Research on Sound Source Localization and Real-Time Facial Expression Recognition for Security Robot

Zhonghua Han and Tingting Li

First Research Institute of the Ministry of Public Security of PRC, Beijing Zhong-Tianfeng Security Protection Technologies Co. Ltd, Beijing, China
Email: 15120072767@139.com; ltting101416@163.com

Abstract. Security robots are intelligent equipment used to perform security tasks, which can play a role in reducing the labor intensity and duty risk of police. In view of the shortcomings of existing police security robots in human-computer interaction and cooperation, the sound source localization and real-time facial expression recognition for the security robot is introduced in this paper. The auditory system of the robot obtains external sound analog signals through a microphone array, and performs corresponding digital signal processing on the collected sound signals. In order to strengthen the information interaction capability between the security robot and operators, the Kinect with deep data processing capabilities is introduced in this paper. The experimental results show that the system can analyze and recognize the position and status of the operators.

Keywords. Security robot; sound source localization; facial expression recognition.

1. Introduction
Security work has always been an important service work. With the development of society, for example in the complex environments such as supermarkets, airports and hospitals, the traditional human security, due to its own limitations such as physiological conditions, individuals service quality and limited threat detection capabilities [1], is difficult to meet the requirements of modern security work. The security robots with the potential to overcome these difficulties emerge at this moment.

After decades of research, a variety of security robots for the surveillance, patrol and service have been developed. The security patrol robot MDARS-I and MDARS-E developed by Cybermotion [2] has the microwave radar and thermal imaging cameras that can monitor smoke, fire and intruders. The K5 Knightscope robot [3] with the thermal imaging system, air detectors and video camera are used to guard in the public areas such as schools and parks. For the iBotGuard developed by Hong Kong Polytechnic University [4], the robot teleoperation is introduced to the iBotGuard system. The ActivMedia Peoplebot security robot improved by Treptow et al. [5] can realize real time monitoring and tracking people. The NUCC Security Warrior developed by National Chung Cheng University [6] utilize visual system to carry out human detection and tracking. The security patrol robot developed by Tahzib et al. [7] can be easily controlled by mobile application. The security reconnaissance and surveillance robot developed by Silvia et al. [8] can surveil through aerial platforms.

Limited by the development status of artificial intelligence, security robots’ ability to perform tasks independently is still insufficient. For example, the K5 Knightscope robot, which is a fully autonomous security robot, has had several accidents such as evicting homeless people, falling into pools and being attacked by people [3]. Therefore, cooperating with people should still be the main working mode of security robots. Aiming to improve the human-machine interaction and cooperation, the sound source...
localization and real-time facial expression recognition for the security robot is introduced into this paper.

This paper is organized as follows: Section 2 introduces the sound source localization method. Section 3 introduces the facial expression recognition based on Kinect. Experiment and results are presented in section 4 and a conclusion is drawn in section 5.

2. Sound Source Localization

The auditory system of the robot obtains external sound analog signals through a microphone array, and performs corresponding digital signal processing on the collected sound signals. The system requirements are as follows:

1. Omnidirectional microphone with high sensitivity and good frequency response characteristics.
2. Multi-channel simultaneous sampling.
3. The sampling frequency of single channel must be at least 44.1 KHz.
4. High real-time implementation.

According to the above requirements, we choose the Re speaker 6-Mic microphone array based on Raspberry Pi to complete the research of sound source localization. The schematic diagram of microphone array is shown in figure 1. $S$ is sound source, $S'$ is projection point of sound source on array plane. $O$ is reference point, the radius of the circle where Mic 1 located is the reference line. $d_i$ is the distance between $S$ and $O$, $d_{mi}$ ($i=1,2,3,\ldots,6$) is the distance between $S$ and Mic $i$. $\varphi_{mi}$ is the angle between the microphone and the reference line.

![Figure 1. Schematic diagram of microphone array.](image)

The main components of the signals received by each element in the microphone array come from the same sound source, and have strong correlation. Considering the real-time requirements of the system, time delay estimation algorithm based on Generalized Gross Correlation (GCC) proposed by Knapp and Carter [9] is adopted. We choose two of the six microphones on the diagonal as a group, using cross correlation function of the two signals to calculate the time difference of sound source arrival. Suppose that the signals received by Mic 1 and Mic 4 are shown as follows:

$$x_1(n) = a_1 s(n - \tau_1) + k_1(n)$$

$$x_4(n) = a_4 s(n - \tau_4) + k_4(n)$$

where $s(n)$ is sound source signal. $a_i$ is attenuation factor of signal. $\tau_i$ is propagation time of sound wave from sound source to Mic $i$. $k_i(n)$ and $k_4(n)$ are uncorrelated noise signals.
According to the definition of cross correlation function, the correlation function \( R_{n_i} (\tau) \) between the two microphones is as follows:

\[
R_{n_i} (\tau) = E(x_i(n)x_i(n-\tau))
\]  

(3)

From the relationship between cross correlation function and cross power spectrum:

\[
R_{n_i} (\tau) = \int_{-\infty}^{+\infty} G_{n_i} (\omega)e^{j\omega\tau} d\omega
\]  

(4)

\[
G_{n_i} (\omega) = X_i(\omega) \ast X_i^*(\omega)
\]  

(5)

\( G_{n_i} (\omega) \) is the cross power spectrum function of \( x_i(n) \) and \( x_i(n) \), \( X_i(\omega) \) and \( X_i(\omega) \) are Fourier transforms of \( x_i(n) \) and \( x_i(n) \), \( X_i^*(\omega) \) is the conjugate of \( X_i(\omega) \).

In order to reduce the impact of noise in the actual environment, the cross power spectrum is weighted by the PHAT weighting function \( \psi_{n_i} (\omega) \) in the frequency domain. After IFFT, the GCC function between two signals is obtained:

\[
R_{n_i} (\tau) = \int_{-\infty}^{+\infty} \psi_{n_i} (\omega)G_{n_i} (\omega)e^{j\omega\tau} d\omega
\]  

(6)

\[
\psi_{n_i} (\omega) = \frac{1}{|G_{n_i} (\omega)|}
\]  

(7)

The peak value of \( R_{n_i} (\tau) \) is the time difference \( \tau_{n_i} \) between two signals. The same method is used to calculate the time difference between the other two groups of microphone signals. The sound source direction can be obtained by spatial geometry calculation.

3. Facial Expression Recognition

In order to strengthen the information interaction capability between the security robot and operators, the Kinect with deep data processing capabilities is introduced in this paper as an image acquiring device. The deep data is obtained by a pair of infrared cameras using structured light binary code. Kinect can capture the skeleton information based on the deep data for skeletal tracking. Figure 2 shows the physical structure of Kinect: RGB Module is color sensor for RGB imaging. IR Projector, Right Imager and Left Imager is used for depth sensing. The system chip with image encoding and logic calculation capability is the core component of Kinect. A three-axis accelerometer is used to maintain Kinect position stability.

3.1. Facial Feature Acquisition

Figure 3 shows the facial feature acquisition process. In this paper, Haar features are used for facial information extraction. Basic Haar features include edge features, linear features, center features and diagonal features, which are combined into feature templates.
In order to calculate Haar features quickly, the integral image method proposed by Viola and Jones [10] is introduced in this paper. The integral image at the pixel point \((x, y)\) is defined as follows:

\[
\mathcal{I}(x, y) = \sum_{i \leq x, j \leq y} \mathcal{I}(x', y')
\]

(8)

where \(\mathcal{I}(x, y)\) is integral image, \(\mathcal{I}(x', y')\) is original image.

\[
s(x, y) = s(x, y-1) + \mathcal{I}(x, y)
\]

(9)

\[
\mathcal{I}(x, y) = \mathcal{I}(x-1, y) + s(x, y)
\]

(10)

\(s(x, y)\) is the cumulative row sum, when \(s(x,-1)=0\), \(\mathcal{I}(-1, y)=0\), the integral image can be calculated by once scan on the original image.

**Figure 3.** Facial feature acquisition process.

### 3.2. Facial Expression Recognition Based on SVM

Support Vector Machine (SVM) is a generalized linear classifier performs binary classification of data in a supervised learning manner. With the development of SVM, the kernel function has become the key to solve the nonlinear problem of facial expression recognition. The kernel function is used to realize the nonlinear transformation of class space, data space and feature space as follows:

\[
(x_i, x_j) \rightarrow K(x_i, x_j) = \Phi(x_i) \bullet \Phi(x_j)
\]

(11)

where \(x_i\) and \(x_j\) are data space sample points, \(\Phi\) is the mapping function from data space to feature space, \(K(x_i, x_j)\) is the kernel function meet the Mercer condition. The Hibert space for \(K(x_i, x_j)\), \(\Phi^d \rightarrow H\) satisfies the following formula:

\[
K(x_i, x_j) = \sum_{n=1}^{d_x} \Phi(x_i) \bullet \Phi(x_j)
\]

(12)

where \(d_x\) is Hibert space dimension. Using the kernel function instead of \((x_i, x_j)\), the SVM model is obtained as follows:
\[ f(x) = \text{sgn}\left( \sum_{i=1}^{n} \alpha_i y_i K(x, x') + b \right) \] (13)

Combined with the facial features obtained in the previous section, facial expression recognition can be realized, and the experimental results are introduced in the next section.

4. Experiments and Results

The security robot experiment platform is shown in figure 4. Figure 4a shows the security robot of Ministry First Research Institute. Figure 4b shows the 6-Mic Circular Array Kit which consists of voice accessory HAT and six microphones circular array. It can pick up omnidirectional sound source signal. Figure 4c shows the depth camera module. It is a stereo tracking solution, offering quality depth for applications.

We tested functional tests on two modules, the test results are shown in figure 5. As can be seen from figure 5a, the sound source localization module can identify whether there are sound signals around and determine the direction of the sound source. As shown in figure 5b, the facial expression recognition module successfully identified several expressions such as surprise, neutral, sad, happy, etc.
Based on the interaction needs between security robots and humans, the direction and state of the interactive person are related to a certain extent, which is reflected in the following aspects:

When a person interacts with a robot, the spatial position of the person relative to the robot can be obtained by sound source localization. At this point, the robot has a basis to turn its front side to the person. It can not only enhance the sense of interactive experience, but also include the person in its visual collection range.

After the face is quickly locked, the facial expression feature points are quickly extracted, which can identify the facial expression classification, reduce the scope of the related semantic expression content, and further output variables such as the frequency of mouth shape change, provide a basis for the real identification of multiple objects in the same direction.

The relationship between sound and image is an effective way to improve the experience of human-robot interaction. Furthermore, it can refine and expand the logical relationship between the human, the event and the object in the two dimensions of time and space, and realize the humanoid application of robots.

5. Conclusion
In this paper the sound source localization and real-time facial expression recognition for the security robot is introduced in this paper to improve the human-machine interaction and cooperation. The microphone array and the Kinect with deep data processing capabilities is introduced in this paper and the feasibility of the method is verified by experiments.

Acknowledgment
The work was funded by the National Key R&D Program of China under Grant 2017YFC0806506.

Figure 5. (a) Sound source localization test results; (b) Facial expression recognition test results.
References

[1] Sean W G 2009 Cuwith: A Curiosity Driven Robot for Office Environmental Security (Auckland University of Technology) pp 1-97.

[2] Shoop B, Johnston M, Goehring R, Moneyhun J and Skibba B 2006 Mobile detection assessment and response systems (MDARS): A force protection physical security operational success The International Society for Optical Engineering 6230 (1) 1-11.

[3] Keijzers M, Kazmi H, Eyssel F and Bartneck C 2019 Teaching robots a lesson: Determinants of robot punishment International Journal of Social Robotics 9 1-14.

[4] Liu J N K, Wang M and Feng B 2005 Ibotguard: An internet-based intelligent robot security system using invariant face recognition against intruder IEEE Transactions on Systems, Man and Cybernetics 35 (1) 97-105.

[5] Treptow A, Cielniak G and Duckett T 2005 Active people recognition using thermal and grey images on a mobile robot IEEE/RSJ Int. Conf. on Intelligent Robots and Systems pp 2103-08.

[6] Ren C L, Yi T C, Chung T L, Chun C L and An C T 2008 NCCU security warrior: An intelligent security robot system IECON 2007-33rd Annual Conf. of the IEEE Industrial Electronics Society pp 2960-65.

[7] Tahzib M, Hasib Z and Maofic F K 2018 Design and implementation of security patrol robot using android application 11th Int. Conf. on Mathematical Modelling and Computer Simulation pp 77-82.

[8] Prophet S, Atman J and Trommer G F 2017 A synergetic approach to indoor navigation and mapping for aerial reconnaissance and surveillance Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN) pp 1-8.

[9] Knapp C and Carter G 2003 The generalized correlation method for estimation of time delay IEEE Transactions on Acoustics Speech and Signal Processing 24 (4) 320-27.

[10] Viola P and Jones M 2004 Robust real-time face detection International Journal of Computer Visio 57 (2) 137-154.