An interactive system for humanoid robot SHFR-III

Xianxin Ke, Bin Cao, Jiaojiao Bai, Wenzhen Zhang and Yujiao Zhu

Abstract
The natural interaction between human and robot is full of challenges but indispensable. In this article, a human–robot interactive system is designed for humanoid robot SHFR-III. The system consists of three subsystems: multi-sensor positioning subsystem, emotional interaction subsystem, and dialogue subsystem. The multi-sensor positioning subsystem is designed to improve the positioning accuracy, the emotional interaction subsystem uses bimodal emotional recognition model and fuzzy emotional decision-making model to realize the emotion recognition and expression feedback to the interactive objects, and the dialogue subsystem with personal information can complete the response consistent with the default information and avoid conflicting responses. The experimental results show that the multi-sensor positioning subsystem has good environmental adaptability and positioning accuracy, the emotional interaction subsystem can achieve human-like emotional feedback, and the dialogue subsystem can achieve more natural, logical, and consistent responses.

Keywords
Interactive system, positioning system, emotional interaction, dialogue system, humanoid robot

Introduction
Human–robot interaction (HRI) was first proposed in 1975. HRI is an important interdisciplinary research field of computer, ergonomics, cognitive science, and other disciplines. It is also an important content of engineering psychology research. At present, HRI is developing toward personification, intellectualization, and naturalization.

Many researchers are now dedicating their efforts to studying interactive modalities such as facial expressions, natural language, and gestures. This phenomenon makes communication between robots and individuals become more natural. Gunes et al. analyzed human participants’ nonverbal behavior and predicted their facial action units, facial expressions, and personality in real time while they interacted with a small humanoid robot. Ali et al. designed a sign language educational humanoid that possess stereo vision and stereo microphones along with stereo audios for intuitive interaction. De Jong Michiel et al. designed a humanoid robot for social interaction, they combined vision, gesture, speech, and input from an onboard tablet, a remote mobile phone, and external microphones. Kumra and Kanan presented a novel robotic grasp detection school of mechatronic engineering and automation, Shanghai University, Shanghai, China.

Corresponding authors:
Xianxin Ke, School of Mechatronic Engineering and Automation, Shanghai University, 99 Shangda Road, Shanghai 200444, China. Email: xinke@staff.shu.edu.cn
Bin Cao, School of Mechatronic Engineering and Automation, Shanghai University, 99 Shangda Road, Shanghai 200444, China. Email: 1573115830@qq.com

School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China.
system that predicts the best grasping pose of a parallel-plate robotic gripper for novel objects using the RGB Depth Map image of the scene. This article considers the sustainability of the product and conducts SHFR-III interactive research from the perspective of positioning, emotion, and dialogue.

Robots should have lightweight, low-energy consumption, and excellent performance, so we designed and built a humanoid emotional robot SHFR-III (see Figure 1). SHFR-III has 22 degrees of freedom and can realize 8 basic expressions, including calm, happiness, and so on.

Target positioning is an important part of the research of humanoid robots. In HRI, first, robots need to recognize the interactive objects from the environment. Laurenzi et al. introduced a set of modules based on visual object localization. Deepak Gala et al. used auditory sensors for positioning. However, the positioning effect of single sensor is greatly affected by external factors, such as auditory positioning is limited by noise and visual positioning is limited by illumination. Therefore, multi-sensor system can reduce this situation.

In real life, people usually do not judge each other’s emotional state based on single modal information. Visual information and voice signal information are very important for emotional judgment. Multi-modal emotion recognition is to recognize emotional states by using multi-modal information. Emotional computing was first proposed by Picard. Emotional computing can measure and analyze the external performance of human emotions and affect them.

As argued by Vinyals and Le, the current conversation systems are still unable to pass the Turing test, and the lack of consistent personal information is one of the most challenging constraints. In recent years, Li et al. learned interactive object-specific conversational styles by embedding users into sequence to sequence model. Al-Rfou et al. used similar user embedding techniques to simulate user personalization. Both studies required conversational data from each user to simulate her/his personality, but a lot of data are needed to mark the information location.

To make HRI more natural and harmonious, this article makes the following contributions for SHFR-III:

- A multi-sensor positioning subsystem is designed to reduce dependence on work environment and improve the overall positioning accuracy by data fusion of multiple sensors.
- The emotional recognition model based on facial expression and speech is used to deal with situations that a single modal would fail. At the same time, fuzzy algorithm is used to simulate emotional decision-making.
- Using default information to solve the problem of inconsistent personal information in dialogue, and maximum mutual information is taken as the objective function to reduce meaningless replies in the dialogue model.

Overview

Our interactive system works as follows (see Figure 2): First, we use multi-sensor positioning subsystem to find the exact location of the interactive objects. Then the robot adjusts the angle between the interactive objects, and the robot then obtains the facial and voice information of the interactive objects through cameras and microphones. The emotional interaction subsystem recognizes their emotional state and emotional decision-making, and SHFR-III will display the results of emotional decision-making with facial expression. The dialogue subsystem with personal information generates responses and presents them in the form of a voice.

Multi-sensor positioning subsystem

In this article, a multi-sensor positioning subsystem is designed, which includes an infrared positioning module, an auditory positioning module, and a binocular vision positioning module.

Design of positioning module

Infrared positioning based on infrared sensor array. Four pyroelectric infrared sensors with the same parameters are selected to form the sensor array as shown in Figure 3. The vertical equidistant distribution of the four sensors and the horizontal angle of the adjacent sensors are 30°. The output results of the four sensors from top to bottom are recorded as \( S_1, S_2, S_3, \) and \( S_4 \), respectively. The sensor array can cover the frontal range of 0–210.

The positioning function \( Fr (S_1, S_2, S_3, S_4) \) is defined as the mapping relationship between the sensor output values \( S_1, S_2, S_3, S_4 \) and the target location information \( \theta \). The output values of \( S_1, S_2, S_3, \) and \( S_4 \) are 1 or 0, which
correspond to the high and low level of sensor output. Therefore, the $F_r(S_1, S_2, S_3, S_4)$ can be described by the truth table as presented in Table 1.

### Table 1. Positioning function truth table.

| $S_4$ | $S_3$ | $S_2$ | $S_1$ | $F(\degree)$ | $\theta(\degree)$ |
|-------|-------|-------|-------|--------------|------------------|
| 0     | 0     | 0     | 1     | 0–30         | 15               |
| 0     | 0     | 1     | 1     | 30–60        | 45               |
| 0     | 1     | 1     | 1     | 60–90        | 75               |
| 1     | 1     | 1     | 1     | 90–120       | 105              |
| 1     | 1     | 1     | 0     | 120–150      | 135              |
| 1     | 1     | 0     | 0     | 150–180      | 165              |
| 1     | 0     | 0     | 0     | 180–210      | 195              |

Auditory positioning based on auditory sensor array. The three sensors are isosceles triangular in the vertical plane. Sensors 2 and 3 are arranged horizontally, and the distance is $p$. Sensor 1 is located on the vertical line of $B$ and $C$, and the distance is $q$. Sound source is recorded as $P(x, y, z)$. The distribution model is shown in Figure 4.

The time difference of sound arrival between sensor 2 and sensor 1 is defined as $t_{21}$, and the time difference of sound arrival between sensor 3 and sensor 1 is $t_{31}$. Based on the sampling frequency of sensor (50 KHz) and sound velocity (340 m/s), the relationship between time difference of sound arrival $T$ and distance difference $D$ is as follows

$$d = 6.8t$$

As we know,

$$\begin{align*}
PB - PA &= d_{21} \\
PC - PA &= d_{31}
\end{align*}$$

where $PA = \sqrt{x^2 + z^2 + q^2}$, $PB = \sqrt{(x-p)^2 + z^2}$, and $PC = \sqrt{(x+p)^2 + z^2}$.

We have the following equation

$$\begin{align*}
\sqrt{(x-p)^2 + z^2} - \sqrt{x^2 + z^2 + q^2} &= d_{21} \\
\sqrt{(x+p)^2 + z^2} - \sqrt{x^2 + z^2 + q^2} &= d_{31}
\end{align*}$$

The distance from the source $P$ to the origin is $\rho = \sqrt{x^2 + z^2}$. If the variable is $k = \sqrt{\rho^2 + q^2}$, we have the following equation
The vision positioning module.

\[
\begin{aligned}
\begin{cases}
\sqrt{(x-p)^2 + z^2} = d_{21} + k \\
\sqrt{(x+p)^2 + z^2} = d_{31} + k
\end{cases}
\end{aligned}
\]

(4)

\[k = \frac{2p^2 - 2q^2 - d_{21}^2 - d_{31}^2}{2(d_{21} + d_{31})}
\]

(5)

\[
\begin{aligned}
\begin{cases}
k^2 - q^2 + p^2 - 2xp = d_{21}^2 + k^2 + 2kd_{21} \\
k^2 - q^2 + p^2 + 2xp = d_{31}^2 + k^2 + 2kd_{31}
\end{cases}
\end{aligned}
\]

(6)

The polar angle of the sound source \( P \) relative to the origin is \( \theta \).

\[\theta = \cos^{-1}\left(\frac{x}{\rho}\right)\]

(7)

where \( x = \frac{(d_{31} - d_{21})(d_{31} + d_{21} + 2k)}{4p} \).

The polar coordinates of the sound source \((p, \theta)\) are as follows

\[
\begin{aligned}
\begin{cases}
\rho = \sqrt{k^2 - q^2} \\
\theta = \cos^{-1}\left(\frac{x}{\rho}\right)
\end{cases}
\end{aligned}
\]

(8)

**Visual positioning based on binocular stereovision.** The model of binocular vision positioning is shown in Figure 5. The focal length of the left and right eyes is \((f_{x1}, f_{y1})\) and \((f_{x2}, f_{y2})\), respectively. The coordinates of the center of light are \((c_{u1}, c_{v1})\) and \((c_{u2}, c_{v2})\), respectively. Binocular spacing is 2B. Left and right eye cameras have the same model and focal length, which can be approximately considered equal.

\[f_{y1} = f_{x1} = f_{y2} = f_{x2} = f\]

(9)

The transformation relationship between coordinates in the plane \((u, v)\) and coordinates in the world coordinate system \((x_w, y_w, z_w)\) is as follows

The parameters of the left and right eyes are introduced into the upper equation. We have the following equation

\[
\begin{aligned}
\begin{bmatrix}
u \\
v \end{bmatrix} &= \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \end{bmatrix} \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x_w \\
y_w \\
z_w \\
1 \end{bmatrix} = M \begin{bmatrix} x_w \\
y_w \\
z_w \\
1 \end{bmatrix}
\end{aligned}
\]

(10)

The working environment of the system is a closed room of \(5 \times 6\) m\(^2\). The coverage area of the positioning subsystem is shown in Figure 6, where O is the positioning center of the multi-sensor subsystem and DEGF represents the whole room. The coverage area of visual positioning module is the triangular area ABC, and polar angle positioning error \((E_\theta)\) and distance positioning error \((E_p)\) increase with the increase of target distance. The coverage of infrared positioning module is MNFG, and \(E_\theta\) decreases with the increase of target distance. The coverage of the auditory positioning module is DEGF, and \(E_p\) increases with the increase of the target distance, while \(E_\theta\) remains unchanged.

According to the working characteristics of multi-sensor positioning subsystem, the multi-level positioning fusion method shown in Figure 7 is proposed. In triangular ABC region, all three positioning modules work normally, and
the three-positioning data are fused. In rectangular MNGF region, only infrared positioning module and auditory positioning module can work, at this time, the two-positioning data are fused. In rectangular DENM region, only the auditory positioning module can work normally, and then the auditory positioning data can be output directly.

**Weighted fusion algorithms with variable weights**

The final result of the multi-sensor positioning subsystem, that is, the coordinates of the interactive target relative to the robot in the horizontal plane are expressed in polar coordinates $\rho, \theta$. The positioning data of infrared positioning module is $(\theta_r, \rho_r)$, the positioning data of auditory positioning module is $(\theta_s, \rho_s)$, and the positioning data of visual positioning module is $(\theta_v, \rho_v)$. The weighted fusion algorithm is constructed as follows

$$\begin{align*}
\theta &= m_r \theta_r + m_s \theta_s + m_v \theta_v \\
\rho &= n_r \rho_r + n_s \rho_s
\end{align*}$$

(12)

where $m_r$, $m_s$, and $m_v$ are the angle weighting coefficients of infrared positioning module, auditory positioning module, and visual positioning module, respectively. $n_r$ and $n_v$ are the distance weighting coefficients of auditory positioning module and visual positioning module, respectively.

In this article, three kinds of positioning modules are tested separately, and the corresponding positioning accuracy is calculated. The average $E_{\theta}$ of infrared positioning module is $15^\circ$ ($E_{\theta_r}$); the average $E_{\theta}$ of auditory positioning module is $3^\circ$ ($E_{\theta_s}$), and the average $E_{\theta}$ of visual positioning module is $1^\circ$ ($E_{\theta_v}$), and the average $E_{\rho}$ of infrared positioning module is $340$ mm ($E_{\rho_r}$); the average $E_{\rho}$ of visual positioning module is $100$ mm ($E_{\rho_v}$).

The weighting coefficients are inversely proportional to the positioning accuracy, and the weighting coefficients are calculated according to the sum of the coefficients being 1.

$$\begin{align*}
m_r : m_s : m_v &= \frac{1}{E_{\theta_r}} : \frac{1}{E_{\theta_s}} : \frac{1}{E_{\theta_v}} \\
n_r : n_s : n_v &= \frac{1}{E_{\rho_r}} : \frac{1}{E_{\rho_s}}
\end{align*}$$

(13)

$$\begin{align*}
m_r + m_s + m_v &= 1 \\
n_r + n_s + n_v &= 1
\end{align*}$$

The weighted fusion equation is as follows

$$\begin{align*}
\theta &= 0.05\theta_r + 0.24\theta_s + 0.71\theta_v \\
\rho &= 0.22\rho_r + 0.78\rho_v
\end{align*}$$

(14)

In equation (3), the weighting coefficient is constant. This equation is only applicable when all three positioning systems are working normally. However, as the location of the interactive target and the external environment change, one or some positioning systems may fail. According to the positioning accuracy of each positioning system determined by experiments, a weighted fusion algorithm with variable weights is proposed.

When the interactive object is out of the location range of visual positioning module or in a dark environment, $\theta_r$ and $\theta_s$ have data, $\theta_v$ has no data, so the equation is follows

$$\begin{align*}
\theta &= 0.17\theta_r + 0.83\theta_s \\
\rho &= \rho_s
\end{align*}$$

(15)

Only when the auditory positioning module works, it indicates that the interactive object is in a dim environment and out of the infrared positioning module. At this time, the positioning subsystem directly outputs the results of the auditory positioning module.

When in a noisy environment, the auditory positioning module stops working. At this time, when $\theta_r$ and $\theta_r$ have data, it means that the interacting object is in the triangular ABC region and the light is abundant.

$$\begin{align*}
\theta &= 0.06\theta_r + 0.94\theta_v \\
\rho &= \rho_v
\end{align*}$$

(16)

Only when the infrared positioning module works, $\theta_s$ is output directly, and then the positioning subsystem has no distance information output.

**Emotional interaction subsystem**

**Bimodal emotion recognition**

In this article, facial expression and voice emotion are fused by decision-level fusion.

**Facial emotion recognition.** Nodulus Facial Expression Analysis System (FaceReader) is used for bimodal emotion recognition. This article is based on discrete emotional classification. FaceReader data output format is based on Paul Ekman’s six basic expressions plus calm expressions.
Speech emotion recognition. Speech emotion recognition is a new research hotspot involving traditional speech signal processing, pattern recognition, human psychology, artificial intelligence, and other fields. The research of speech emotion recognition is based on discrete emotion classification system.

Feature extraction: The main emotional features in speech signal are prosodic features, spectrum-based features and sound quality features. Based on the study of emotional features of expression, this article extracts the acoustic parameters including the root mean square of energy, zero-crossing rate, fundamental frequency, vocal probability, MFCC, frequency and bandwidth of first, second and third formants, and the corresponding first-order difference as the dynamic parameters of speech signal. Statistical features of acoustic parameters and dynamic parameters are used as feature vectors for speech emotion recognition, and a total of 382-dimensional features are obtained.

Hierarchical support vector machine (SVM) classifier: First, all categories are divided into two subclasses, and then these two subclasses are further divided into subclasses until a separate category is obtained. This method can classify subclasses according to the degree of confusion between classes.

The degree of confusion between category $i(G_i)$ and category $j(G_j)$ is $Mix_{ij}$.

$$Mix_{ij} = \frac{(P(r = i|xG_j) + P(r = j|xG_i))}{2} \quad (17)$$

The higher the level of confusion, the more difficult it is to distinguish between category $i$ and category $j$. When determining the first level of classification, if the degree of confusion is greater than 0.1, the two categories can be divided into one subclass. If all confusion is less than 0.02, this category is treated as a subclass alone.

When the degree of confusion between a certain category and other categories is greater than 0.02 and less than 0.1, the classification situation cannot be directly judged. By calculating the total degree of confusion between the category and a subclass, the classification with high degree of total confusion can be selected.

$$T(a,B) = \sum_{i=1}^{m} Mix_{ai} \quad (18)$$

where $a$ is an emotional category, $B$ is a subclass, and $B$ includes emotional categories.

Decision-level fusion. This article chooses the weighted summation method through experiments.

| Table 2. Fuzzy emotional decision rules. |
| --- |
| Calm | Hap. | Sad. | Ang. | Sur. | Fear |
| Calm | Calm | 1.0 | Hap. | 0.9 | Sad. | 0.9 | Ang. | 0.8 | Sur. | 0.7 | Fear | 0.8 |
| Hap. | Hap. | 0.7 | Hap. | 0.95 | Ang. | 0.6 | Ang. | 0.95 | Sur. | 0.8 | Fear | 0.6 |
| Sad. | Sad. | 0.6 | Sad. | 0.7 | Sad. | 0.9 | Ang. | 0.6 | Sur. | 0.9 | Sad. | 0.75 |
| Ang. | Sur. | 0.7 | Sur. | 0.7 | Sur. | 0.7 | Ang. | 0.95 | Sur. | 0.5 | Fear | 0.6 |
| Sur. | Sur. | 0.8 | Sur. | 0.8 | Sad. | 0.8 | Ang. | 0.6 | Sur. | 0.95 | Fear | 0.65 |
| Fear | Fear | 0.8 | Sur. | 0.8 | Fear | 0.6 | Sur. | 0.8 | Fear | 0.9 | Fear | 0.9 |

Fuzzy emotional decision-making model

Fuzzy emotional decision-making takes the initial emotional state and the emotional state of the current interactive object as input and combines the input to make the fuzzy emotional decision through the fuzzy reasoning rules to generate the robot’s emotions.

This article refers to the way of emotional quantification in reference to quantify this emotional state. Seven emotional states are quantified as interval values. Considering external stimuli, the interval equivalence ratio is expanded to [0,7]. The Mamdani algorithm is used to construct a fuzzy affective decision model.

After analysis, the orthodox distribution curve of Gauss function is more in line with the characteristics that the affective control range is gradually weakening from the central point to the surrounding area. It shows that an emotional input belongs to the membership degree of the fuzzy emotional subset, which is convenient for calculation and processing. Assuming that all the emotional centers have the same influence range and ability, the Gauss function is as follows

$$f(x, c, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-c)^2}{2\sigma^2}} \quad (19)$$

where $c$ is the position of the central point of the function, $\sigma$ is the width of the number curve, $\sigma$ of each fuzzy subset is the same, and $c$ is different.

Based on a male volunteer, the fuzzy rules of orthogonal combination are equationed and the probabilities of each rule are evaluated. As presented in Table 2, the horizontal row represents the emotional state of the robot at the front moment, the vertical row represents the external stimulus, and the numerical value of the emotional state in the table represents the proportion of the rule.

Dialogue subsystem with personal information

In this article, a dialogue subsystem with personal information is proposed (see Figure 8). By giving the chat robot specific personal information, the robot can generate a response consistent with its given information. The system first uses the question classifier to distinguish whether the question needs personal information dialogue model to deal with. If yes, the model retrieves the most similar question
Questions (x) \hspace{1cm} \text{Classifier} (D) \hspace{1cm} \text{Open domain} \hspace{1cm} \text{Response (y)}

ps

Response (y)

margin

seq2seq model is prone to generate security answers like “I don’t know”, “Ha ha ha”, and “well”. So we use the anti-language model proposed by Li et al. and take maximum mutual information as the objective function of seq2seq, as shown in the following equations

\[ p(T|S) = \prod_{k=1}^{N} p(y_k|x_1, \ldots, x_r, y_1, \ldots, y_{k-1}) \] \hspace{1cm} (21)

\[ \hat{\theta} = \arg \max_{\theta} \left( \log P(S, T) \right) \] \hspace{1cm} (22)

\[ \hat{\theta} = \arg \max_{\theta} \left( \log P(T|S) - \lambda \log P(T) \right) \] \hspace{1cm} (23)

The log \( P(T|S) \) has the same representation as the maximum logarithmic likelihood model. \( \lambda \log P(T) \) is regarded as a penalty for candidate words with high probability for any input, and the penalty is controlled by parameter \( \lambda \).

Personal information dialogue model

This article constructs a personal information dialogue model based on the twin network idea. First, the two objects to be matched are represented by the deep learning model, and then the matching degree of the two objects can be output by calculating the similarity between the two representations. This article uses bidirectional long short-term memory (BiLSTM) to represent the semantic information of sentences.

The loss function used is the comparative loss function, which is often used in twin neural networks. This loss function can effectively deal with the relationship between paired data in twin neural networks. The expression of the comparative loss function is as follows

\[ L = \frac{1}{2N} \sum_{n=1}^{N} yd^2 + (1-s) \max(margin-d, 0)^2 \] \hspace{1cm} (20)

where \( d = ||a_n - b_n||_2 \) represents the Euclidean distance of two sample features, \( s \) is the label of whether two samples match, and \( margin \) is the set threshold.

Dialogue model based on maximum mutual information

The open domain dialogue model is based on seq2seq. However, if we only rely on the maximum likelihood estimation, even if we train with a large number of data, the seq2seq model is prone to generate security answers like “I don’t know”, “Ha ha ha”, and “well”. So we use the anti-language model proposed by Li et al. and take maximum mutual information as the objective function of seq2seq, as shown in the following equations

\[ p(T|S) = \prod_{k=1}^{N} p(y_k|x_1, \ldots, x_r, y_1, \ldots, y_{k-1}) \] \hspace{1cm} (21)

\[ \hat{\theta} = \arg \max_{\theta} \left( \log P(S, T) \right) \] \hspace{1cm} (22)

\[ \hat{\theta} = \arg \max_{\theta} \left( \log P(T|S) - \lambda \log P(T) \right) \] \hspace{1cm} (23)

The log \( P(T|S) \) has the same representation as the maximum logarithmic likelihood model. \( \lambda \log P(T) \) is regarded as a penalty for candidate words with high probability for any input, and the penalty is controlled by parameter \( \lambda \).

Pre-generative words have a greater impact on sentence diversity than post-generative words. In order to ensure the fluency of sentences as much as possible, only the high probability candidate words generated in the early stage are punished in the process of sentence generation by decoder. The \( \gamma \) in equation is set to 1. Only the first word of the sentence is punished to ensure the coherence of the sentence as far as possible.

The model consists of encoder and decoder. In the encoder part, two layers of BiLSTM neural network are adopted, the number of units is 512, and the dimension of word vector is set to 300. Bahdanau’s attention mechanism was used. In the training process, dropout mechanism is adopted with a retention rate of 0.5, Adam learning rate is set to 0.001, batch size is set to 32, and the number of data iterations is 128.

Experiment

Experimental results of multi-sensor positioning subsystem

Each positioning module experiments in different environments. To verify that the multi-sensor positioning system can be applied to various working environments, target

![Diagram](image-url)
Positioning experiments are carried out in normal lighting environment, dark environment and noise environment respectively. The experimental results are given in Table 3.

From the experimental results, we can see that the some positioning module designed in this article will fail in different scenarios, but other positioning modules still work normally and have good stability.

Fusion experiment results of multi-sensor positioning subsystem.

The multi-sensor positioning system is used to locate points in the environment, and the results are fused. Some statistical results are given in Table 4.

The experimental results show that the positioning accuracy after fusion is higher than that of single positioning system, and the stability of positioning has been greatly improved.

Experiments on emotional interaction

Speech emotion recognition. This article uses CASIA Chinese Emotional Corpus for model training, openSMILE for feature extraction, principal component analysis (PCA) contribution selection 95%, and LIBSVM toolkit developed by Professor Lin of Taiwan University (Kernel function is three-degree polynomial). The degree of confusion obtained by experiments is given in Table 5.

According to the degree of confusion, the hierarchical SVM as shown in Figure 9 is designed.

The comprehensive recognition rate and recognition rate of each level classifier are given in Table 6.

Bimodal emotional fusion. Some data in eNTERFACE’05 multi-modal emotion database is used to verify the validity. The five emotional expressions of happiness, surprise, fear, sadness, and anger are screened by scoring principle. Thirty pieces of data are selected for each emotion for bimodal emotion recognition.

The results of single-modal and bimodal emotion recognition are compared as given in Table 7. The experimental results show that the performance of bimodal emotion recognition is better than single-modal, with an average recognition rate of 59.34%. Table 8 gives the result of emotion recognition.

| Table 3. Positioning results in different environment. |
|-----------------------------------------------|
| Env. | Infra. | Aud. | Vision | Real. |
|------|--------|------|--------|-------|
| Nor. | θ (°)  | 75   | 82.1   | 80.2  | 80    |
|      | ρ (mm) | None | None   | 826   | 800   |
| Dark | θ (°)  | 75   | 63.7   | Fail  | 65    |
|      | ρ (mm) | None | 1181.6 | 1300  |
| Noi. | θ (°)  | 75   | Fail   | 74.5  | 70    |
|      | ρ (mm) | None | 1250.1 | 1300  |

| Table 4. Fusion experiment results of multi-sensor positioning system. |
|-----------------------------------------------|
| Point | Infra. | Aud. | Vision | Fusion | Real. |
|-------|--------|------|--------|--------|-------|
| θ (°) | 1      | 105  | 93.9   | 95.4   | 95    |
| ρ (mm)| None   | 728.3| 1026.7 | 980    | 100   |
| θ (°) | 2      | 75   | 88     | 76     | 77.8  |
| ρ (mm)| None   | 1467.5| 1632.4 | 1595.4 | 1600  |

| Table 5. The confusion between emotional categories. |
|-----------------------------------------------|
| Calm | Hap. | Sur. | Fear | Sad. |
|------|------|------|------|------|
| Hap. | 0.1  |      |      |      |
| Sur. | 0.025| 0.075| 0.1  |      |
| Fear | 0    | 0.05 | 0.1  |      |
| Sad. | 0    | 0    | 0.2  |      |
| Ang. | 0    | 0.05 | 0.05 | 0.05 |

| Table 6. The recognition rate of each SVM and hierarchical model. |
|-----------------------------------------------|
| Level | 1 | 2 | 3 | 4 | Hierarchical SVM |
|-------|---|---|---|---|-----------------|
| Rate  | 85.83% | 90% | 85% | 73.33% | 66.67% |

| Table 7. Comparison of single-mode and double-mode recognition results. |
|-----------------------------------------------|
| Method | Rate (%) |
|--------|----------|
| Face expression recognition | 52.67 |
| Speech emotion recognition | 38.67 |
| Bimodal emotional recognition | 59.34 |

| Table 8. Different emotional recognition results. |
|-----------------------------------------------|
| Hap. | Sur. | Fear | Sad. | Ang. | Rate (%) |
|------|------|------|------|------|----------|
| Hap. | 23   | 4    | 3    |      | 76.67    |
| Sur. | 1    | 21   | 8    |      | 70       |
| Fear | 1    | 9    | 15   | 5    | 50       |
| Sad. | 11   | 3    | 16   |      | 53.33    |
| Ang. | 3    | 9    | 1    | 3    | 14       | 46.67    |
When the probability of detecting disgust is the highest, speech emotion recognition is not carried out, and disgust is taken as the final recognition result.

**Fuzzy emotional decision-making.** After setting the initial emotional state of the robot, with the change of external stimulus, the current emotional state of the robot is simulated and calculated.

When the initial emotional state of the robot is calm, the emotional state of the interactive object changes in turn. The emotional change curve of the robot is given in Figure 10(a). When the robot changes to the state of surprise, as the external stimulus gradually turns to fear, the robot will change to fear. When it encounters the external stimulus of sadness, the robot will then turn to sadness, and keep the sad state under the stimulus of disgust. Figure 10(b) shows the simulation of experimental results with sad initial state.

The simulation results show that under the fuzzy affective decision-making model, the change of the robot’s emotional state under the external stimulus is slow and continuous, and the calculation results are in accordance with the reasoning rules and human’s emotional changes.

**Experiments of dialogue subsystem**

In this article, the accuracy, F1 Score (F1) and Area under the curve (AUC) values are selected to evaluate the performance of the question classification model. The values are 87.69%, 0.8754, and 0.8797, respectively.

In the personal information reply model, since five identities are set in this article, accuracy refers to whether the category of sentences with the greatest similarity belongs to the label category, which is 87.4%.

The open domain dialogues are evaluated by manual and bilingual evaluation understudy (BLEU) methods. In the experiment, the penalty coefficient of the maximum mutual information model is 0.5 and γ is set to 1. The BLEU results of the two models are 0.17 and 0.25, the mutual information model is better. In the manual evaluation, most people think that the results of mutual information model are similar to those of maximum likelihood model, 30% think mutual information model is better than maximum logarithmic likelihood model.

In the overall dialogue system, this article randomly selected some dialogues (the dialogues are in Table 9), and asked volunteers to evaluate the following aspects:

- **Naturalness**: whether the generated response is natural and smooth.
- **Logic**: whether the generated response is logically related to the problem.
- **Information consistency**: whether the response to personal information is consistent.
- **Variety**: whether there are multiple ways to respond to a question.

From Table 10, we can see that our model is superior to the ordinary seq2seq model in each index, especially in information consistency, because this article adds personal information reply model to ensure the consistency of personal information.

---

**Table 9. Samples of dialogues.**

| Chinese                        | English (translated) |
|-------------------------------|----------------------|
| 你性别是什么？                 | What is your gender? |
| 女孩                          | Girl                 |
| 你会唱歌吗？                   | Can you sing?        |
| 不会                          | No                   |
| 吃饭了吗                      | Did you have your meal? |
| 你想请我去吃饭                  | Would you like to invite me to dinner? |

**Table 10. Evaluation of responses.**

| Method  | Naturalness (%) | Logic (%) | Consistency (%) | Variety (%) |
|---------|-----------------|-----------|-----------------|-------------|
| Seq2seq | 71.4            | 38.7      | 2.1             | 1.6         |
| Ours    | 75.9            | 52.7      | 75.7            | 22.2        |
Conclusion and future work

This article designs an interactive system for humanoid robot SHFR-III. The system can use multi-sensor positioning subsystem to locate accurately in complex environment, use bimodal emotion recognition model and fuzzy emotion decision-making model to complete human–robot emotional interaction in the form of robot facial expression, and the dialog subsystem with personal information can complete the response consistent with the default information. The system has the advantages of easy implementation and good interactivity, and can be applied in the fields of elder group’s care, we can understand their emotional world through HRI, and use chatting to relieve their loneliness. Besides, it can also be used in the field of robot teaching and autism treatment and so on.

Our work is only a small step toward achieving a harmonious HRI, and there are many future directions:

- **Positioning system**: Constructing three-dimensional auditory sensor array to improve positioning accuracy and optimize fusion strategy.
- **Emotional interaction**: The research of emotional recognition in this article is the discrete emotional recognition. Later, we can study dimension emotional recognition, which is convenient for the joint research with the dimension emotional model, taking into account personal, psychology, morality, and other factors.
- **Dialogue system**: Integrating emotional information into the dialogue system to generate a response with emotional information style. Introducing reinforcement learning into dialogue system, and improving the naturalness and logic of response by adding a “teacher.”

In addition, specific subsystems should be added for different application scenarios, such as robot hand grasping, motion trajectory control, smart home control, and so on.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Bin Cao https://orcid.org/0000-0003-4708-4021

References

1. Zhao Y. Survey of human-computer interaction research. *Chin Comput Commun* 2017; 23: 24–25+28.
2. Song S and Yamada S. Expressing emotions through color, sound, and vibration with an appearance-constrained. In: ACM/IEEE international conference on human-robot interaction part FI27194, Vienna, Austria, 6–9 March 2017, pp. 2–11.
3. Gunes H, Celiktutan O, and Sarıyani E. Live human-robot interactive public demonstrations with automatic emotion and personality prediction. *Philos T R Soc B* 2019; 374(1771): 20180026.
4. Ali M, Minoo A, and Mohammad Z. Design and realization of a sign language educational humanoid robot. *J Intell Robot Syst* 2016; 95(1): 3–17.
5. Michiel DJ, Kevin Z, and Aaron MR. Towards a robust interactive and learning social robot. In: 17th international conference on autonomous agents and MultiAgent systems, Stockholm, Sweden, 10–15 July 2018, pp. 883–891.
6. Kumra S and Kanam C. Robotic grasp detection using deep convolutional neural networks. In: IEEE international conference on intelligent robots and systems, Vancouver, Canada, 24–28 September 2017, pp. 769–776. IEEE.
7. Bin H, Ting L, and Shan H. Product sustainability assessment for product life cycle. *J Clean Prod* 2019; 206: 238–250.
8. He B, Wang S, and Liu YJ. Underactuated robotics: a review. *Int J Adv Robot Syst* 2019; 16(4): 1–29.
9. Xin J, Ke X, Yang Y, et al. Development of the system of a humanoid robot head with facial expressions [in Chinese]. *CAAI Trans Intell Syst* 2015; 10(04): 555–561.
10. Cid F, Moreno J, Bustos P, et al. Muecas: a multi-sensor robotic head for affective human robot interaction and imitation. *Sensors* 2014; 14(5): 7711–7737.
11. Laurenzi A, Kanoulas D, Mingo Hoffman E, et al. Whole-body stabilization for visual-based box lifting with the COMAN+ robot. In: *Proceedings – 3 rd IEEE international conference on robotic computing*, Naples, Italy, 25–27 February 2019, pp. 445–446. IEEE.
12. Gala D, Lindsay N, and Sun L. Realtime active sound source localization for unmanned ground robots using a self-rotational bi-microphone array. *J Intell Robot Syst* 2019; 95(3-4): 935–954.
13. Yang Y, Liu Y, Jiao X, et al. Research progress on human body localization in smart home system. *Chin J Biomed Eng* 2013; 7(5): 467–470.
14. Zhu Y. The research of multimodal human-robot interaction of humanoid emotional robot [in Chinese]. *Sahnghai University* 2018, pp. 46.
15. Picard RW. *Affective computing*. London: MIT Press, 1997.
16. Vinylas O and Le Q. A neural conversational model. arXiv preprint arXiv:1506.05869, 2015.
17. Li J, Galley M, Brockett C, et al. A persona-based neural conversation model. In: *Proceeding 54th annual meeting of the association for computational linguistics*, Berlin, Germany, 7–12 August 2016, pp. 944–1003.
18. Al-Rfou R, Pickett M, Snaidr J, et al. Conversational contextual cues: the case of personalization and history for response ranking. arXiv preprint arXiv:1606.00372, 2016.
19. Qian Q, Huang M, Zhao H, et al. Assigning personality/identity to a chatting machine for coherent conversation generation. arXiv preprint arXiv:1706.02861, 2017.
20. Delic V, Bojanic M, Gnjatovic M, et al. Discrimination capability of prosodic and spectral features for emotional speech recognition. Electron Electr Eng 2012; 18(9): 51–54.
21. Wu CH and Liang WB. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. In: 2015 international conference on affective computing and intelligent interaction (ACII), Xi’an, China, 21–24 September 2015, pp. 477–483.
22. Yu D. Research of key technology of affective computing [in Chinese]. Donghua University 2009, pp. 27.
23. Guan P, Li B, Lv X, et al. Attention enhanced bi-directional LSTM for sentiment analysis. J Chin Inform Proc 2019; 33 (2): 105–111.
24. Hadsell R, Chopra S, and LeCun Y. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06), vol. 2, New York, NY, USA, 17–22 June 2006, pp. 1735–1742.
25. Li J, Galley M, Brockett C, et al. A diversity—promoting objective function for neural conversation models. Comput Sci 2015, arXiv preprint arXiv:1510.03055.