{R}^{-1}\boldsymbol{v}_s}{\boldsymbol{v}_s^H \boldsymbol{R}^{-1} \boldsymbol{v}_s} \tag{16}$$

that is substituted into Eq. (14) to get the power estimate:

$$P_{MVDR} = \frac{1}{\boldsymbol{v}_s^H \boldsymbol{R}^{-1} \boldsymbol{v}_s} \tag{17}$$

However, the actual array steering vector often has a certain deviation, causing power loss of the SOI. The RABF algorithm estimates the true steering vector $\bar{\boldsymbol{v}}$ and replaces it with the estimated $\boldsymbol{v}$. The smaller the deviation between $\boldsymbol{v}$ and $\bar{\boldsymbol{v}}$, the larger the power output value of beamforming. Therefore, the RABF algorithm based on array steering vector estimation can be converted to the following quadratic optimization:

$$\max_{\boldsymbol{v}} \frac{1}{\boldsymbol{v}^H \boldsymbol{R}^{-1} \boldsymbol{v}} \quad \text{subject to} \quad \|\boldsymbol{v} - \bar{\boldsymbol{v}}\|^2 = \boldsymbol{\varepsilon} \tag{18}$$

where $\boldsymbol{\varepsilon}$ is the upper norm of the error of the steering vector, which depends on the error between the theoretical and the actual steering vector. Similarly, this problem can be solved by using the Lagrange multiplier methodology, that is given by

$$f(\boldsymbol{v}, \lambda) = \boldsymbol{v}^H \boldsymbol{R}^{-1} \boldsymbol{v} + \lambda(\|\boldsymbol{v} - \bar{\boldsymbol{v}}\|^2 - \boldsymbol{\varepsilon}) \tag{19}$$

in which $\lambda \geqslant 0$ is the Lagrange multiplier. Find the partial derivative of the above equation with respect to $\boldsymbol{v}$, and let the derivative be 0, then get the best steering vector estimate as:

$$\hat{\boldsymbol{v}} = \left(\frac{\boldsymbol{R}^{-1}}{\lambda} + \boldsymbol{I}\right)^{-1} \bar{\boldsymbol{v}} = \bar{\boldsymbol{v}} - (I + \lambda \boldsymbol{R})^{-1} \bar{\boldsymbol{v}} \tag{20}$$

Substitute the above equation into constraint $\|\boldsymbol{v} - \bar{\boldsymbol{v}}\|^2 = \boldsymbol{\varepsilon}$, there is

$$g(\lambda) \triangleq \left\| (I + \lambda \boldsymbol{R})^{-1} \bar{\boldsymbol{v}} \right\|^2 = \boldsymbol{\varepsilon} \tag{21}$$

Let

$$\boldsymbol{R} = \boldsymbol{U} \boldsymbol{\Gamma} \boldsymbol{U}^H \tag{22}$$

where $\boldsymbol{U}$ consists of eigenvectors of $\boldsymbol{R}$, and the eigenvalues of $\boldsymbol{R}$ constitute the diagonal elements of the diagonal matrix $\boldsymbol{\Gamma}$, in which $\gamma_1 \geqslant \gamma_2 \geqslant \cdots \geqslant \gamma_M$. Let

$$\boldsymbol{z} = \boldsymbol{U}^H \bar{\boldsymbol{v}} \tag{23}$$

The solution λ to Eq. (21) is unique and it can be calculated that λ belongs to the following interval:

$$\frac{\|\bar{\boldsymbol{v}}\| - \sqrt{\varepsilon}}{\gamma_1 \sqrt{\varepsilon}} \leqslant \lambda \leqslant \min\left\{ \frac{\|\bar{\boldsymbol{v}}\| - \sqrt{\varepsilon}}{\gamma_M \sqrt{\varepsilon}}, \left(\frac{1}{\varepsilon} \sum_{m=1}^{M} \frac{\|z_m\|^2}{\gamma_m^2} \right)^{\frac{1}{2}} \right\} \tag{24}$$

where z_m denotes the m-th element of $\boldsymbol{z}$. Once λ is determined, the best steering vector $\hat{\boldsymbol{v}}$ can be calculated by Eq. (20). Substitute the estimate of the steering vector into Eq. (16) to obtain the weight vector:

$$\begin{aligned} \boldsymbol{\omega}_{RABF} &= \frac{\boldsymbol{R}^{-1}\hat{\boldsymbol{v}}}{\hat{\boldsymbol{v}}^H \boldsymbol{R}^{-1} \hat{\boldsymbol{v}}} \\ &= \frac{\left(\boldsymbol{R} + \frac{1}{\lambda}\boldsymbol{I}\right)^{-1} \bar{\boldsymbol{v}}}{\bar{\boldsymbol{v}}^H \left(\boldsymbol{R} + \frac{1}{\lambda}\boldsymbol{I}\right)^{-1} \boldsymbol{R} \left(\boldsymbol{R} + \frac{1}{\lambda}\boldsymbol{I}\right)^{-1} \bar{\boldsymbol{v}}} \end{aligned} \tag{25}$$

Finally, the power estimate can be calculated by Eq. (17), as follows:

$$P_{RABF} = \frac{1}{\|\bar{\boldsymbol{v}}\|^2} \frac{\sum\limits_{i=1}^{M} \left(\frac{\lambda\gamma_m}{1+\lambda\gamma_m}\right)^2 |z_i|^2}{\sum\limits_{i=1}^{M} \left(\frac{\lambda}{1+\lambda\gamma_m}\right)^2 \gamma_m |z_i|^2} \tag{26}$$

The overall broadband beamforming algorithm can be illustrated as FIGURE 2. The VAD processing is to distinguish whether the sound signal is valuable, so it should run on all 7 channels before beamforming. The received signal of each array element is preprocessed and the data shell is decomposed into K subbands by discrete Fourier transform (DFT), and then signal in each subband can be treated as a narrow band to perform beamforming operation. Finally, we can obtain the output signal in time domain by inverse discrete Fourier transform (IDFT).

K-point DFT is performed on the M sampling channels in order to get K frequency bins. A new vector of the result in the same frequency bin is made up so that K vectors can be expressed as:

$$\boldsymbol{X}(k) = \begin{bmatrix} \boldsymbol{X}_0(k) \\ \boldsymbol{X}_1(k) \\ \vdots \\ \boldsymbol{X}_{M-1}(k) \end{bmatrix} \quad k = 0, 1, \cdots, K-1 \tag{27}$$

where $X_m(k)$ is the signal of m-th (m=1,2,$\cdots$,M-1) array element in the frequency domain.

After M signals are weighted and summed, the results of K frequency bands can be obtained as follow:

$$\boldsymbol{Y}(k) = \boldsymbol{\omega}^H(k) \cdot \boldsymbol{X}(k) \tag{28}$$

where $\boldsymbol{\omega}^H(k)$ is the weight vector in each subband, it can be expressed as:

$$\boldsymbol{\omega}(k) = \begin{bmatrix} \boldsymbol{\omega}_0(k) \\ \boldsymbol{\omega}_1(k) \\ \vdots \\ \boldsymbol{\omega}_{M-1}(k) \end{bmatrix} \quad k = 0, 1, \cdots, K-1 \tag{29}$$

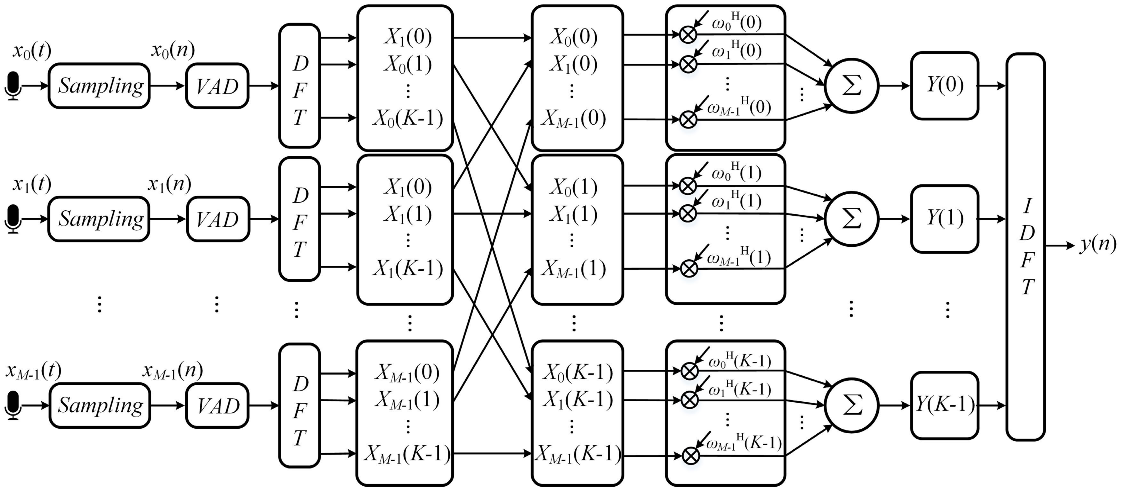


FIGURE 2. The overall data flow of the broadband beamforming algorithm.

K-point inverse Fourier transform of $\boldsymbol{Y}(k)$ is performed to obtain the beam output $\boldsymbol{y}(n)$ in the time domain. Calculate the beamforming output power and scanning power-azimuth spectrum can be plotted.

B. VISUAL TARGET DETECTION AND TRACKING

According to the result of acoustic localization, the pan-tilt camera turns towards the target outside the camera's FOV. Then the frame-difference method is used to detect the moving object and the mean-shift tracking algorithm is applied to track the detected target.

1) MOVING TARGET DETECTION BY FRAME-DIFFERENCE

The principle of the frame-difference method is to subtract two adjacent frames of video images in order to detect the moving target. The difference of two adjacent frames of video images is expressed in the form of

$$\Delta I(x, y) = |I(x, y, t) - I(x, y, t-1)| \tag{30}$$

where $I(x, y, t)$ represents the gray value at the (x, y) point and time t, and $\Delta I(x, y)$ is the difference of grayscale of the two adjacent frames of images. Then binarize the difference image so that we can detect the target and locate its position.

2) MEAN-SHIFT TRACKING

After the target position is located by frame-difference, the mean-shift algorithm is used for its tracking. The algorithm tracks the target motion using gradient information, with linear convergence rate, which makes the iterative calculations small and it is easy to be applied in real-time. The mean-shift vector about the target model is described according to the candidate model with the largest similarity to the target, which is the vector of the target movement. Due to the fast convergence of the mean-shift algorithm, by continuously iteratively computing the mean-shift vector, the algorithm will eventually converge to the true position of the target. The mean-shift algorithm takes the moving target obtained by the frame-difference method as the tracking target, and the RGB feature is used to model the probability density estimation.

Suppose that there are n pixels in the target area which position is $x_i(i = 1, \cdots, n)$. The color space of the target is divided into m intervals. The probability density of the target model is expressed as $\boldsymbol{q} = [q_1 \quad q_2 \quad \cdots \quad q_m]^T$. Suppose that x is the position of the center of the target region. In each interval it is given by

$$q_u = C_1 \sum_{i=1}^{n} k\left(\left\|\frac{x - x_i}{h}\right\|^2\right)\delta(i, u) \quad u = 1, 2, \cdots, m \tag{31}$$

In the Eq. (31), h is the window width of kernel function. $\delta(i, u)$ is the Kronecker function which determines whether x_i belongs to the u interval, and returns 1 or 0. k is the profile function of Epanechikov kernel function $K_E(z)$. C_1 is a normalized coefficient expressed as:

$$C_1 = \frac{1}{\sum\limits_{i=1}^{n} k\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} \tag{32}$$

There is the formula as $K_E(z) = k(\|z\|^2)$. Suppose the volume of the d-dimensional unit ball is c_d and it can be described as:

$$K_E(z) = \begin{cases} \frac{1}{2} c_d^{-1}(d+2)(1 - \|z\|^2) & \|z\| < 1 \\ 0 & \text{others} \end{cases} \tag{33}$$

The pixels in target candidate region is assumed as $x_i(i = 1, 2, \cdots, n)$ and the probability density estimation of candidate is expressed as $\boldsymbol{p} = [p_1 \quad p_2 \quad \cdots \quad p_m]^T$. In each interval it can be given by

$$p_u(y) = C_2 \sum_{i=1}^{n} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right)\delta(i, u) \quad u = 1, 2, \cdots, m \tag{34}$$

where y is the position of the center of the candidate. C_2 is a normalized coefficient expressed as:

$$C_2 = \frac{1}{\sum_{i=1}^{n} k\left(\left\|\frac{y-x_i}{h}\right\|^2\right)} \tag{35}$$

The Bhattacharyya coefficient is used to measure the similarity between the target and the candidate model, as follow:

$$\rho[\boldsymbol{p}(y), \boldsymbol{q}\,] = \sum_{u=1}^{m} \sqrt{p_u(y)q_u} \tag{36}$$

The larger the value of $\rho[\boldsymbol{p}(y), \boldsymbol{q}\,]$, the higher the similarity. Let y_s be the target initial position, and perform a first-order Taylor expansion on $\rho[\boldsymbol{p}(y), \boldsymbol{q}\,]$ at y_s and sort it to get the Eq. (37).

$$\rho[\boldsymbol{p}(y), \boldsymbol{q}\,] = \frac{1}{2}\sum_{u=1}^{m}\sqrt{p_u(y_s)q_u} + \frac{1}{2}\sum_{i=1}^{n}\omega_i k\left(\left\|\frac{y-x_i}{h}\right\|^2\right) \tag{37}$$

in which

$$\omega_i = \sum_{u=1}^{m} \delta(i, u)\sqrt{\frac{q_u}{p_u(y_s)}} \tag{38}$$

In Eq. (37), the first term on the right side of the equation does not contain y. The second term represents a kernel probability density estimation with a profile function of k and a weight of ω_i. Let $T(y)$ represent the second term. Calculate its gradient with respect to y and sort it to get the Eq. (39).

$$\nabla T(y) = -\frac{2}{h^2}\left[\frac{1}{2}\sum_{i=1}^{n} k'\left(\left\|\frac{y-x_i}{h}\right\|^2\right)\right]\left[\frac{\sum_{i=1}^{n} x_i\omega_i k'\left(\left\|\frac{y-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n}\omega_i k'\left(\left\|\frac{y-x_i}{h}\right\|^2\right)} - y\right] \tag{39}$$

where k' represents the derivative of k. It can be noted that the term in the first bracket can be regarded as the kernel probability density estimation with the profile function of k, and the term in the second bracket represents the vector of the mean shift. The maximization of $T(y)$ can be completed by the following mean-shift iterative process.

$$y_{s+1} = \frac{\sum_{i=1}^{n} x_i\omega_i k'\left(\left\|\frac{y_s-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n}\omega_i k'\left(\left\|\frac{y_s-x_i}{h}\right\|^2\right)} \tag{40}$$

When the condition $y_{s+1} - y_s \leq \varepsilon$ or the number of iterations $\lambda \geq N$ are satisfied, the iteration is considered to be the end and the target location is updated. Take $\varepsilon = 0.5$ pixels and $N = 20$ in the algorithm.

III. SYSTEM ARCHITECTURE

This section describes the system architecture depicted in FIGURE 3. It is worth mentioning that the rotation range of the camera is 0 to 360 degrees in the two-dimensional plane. The array shape should not only match its detection range, but also adopt the minimum system volume. The resolution of the circular array is the same in all directions, so it perfectly matches the rotation range of the camera. As shown in FIGURE 4, the overall system is integrated on a circular substrate with a diameter of 10 cm. It is connected to the host computer through Ethernet or USB port. The platform is based on FPGA data acquisition and ARM system control. The auditory module is mainly divided into three parts: acoustic sensor, FPGA data acquisition and ARM control module.

1) The acoustic sensing module in the system is a centrally symmetrical circular array structure consisting of 7 acoustic digital MEMS sensors ADMP441.
2) FPGA synchronously acquires digital audio signals from 7 MEMS acoustic sensors via I^2S bus.
3) ARM aggregates and processes the data received by the FPGA through the FMC bus.

The ADMP441 (ADI, Massachusetts, America) microphone consists of MEMS sensor, ADC circuit, power management and industry standard 24-bit I^2S interface. It has a flat response curve from 60 Hz to 15 kHz and can be directly connected to the FPGA using the I^2S interface. Therefore, no audio encoder and decoder are required in the system.

The main function of the FPGA is to provide a synchronous multi-channel I^2S interface design. The low-power and low-cost chip EP2C5T144C8 (Altera, California, America) in the Cyclone II family is selected as the FPGA controller. It needs to complete the timing simulation of four I^2S buses, a buffer area and a data transmission interface for the task of collecting data in parallel. The 48 MHz active crystal oscillator is used as the clock input of FPGA. The serial configurator EPCS16 with 16 Mbit storage capacity is used in program storage and chip configuration, and the USB blaster is used as the emulator for debugging and downloading programs. The two AMS117 linear voltage stabilizer chips provide the core voltage of 1.2 v and 1.8 v respectively for the FPGA. To keep the microphone signals acquired synchronously, the same set of connecting lines SCK and WS are used between the FPGA and microphones.

The main function of ARM chip in this system is to process the data collected by FPGA through FMC bus. Communication with FPGA requires a high-speed communication bus. Meanwhile, the underlying hardware and drivers supporting USB or ethernet are needed to transmit data with the upper computer. The ARM microprocessor Cortex-M7 series is used in real-time control, in which the STM32F746ZGT6 (STMicroelectronics, Geneva, Switzerland) has a 216 MHz main frequency, 320 KB RAM and 1 MB FLASH storage space. Most importantly it supports FMC function, USB interface and ethernet communication. In this design, STM32 mainly writes the data in SRAM into the buffer through the FMC bus, and then transmits packaged data to the host computer through the USB bus to realize the function of the USB microphone array. The main work is the implementation of the STM32 and FPGA communication interface, as well as the adaptation of the USB audio class specification.

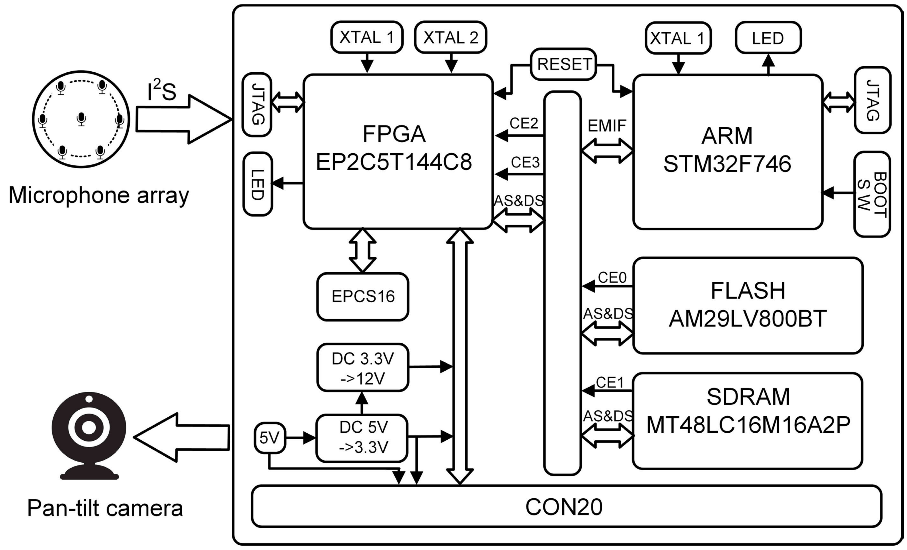


FIGURE 3. System architecture.

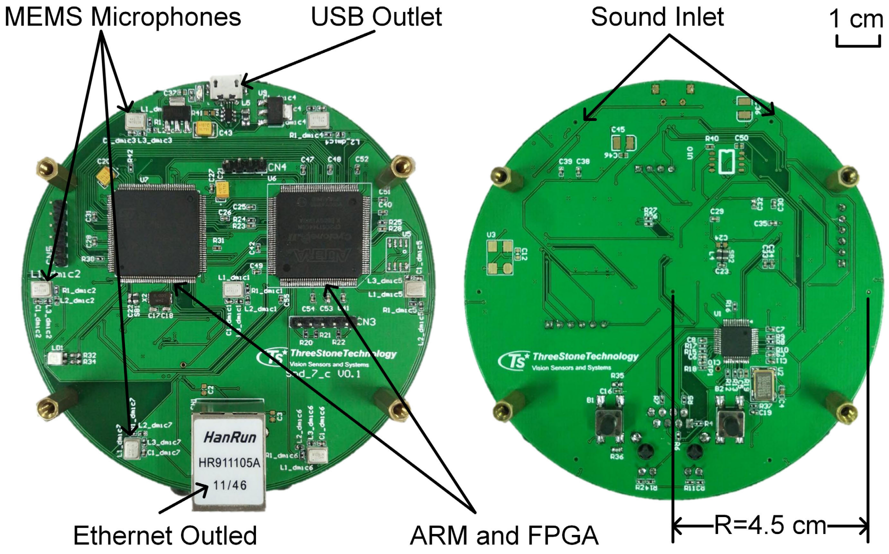


FIGURE 4. Photograph of the system.

STM32 and FPGA initiate data transmission through a 16-bit data bus, and data transmission is performed by means of time division multiplexing of data lines and address lines.

IV. RESULT AND DISCUSSION

In this section, the experiment setup as shown in FIGURE 5 (a) consists of the MEMS microphone array mentioned in Section III, a PTZ camera (DS-2DC7120IW-A, Hikvision, Hangzhou, China), and a computer. The system was tested in a normal office room with the size of 12 m × 7.5 m × 4 m. FIGURE 5 (b) illustrates the experimental scenario. The door opening sound activated the auditory module of the system, and then the visual module was driven to detect and track the target person according to the azimuth information provided by the auditory module. Auditory module was separately tested because of its significance to the overall system.

A. AUDITORY LOCALIZATION

In the auditory module test, seven channels of sound signals are received by the MA. The phase difference of these seven signals is used to calculate the source direction. A snapshot of these seven signals acquired by the MA in time domain is shown in FIGURE 6. FIGURE 7 demonstrates a snapshot of the sound signal and the VAD result based on short-term energy and ZCR. The part between the two red lines is judged to be a valid voice.

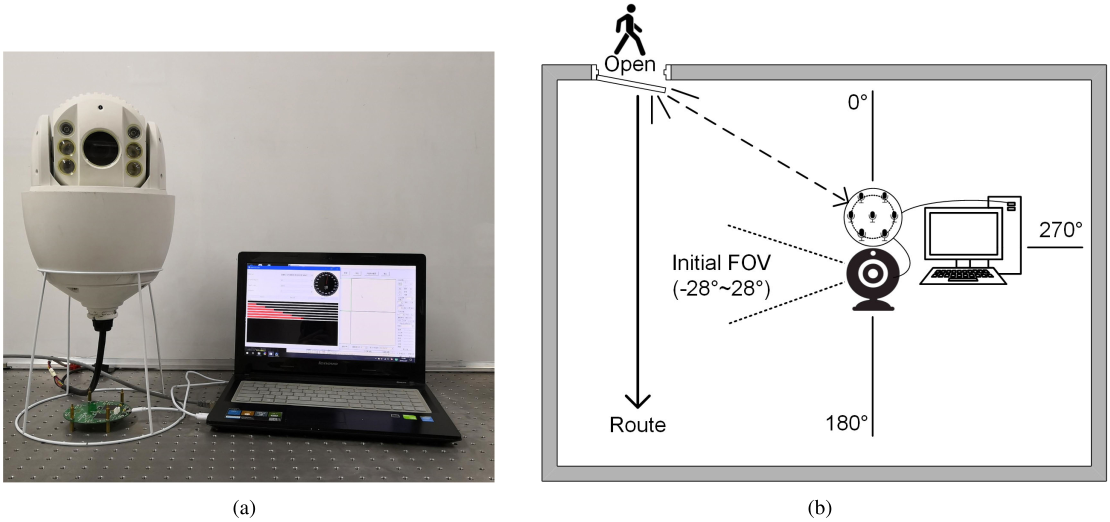


FIGURE 5. Experimental setup. (a) A picture of the experiment setup. (b) The experimental situation.

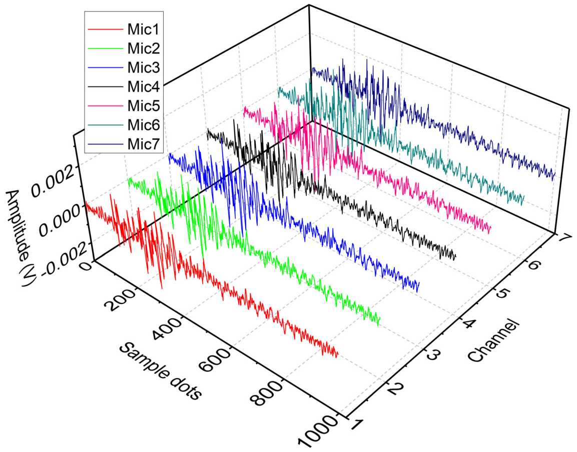


FIGURE 6. A snapshot of signals received by multiple microphones.

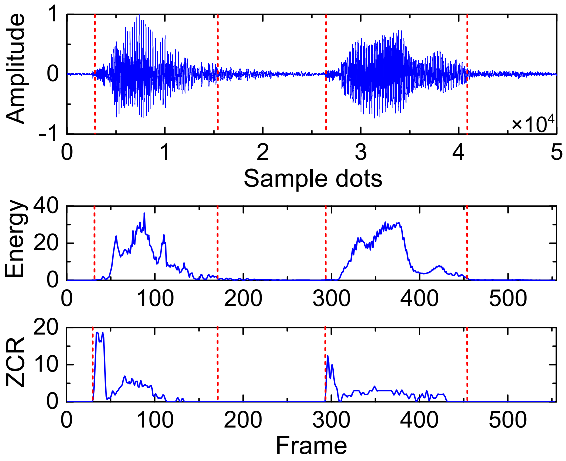


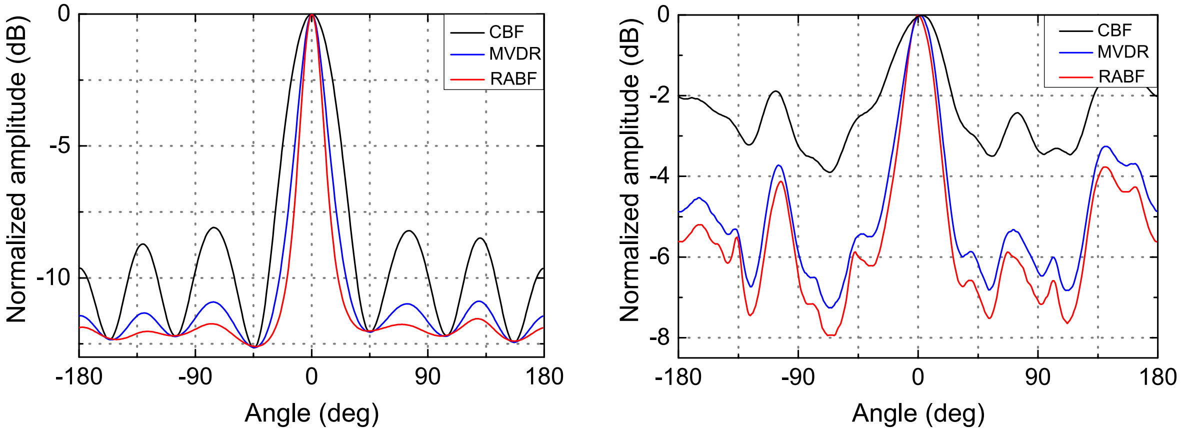
FIGURE 7. A snapshot of VAD results.

TABLE 1. Localization results of different angle.

Actual direction(°)	Average(°)	Variance(°)
0	2.14	4.03
90	91.94	1.47
180	180.60	2.33
270	271.23	2.94

The power-azimuth spectrum estimates of CBF, MVDR and RABF are plotted on a graph with a theoretical R and a sample R respectively, as is shown in FIGURE 8. The circular MA is used in this design with omnidirectional microphones, that is why the SOI power will not be different depending on the azimuth of the source. Therefore, the SOI azimuth in the simulation is set to 0°, while the azimuth angle of the experimental sound source relative to the MA is assumed to be 0°. Same as the MA, the number of array elements M=7, radius r=0.045 m, and array spacing d=0.045 m are set. As is well known to us, in the time domain signal processing, the under-sampling of the signal will lead to the generation of the gate lobes. The appearance of the grating lobes will lead to peak response blur. This is Nyquist's sampling law. Similarly, in order to satisfy the spatial sampling law to eliminate lobe aliasing, it must be satisfied that $d/\lambda \leq 1/2$, that is, the signal frequency is lower than 3777 Hz. The frequency of the simulated signal is set to 3600 Hz. Therefore, the simulated signal source is composed of 3600 Hz sine wave and Gaussian white noise. Under the comparison of the SOI power estimates of these three algorithms, it is obvious that RABF obtains higher SOI power than the other two algorithms.

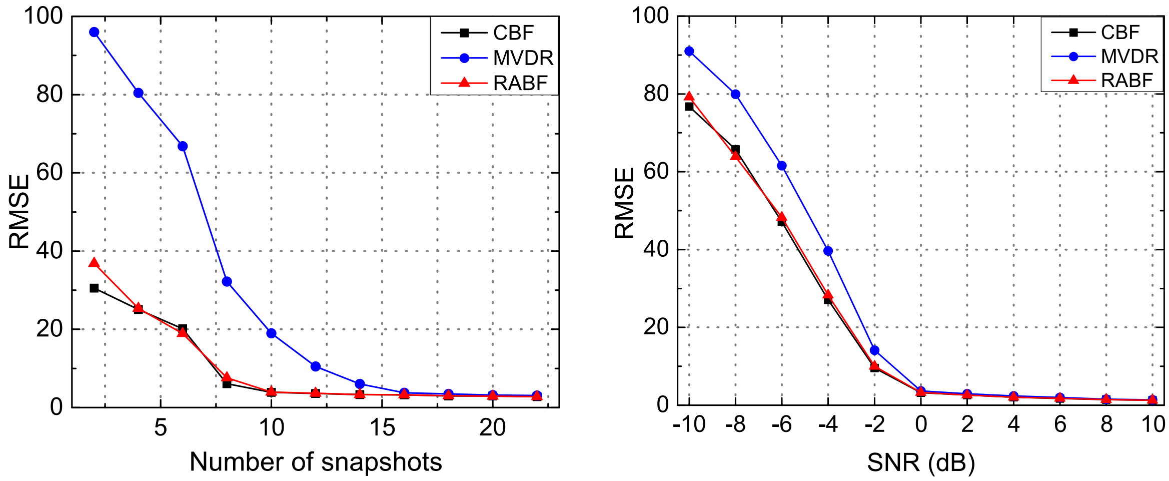
Further more, in order to study the robustness of the algorithms, we also calculate the root-mean-square error (RMSE) performance of these three algorithms from 500 Monte-Carlo runs, respectively. In the simulation, for the sake of generality,



(a)

(b)

FIGURE 8. Results of auditory localization test. (a) Power estimation with a theoretical R. (b) Power estimation with a sample R.



(a)

(b)

FIGURE 9. RMSE performance of each DOA estimation method. (a) RMSE versus SNR with the number of snapshots N=16. (b) RMSE versus number of snapshots with SNR=0 dB.

the SOI azimuth is given randomly from 0° to 359°. The SNR is set to be 0 dB while changing the number of snapshots, and the number of snapshots is set to be 16 while changing the SNR. FIGURE 9 (a) illustrates that RABF and CBF are superior to MVDR when the number of snapshots is less than 16 and RABF and CBF have similar performance. FIGURE 9 (b) shows that CBF and RABF are slightly better than MVDR when the SNR is lower than 0 dB and also RABF and CBF have similar performance. There is no doubt that CBF is of superior robustness. FIGURE 9 demonstrates that RABF has almost the same performance as CBF in both less snapshots number and low SNR. The above experiments demonstrates that RABF possesses superior robustness while having higher SOI power. The RMSE is less than 3° in the case of 0 dB SNR with enough sampling points.

Finally, the localization accuracy of the sound module was tested in the conference room environment. The microphone array was placed in the center of the room, and the sound sources was placed at 0°, 90°, 180°, and 270°, respectively, at a distance of 3 m from the microphone array. The algorithm is tested 30 times at each direction, and the estimation results are described in TABLE 1. The experiment results prove that the error of the averaged angle is less than 3° and it is helpful for visual module to detect and track the target.

B. AUDITORY-VISUAL TRACKING TEST

The experiment of auditory-visual joint tracking was performed in accordance with the scenario of FIGURE 5 (b). The system was modeled as human perception system. At the beginning, the camera was in a normal state and the target was out of sight. When the system received the valid audio input, that was door opening sound and footsteps, the auditory localization algorithm was used to detect the target angle and the PTZ camera rotated to make the target emerge in the FOV of camera. Then the visual module could detect and track the target. The pan-tilt rotation was controlled according to

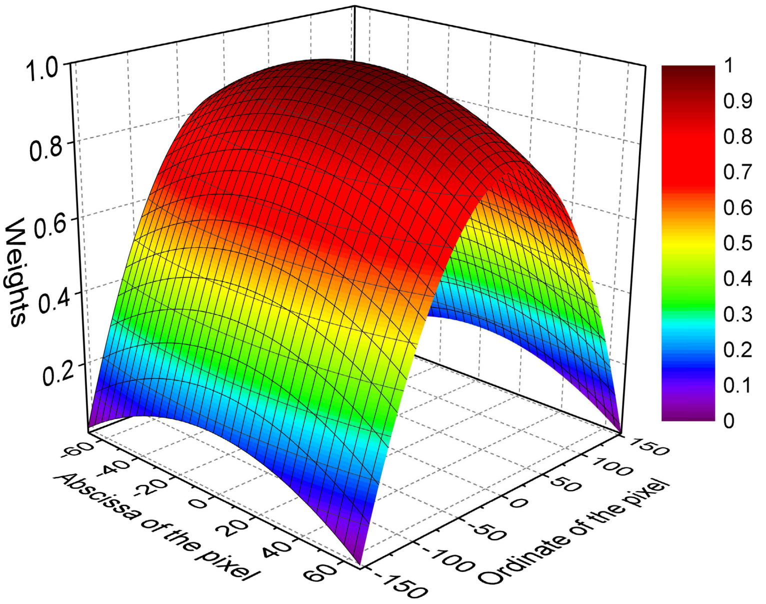
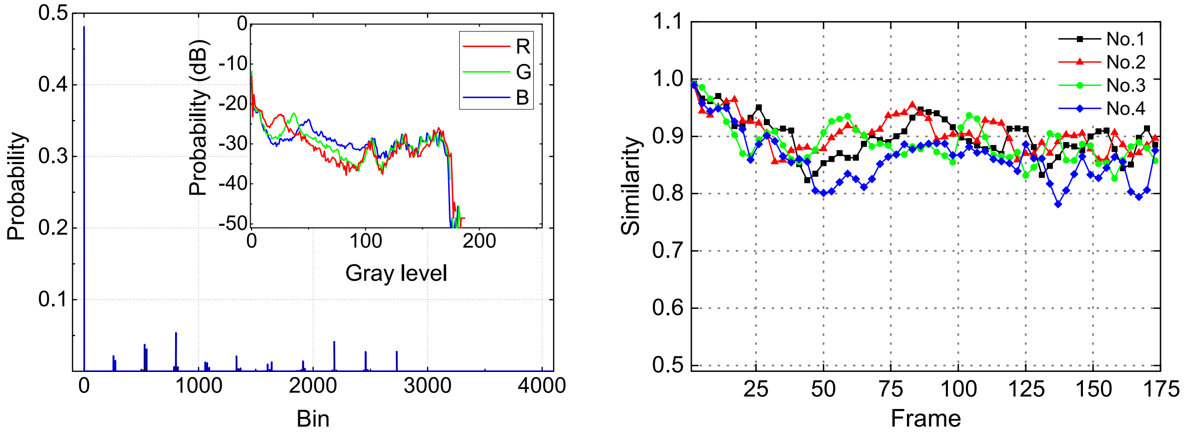


FIGURE 10. Weight distribution of the target area.



(a)

(b)

FIGURE 11. Results of visual tracking. (a) Color histogram of the target in FIGURE 10. (b) The similarity between the target and candidate.

TABLE 2. The time cost of the algorithms.

Algorithm	N	Time cost (ms)
VAD×7	1024	0.031
FFT×7	1024	0.077
Beamforming	1024	9.205
Mean-Shift	310 × 360 (pixels)	25.583

the target position in each frame image. The audio module is a supplement to its working state for necessary situation. The detection rate of the vision module is 25 FPS. In the experiment, the target moved at a speed of about 1 m/s, and the distance between the target's straight path and the system was about 2.5 m. In this design, the sound source module is always in working state to provide azimuth information for the system. The priority of visual tracking is higher than that of auditory, because the information obtained by visual module is larger than that of auditory module. Until the video

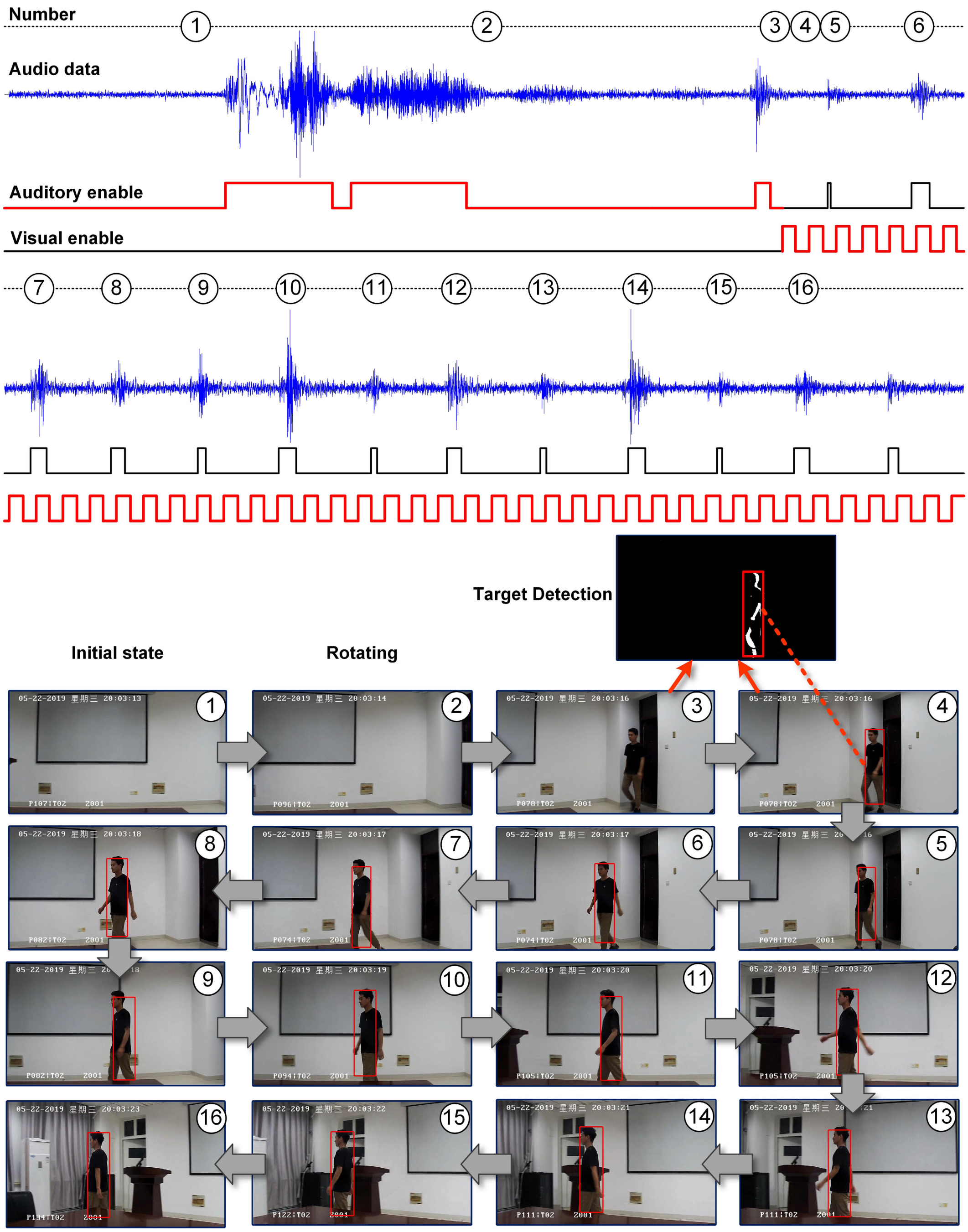


FIGURE 12. The results of real-time auditory-visual tracking.

tracking loses its target, it will be tracked by the audio module again. If neither has an enabling signal, the system works like a normal camera.

For the visual tracking, referring to Eq. (33), the probability weight distribution of each pixel of the target region is shown in FIGURE 10. In the target area of 310 × 136 pixels, the weight of the center portion is greater than edge portion. Referring to the Eq. (31), FIGURE 11 (a) expresses the color histogram of the object model, and the RGB color space is divided into 16 × 16 × 16 bins. The sub-figure shows the gray histogram of each color. Referring to Eq. (36), FIGURE 11 (b) is the curve showing the similarity between the target and the candidates of four sets of experiments. The target tracking in the experiment indicates the accuracy of the visual module, as shown in FIGURE 12. The execution time and frame length of the implemented algorithms is shown in TABLE 2. The VAD, FFT and beamforming all run on 7 channels. N is the length of audio data per frame or the number of pixels of image data per frame. The computational cost represents the number of times the system performs real multiplication.

Obviously, the beamforming and mean-shift algorithms takes up a lot of calculations and the time cost of time is 9.205 ms and 25.583 ms, respectively. The overall time cost meets the real-time requirements.

The process of the overall test is shown in FIGURE 12. The time process was represented in numerical order. From frame 1 to 3, the video module was initially in sleep or normal state and it was triggered by the door opening sound signal. The pan-tilt rotation was controlled by auditory azimuth signal, based on the data of the auditory localization, until the target in the video sequence was detected as frame 4. When the target was detected, visual enable signal took over the control of pan-tilt rotation according to the data of mean-shift tracking, as frame 5 to 16. When the target was lost by visual module, control was taken over by the auditory module again. If neither module provide an enable signal, the tracking ended and the visual module re-entered the sleep or normal state until it was triggered by an audio signal. The experiment was repeated 20 times and the targets were accurately detected and tracked in all groups.

V. CONCLUSION

In this article, for the purpose of solving problems that conventional cameras fail to monitor objects in their blind zones, we proposed an acoustic enhanced camera tracking system for real-time monitoring. It is developed based on a small aperture MEMS microphone array with a 4.5 cm radius, that makes the system miniaturized and integrated. The demo is inspired by the joint action of human beings from hearing to vision. The broadband beamforming and mean-shift algorithm were mainly implemented in sound and image data processing. The acoustic localization results prove that the error of the averaged angle is less than 3°, demonstrating that it is helpful to locate the target outside FOV of camera for the visual tracking module. In the joint tracking experiments, targets were accurately detected and tracked in all groups. The system extends the detection angle of camera tracking system to all directions and performs well in the real-time and robustness. In practical applications, worse scenarios is often faced with, such as lower SNR, occlusion environment and multi-source situation. The applied algorithms need to be optimized to improve the system's performance in worse environment. An improved tracking strategy is also indispensable to meet the demands of more complex situations.

REFERENCES

[1] W. S. M. Sanjaya, D. Anggraeni, K. Zakaria, A. Juwardi, and M. Munawwaroh, "The design of face recognition and tracking for human-robot interaction," in *Proc. 2nd Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Yogyakarta, Indonesia, Nov. 2017, pp. 315–320.

[2] X. Yang, P. Chen, S. Gao, and Q. Niu, "CSI-based low-duty-cycle wireless multimedia sensor network for security monitoring," *Electron. Lett.*, vol. 54, no. 5, pp. 323–324, Mar. 2018.

[3] H. Liu, Z. Wei, Y. Chen, J. Pan, L. Lin, and Y. Ren, "Drone detection based on an audio-assisted camera array," in *Proc. IEEE 3rd Int. Conf. Multimedia Big Data (BigMM)*, Laguna Hills, CA, USA, Apr. 2017, pp. 402–406.

[4] M. Yazdi and T. Bouwmans, "New trends on moving object detection in video images captured by a moving camera: A survey," *Comput. Sci. Rev.*, vol. 28, pp. 157–177, May 2018.

[5] Z. Pan, S. Liu, and W. Fu, "A review of visual moving target tracking," *Multimedia Tools Appl.*, vol. 76, no. 16, pp. 16989–17018, Aug. 2017.

[6] R. Kalsotra and S. Arora, "A comprehensive survey of video datasets for background subtraction," *IEEE Access*, vol. 7, pp. 59143–59171, 2019.

[7] W. Yu, Z. Hou, D. Hu, and P. Wang, "Robust mean shift tracking based on refined appearance model and online update," *Multimedia Tools Appl.*, vol. 76, no. 8, pp. 10973–10990, Apr. 2017.

[8] Y. Liu, X.-Y. Jing, J. Nie, H. Gao, J. Liu, and G.-P. Jiang, "Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos," *IEEE Trans. Multimedia*, vol. 21, no. 3, pp. 664–677, Mar. 2019.

[9] H. Wang, X. Wang, L. Yu, and F. Zhong, "Design of mean shift tracking algorithm based on target position prediction," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Tianjin, China, Aug. 2019, pp. 1114–1119.

[10] *Instructors' Handbook, A. FAA-H-8083-9A*, Dept. Transp., Federal Aviation Admin., Flights Standards Service, Washington, DC, USA, 2008, pp. 2–24.

[11] F. R. D. Amaral, J. C. Serrano Rico, and M. A. F. D. Medeiros, "Design of microphone phased arrays for acoustic beamforming," *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 40, no. 7, p. 354, Jul. 2018.

[12] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization – reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 120, pp. 422–448, Apr. 2019.

[13] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H. Okuno, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, Nov. 2017.

[14] B. Liao, A. Madanayake, and P. Agathoklis, "Array signal processing and systems," *Multidimensional Syst. Signal Process.*, vol. 29, no. 2, pp. 467–473, 2018.

[15] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, vol. 2, Oct. 2003, pp. 1228–1233.

[16] Z.-H. Michalopoulou and R. Jain, "Particle filtering for arrival time tracking in space and source localization," *J. Acoust. Soc. Amer.*, vol. 132, no. 5, pp. 3041–3052, Nov. 2012.

[17] X. Zhang, J. Huang, E. Song, H. Liu, B. Li, and X. Yuan, "Design of small MEMS microphone array systems for direction finding of outdoors moving vehicles," *Sensors*, vol. 14, no. 3, pp. 4384–4398, Mar. 2014.

[18] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[19] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[20] J. L. Odom, J. L. Krolik, and J. S. Rogers, "Maximum-likelihood spatial spectrum estimation in dynamic environments with a short maneuverable array," *J. Acoust. Soc. Amer.*, vol. 133, no. 1, pp. 311–322, Jan. 2013.

[21] A. Manikas, Y. I. Kamil, and M. Willerton, "Source localization using sparse large aperture arrays," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6617–6629, Dec. 2012.

[22] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory.* Hoboken, NJ, USA: Wiley, 2004.

[23] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. Boston, MA, USA: McGraw-Hill, 2000, p. 656.

[24] R. J. Mailloux, *Phased Array Antenna Handbook*. Norwood, MA, USA: Artech House, 2017.

[25] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[26] P. Stoica, Z. Wang, and J. Li, "Robust capon beamforming," *IEEE Signal Process. Lett.*, vol. 10, no. 6, pp. 172–175, Jun. 2003.

[27] L. Tzafri and A. J. Weiss, "High-resolution direct position determination using MVDR," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6449–6461, Sep. 2016.

[28] M. Wax and Y. Anu, "Performance analysis of the minimum variance beamformer in the presence of steering vector errors," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 938–947, Apr. 1996.

[29] A. Pezeshki, B. D. Van Veen, L. L. Scharf, H. Cox, and M. Lundberg Nordenvaad, "Eigenvalue beamforming using a multirank MVDR beamformer and subspace selection," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1954–1967, May 2008.

[30] W. Zhang, J. Wang, and S. Wu, "Robust capon beamforming against large DOA mismatch," *Signal Process.*, vol. 93, no. 4, pp. 804–810, Apr. 2013.

[31] X. Mestre and M. A. Lagunas, "Finite sample size effect on minimum variance beamformers: Optimum diagonal loading factor for large arrays," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 69–82, Jan. 2006.

[32] C.-Y. Tseng, D. D. Feldman, and L. J. Griffiths, "Steering vector estimation in uncalibrated arrays," *IEEE Trans. Signal Process.*, vol. 43, no. 6, pp. 1397–1412, Jun. 1995.

[33] J.-H. Lee and C.-C. Wang, "Adaptive array beamforming with robust capabilities under random sensor position errors," *IEE Proc.-Radar, Sonar Navigat.*, vol. 152, no. 6, pp. 383–390, Dec. 2005.

[34] J. Li and P. Stoica, *Robust Adaptive Beamforming*. Hoboken, NJ, USA: Wiley, 2006.

[35] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.

[36] Y. Li, H. Ma, D. Yu, and L. Cheng, "Iterative robust capon beamforming," *Signal Process.*, vol. 118, pp. 211–220, Jan. 2016.

[37] J. Zhuang, B. Shi, X. Zuo, and A. H. Ali, "Robust adaptive beamforming with minimum sensitivity to correlated random errors," *Signal Process.*, vol. 131, pp. 92–98, Feb. 2017.

[38] Y. Ke, C. Zheng, R. Peng, and X. Li, "Robust adaptive beamforming using noise reduction preprocessing-based fully automatic diagonal loading and steering vector estimation," *IEEE Access*, vol. 5, pp. 12974–12987, 2017.

[39] L. Huang, J. Zhang, X. Xu, and Z. Ye, "Robust adaptive beamforming with a novel Interference-Plus-Noise covariance matrix reconstruction method," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1643–1650, Apr. 2015.

[40] F. Chen, J. Song, and F. Shen, "Robust adaptive beamforming using low-complexity correlation coefficient calculation algorithms," *Electron. Lett.*, vol. 51, no. 6, pp. 443–445, Mar. 2015.

[41] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust capon beamformer," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2407–2423, Sep. 2004.

[42] Z. Wang, Q. Zou, Q. Song, and J. Tao, "The era of silicon MEMS microphone and look beyond," in *Proc. Transducers 18th Int. Conf. Solid-State Sensors, Actuat. Microsyst. (TRANSDUCERS)*, Anchorage, AK, USA, Jun. 2015, pp. 375–378.

[43] B. da Silva, A. Braeken, K. Steenhaut, and A. Touhafi, "Design considerations when accelerating an FPGA-based digital microphone array for sound-source localization," *J. Sensors*, vol. 2017, pp. 1–20, Jan. 2017.

[44] B. da Silva, A. Braeken, and A. Touhafi, "FPGA-based architectures for acoustic beamforming with microphone arrays: Trends, challenges and research opportunities," *Computers*, vol. 7, no. 3, p. 41, Aug. 2018.

[45] Y. Song, "Design, analysis and characterization of silicon microphones," State Univ. New York Binghamton, Binghamton, NY, USA, Tech. Rep. 3310727, 2008.

[46] J. Lewis and B. Moss, "MEMS microphone: The future for hearing aids," *Analog Dialogue*, vol. 47, no. 11, pp. 3–5, 2013.

[47] S.-S. Yun, Q. Nguyen, and J. Choi, "Recognition of emergency situations using audio–visual perception sensor network for ambient assistive living," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 1, pp. 41–55, Jan. 2019.

[48] E. D'Arca, N. M. Robertson, and J. R. Hopgood, "Robust indoor speaker recognition in a network of audio and video sensors," *Signal Process.*, vol. 129, pp. 137–149, Dec. 2016.

[49] R. Viciana-Abad, R. Marfil, J. Perez-Lorenzo, J. Bandera, A. Romero-Garces, and P. Reche-Lopez, "Audio-visual perception system for a humanoid robotic head," *Sensors*, vol. 14, no. 6, pp. 9522–9545, May 2014.

[50] K. Wilson, V. Rangarajan, N. Checka, and T. Darrell, "Audiovisual arrays for untethered spoken interfaces," in *Proc. 4th IEEE Int. Conf. Multimodal Interfaces*, Oct. 2002, pp. 389–394.

[51] X. Qian, A. Xompero, A. Cavallaro, A. Brutti, O. Lanz, and M. Omologo, "3D mouth tracking from a compact microphone array co-located with a camera," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 3071–3075.

[52] H. Ha, S. Han, and J. Lee, "Fault detection on transmission lines using a microphone array and an infrared thermal imaging camera," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 1, pp. 267–275, Jan. 2012.

[53] M. Goseki, M. Ding, H. Takemura, and H. Mizoguchi, "Combination of microphone array processing and camera image processing for visualizing sound pressure distribution," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2011, pp. 139–143.

LEI LI received the B.S. degree in electronic engineering from Zhengzhou University, Zhengzhou, China, in 2004, and the Ph.D. degree in electronic engineering from the Institute of Acoustics, Chinese Academy of Sciences, in 2009. He is currently an Associate Professor with the School of Physics and Engineering, Zhengzhou University. His current research interests include array signal processing, machine learning, and computer vision.

KECHAO LIAN received the B.S. degree in electrical engineering from Zhengzhou University, Zhengzhou, China, in 2018, where he is currently pursuing the master's degree with instrumentation engineering. His current research interests include array signal processing and computer vision.

JINTAO FU is currently pursuing the bachelor's degree in electrical engineering with Zhengzhou University, Zhengzhou, China. His current research interests include digital signal processing, digital image processing, and sensor principle and application.

PENGFEI ZHU received the B.S. and M.S. degrees in electrical engineering from Zhengzhou University, Zhengzhou, China, in 2014 and 2018, respectively. He is currently an Algorithmic Engineer with the 22nd Research Institute of China Electronics Technology Group Corporation. His research interests cover signal processing, satellite navigation, and remote sensing.

ZHIYONG HU received the B.S. degree in electrical engineering from the Zhengzhou University, Zhengzhou, China, in 2018, where he is currently pursuing the master's degree with instrumentation engineering. His current research interests include ultrasonic gas flow meter, signal processing, and machine learning.

CE GUO received the B.S. degree in electrical engineering from the Zhengzhou University of Light Industry, Zhengzhou, China, in 2016. He is currently pursuing the master's degree with instrumentation engineering with Zhengzhou University, Zhengzhou, China. His current research interests include array signal processing, machine learning, and computational vision.