

A Sociable Human-robot Interaction Scheme Based on Body Emotion Analysis

Tehao Zhu, Zeyang Xia*, Jiaqi Dong, and Qunfei Zhao

Abstract: Many kinds of interaction schemes for human-robot interaction (HRI) have been reported in recent years. However, most of these schemes are realized by recognizing the human actions. Once the recognition algorithm fails, the robot's reactions will not be able to proceed further. This issue is thoughtless in traditional HRI, but is the key point to further improve the fluency and friendliness of HRI. In this work, a sociable HRI (SoHRI) scheme based on body emotion analysis was developed to achieve reasonable and natural interaction while human actions were not recognized. First, the emotions from the dynamic movements and static poses of humans were quantified using Laban movement analysis. Second, an interaction strategy including a finite state machine model was designed to describe the transition regulations of the human emotion state. Finally, appropriate interactive behavior of the robot was selected according to the inferred human emotion state. The quantification effect of SoHRI was verified using the dataset UTD-MHAD, and the whole scheme was tested using questionnaires filled out by the participants and spectators. The experimental results showed that the SoHRI scheme can analyze the body emotion precisely, and help the robot make reasonable interactive behaviors.

Keywords: Body emotion analysis, finite state machine, fuzzy inference, human-robot interaction, Laban movement analysis.

1. INTRODUCTION

Human-robot interaction (HRI) is one of the most popular research fields in robotics. Unlike human-computer interaction, robots in HRI have similar profiles to humans (partially or integrally). With the rapid development of artificial intelligence and robot technology, HRI robots are expected to possess better social and interactive skills.

Much research has been devoted to various key technologies of HRI. In the field of human perception, study areas include data capturing and processing of multimodal information, e.g., image [1, 2], sound [3], and depth [4-6]. In the field of action recognition, study areas include movement [7, 8], expression [9], speech [10], and intonation [11] recognition. In the field of intention inference and interaction strategy, study areas include multi-agent systems [12], neural network [13, 14], fuzzy inference [15, 16], and deep learning [17].

Vision-based HRI can achieve recognition of human action. Several typical actions, e.g., greeting, hand shaking, hugging, throwing, can be well recognized by many algorithms. Once an action is recognized, the appropriate in-

teractive behavior can be selected or generated. However, the action recognition algorithm might fail when the class of the performed action is not included in the training samples. Nowadays, studies on the interaction strategy in such case are inadequate. In fact, solving the HRI when the human action cannot be recognized will be a key step in the development of comprehensive interaction logic, and will make the interactive behavior of the robot friendlier and more natural.

Previous studies have attempted to tackle this problem. Many of them addressed the importance of emotion analysis for improving robot social skills. Bohus *et al.* [18] used linguistic hesitation actions to signal the system's state of confusion, which can generate additional time for collecting evidence and resolving uncertainties. Aly *et al.* [19] developed an adapted customized verbal-nonverbal robot's behavior based on personality dimensions. They proposed a behavior expression animation toolkit using linguistic and contextual information to generate a corresponding synchronized set of gestures. Glowinski *et al.* [20] characterized the expressions of emotions by means of movement and gesture. They adopted a layered ap-

Manuscript received July 20, 2017; revised January 17, 2018; accepted September 27, 2018. Recommended by Associate Editor Myung Geun Chun under the direction of Editor Euntai Kim. This work was supported by National Natural Science Foundation of China (No.61773365), Major Research Plan of the National Natural Science Foundation of China (No. 91646205), Major Project of Guangdong Province Science and Technology Department (No. 2014B090919002), Shenzhen Research Project (No. GJHS20160331190459402).

Tehao Zhu, Jiaqi Dong and Qunfei Zhao are with Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mails: {zthjoe, saberfate, zhaoqf}@sjtu.edu.cn). Zeyang Xia is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: zy.xia@siat.ac.cn).

* Corresponding author.

proach to model the position and dynamics of the head and hands from low-level physical measures toward descriptors of overall motion features. Liu *et al.* [21] modeled the communication atmosphere based on the emotional states of humans and robots. They estimated the human emotion from speech and gestures using weighted fusion and fuzzy inference.

In this work, we developed a sociable HRI (SoHRI) scheme based on body emotion analysis to achieve reasonable and natural interaction when dealing with unrecognized human actions. The major contributions of our work include the following aspects: 1) The emotions from the dynamic movements and static poses of humans were quantified using Laban movement analysis (LMA); 2) A finite state machine (FSM) model was constructed to describe the transition regulations of the human emotion state based on the quantified body emotions, and then appropriate interactive behavior was selected according to the inferred human emotion state. Differing from previous studies, the proposed SoHRI scheme can guarantee the continuity of interaction procedures, so that more information will be obtained, which is valuable for the subsequent recognition and interaction. That is the meaning of “sociable”.

The remainder of this paper is organized as follows: Section 2 introduces the overall process of SoHRI scheme. Section 3 discusses the body emotion analysis algorithm, including movement emotion quantification and torso pose emotion labeling. Section 4 first constructs an FSM model to describe the transition regulations of the human emotion state, and then designs the interaction strategy. Section 5 shows the experimental results, and Section 6 presents a summary.

2. SCHEME OF SOHRI

The proposed SoHRI scheme is shown in Fig. 1. A Microsoft Kinect [22] is equipped as the capture device to perceive the human joint position data, and the humanoid robot NAO [23] is adopted as the interactive robot. If the captured human action can be recognized by traditional algorithms, the corresponding robot reactions are chosen directly; otherwise, our SoHRI scheme can analyze the emotion contained in the movements, and finally help the robot to perform suitable interactive movements and speeches.

The joint positions captured by Kinect are preprocessed for body emotion analysis. The time series of the joint position data are transformed into several angular velocities of the arms, several linear velocities of the body movement, and the tilt angles of the spine and shoulder. We denote the set of angular velocities by W , the set of linear velocities by V , the tilt angle of spine by θ_{T_0} , and tilt angle of shoulder by θ_{T_1} , respectively. Detailed descriptions of W , V , θ_{T_0} , and θ_{T_1} will be given in Section 3.

Describing and modeling the relationships between

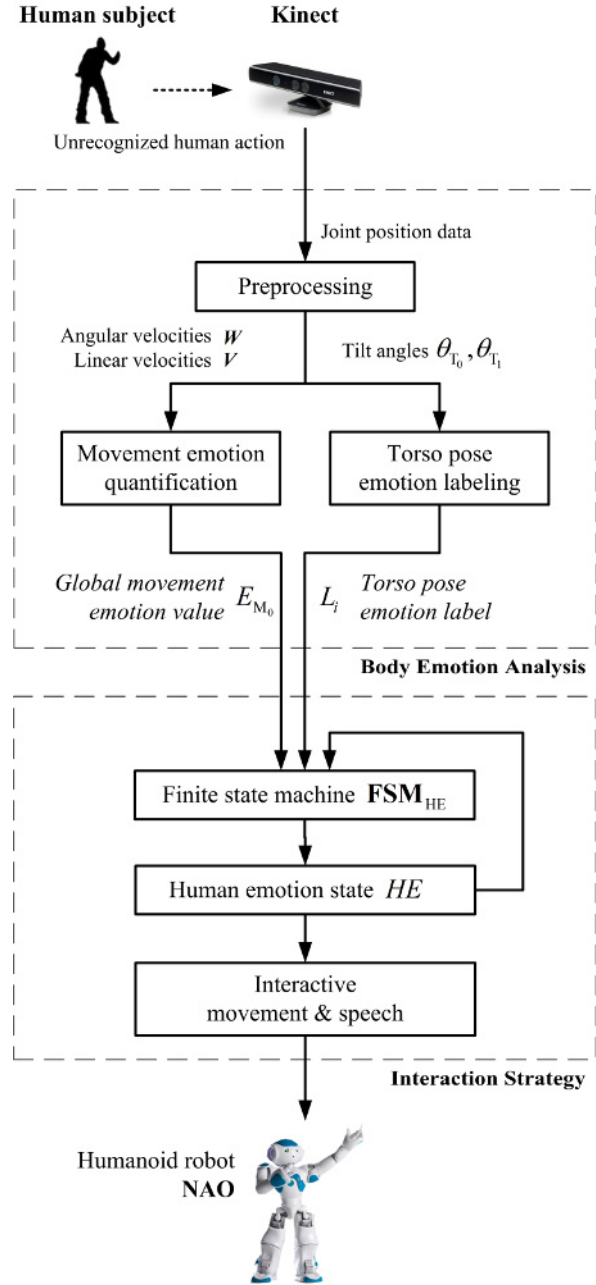


Fig. 1. Scheme of SoHRI.

human movements and emotions is a crucial phase of SoHRI. We use the Laban movement analysis (LMA) [24] as the basis of emotion analysis. As seen in Fig.1, there are two operations in the emotion analysis phase: the movement emotion quantification and the torso pose emotion labeling. The former is performed to determine a global movement emotion depending on W and V , and the latter is to give an emotion label depending on θ_{T_0} and θ_{T_1} . The global movement emotion, denoted by E_{M_0} , changes within $[-1,1]$. The torso pose emotion label includes six kinds, which are represented by L_i ($i = 1, 2, \dots, 6$) here-

after. E_{M_0} and L_i are the quantification results of body emotion analysis.

In the interaction strategy phase, the human emotion state is categorized into five gradations: very negative, negative, normal, positive, and very positive. An FSM model is constructed to describe the transition regulations of the human emotion state. E_{M_0} and L_i obtained from body emotion analysis constitute the input events of the FSM model, so the current human emotion state, represented by HE , can be inferred afterwards. The state-transition functions of the FSM model are specified in Section 4. Finally, appropriate interactive behavior including interactive movement and speech is selected for the robot to perform according to HE .

3. BODY EMOTION ANALYSIS

LMA method was presented by a Hungarian dancer, Rudolf Laban [24]. LMA describes and interprets various human movements. In recent years, various studies have used LMA to analyze the emotion representation in the human body. Kim *et al.* [22] extracted the low-level features of the trajectories of body joint positions, and developed an emotional motion representation through LMA. Cheng [25] explained the connection between robot action organization elements and the user's understanding based on LMA and perceptual learning. She designed three set of robot movement rules to integrate with the scenarios. Juan [26] studied the characteristics of a set of motions with special style and presented a motion style synthetic method based on LMA.

LMA contains four main components: body, effort, shape, and space harmony [24, 27, 28]. Space harmony explores the relationship between people and their surroundings. It regulates the body movement in a kinesphere. In each plane of kinesphere, a pair of opposite emotions is evaluated based on the direction and extent of the movement. Furthermore, the torso is the root of the human skeleton. Several appearances of the torso are considered to associate with specific emotions.

According to the above description, space harmony coincides with the requirement of our work. So it is chosen as the measure for analyzing body emotion. With the help of it, movement emotion quantification and torso pose emotion labeling are developed to achieve the body emotion analysis.

3.1. Movement emotion quantification

Movement projection and emotion matching: All movements can be modeled using the spatial Cartesian coordinates. Space harmony regulates the human movement in a kinesphere, which contains three orthogonal planes: the horizontal plane, the wheel plane, and the vertical plane [27]. As the name suggests, the three planes correspond to the XOZ, XOY, and YOZ planes in the Kinect

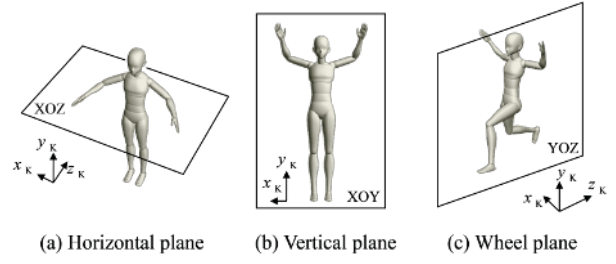


Fig. 2. Three projection planes for movement emotion quantification.

coordinate system (shown in Fig. 2).

In each plane, the projected movements are matched with a pair of opposite emotions. Based on the space harmony of LMA, the relationships are summarized in Table 1 [27].

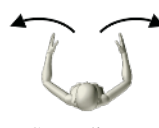

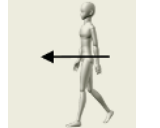

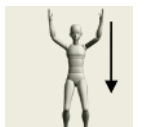

Emotion quantification: Table 1 gives the concepts of the movement emotions only in a qualitative way. The detailed methodology for quantifying the value of the movement emotion in each plane should be further given. We use the specific movement data of several joints to quantify the movements given in Table 1. In each plane, an emotion value is derived from the directions and values of the joint movements to describe the positive or negative degree of the movement. The emotion values in the horizontal plane, the vertical plane, and the wheel plane are denoted by E_H , E_V , and E_W , respectively. By convention, hospitable, encouraged, and active are positive emotions, while impassive, distressed, and scared are negative emotions. So the movements of spreading, ascending, and advancing show positive emotions, and the movements of enclosing, descending, and retreating show negative emotions.

Fig. 3 illustrates the associated angular and linear velocities, which can be derived by the geometrical transformation and the difference between the adjacent values of the time series of joint position data perceived by Kinect. In Fig. 3(a), the angular velocities of the upper and lower arms in the horizontal plane are represented by $\omega^* \in W$. The positive direction corresponds to clockwise rotation in the top view. In Fig. 3(b), the velocities of the elbows, wrists, and center hip along the y-axis are represented by $v_y^* \in V$. In Fig. 3(c), the center hip velocity along the z-axis is represented by $V_z^{CH} \in V$, which stands for the general displacement of the whole body. Note that moving forward results in the reduction of V_z^{CH} because of the definition of the Kinect coordinate system, which differs slightly from the usual cases.

Based on Table 1 and the above velocity definitions, we derive the following formulas for achieving E_H , E_V , and E_W :

$$E_H = S_H[(\omega^{RU} + \omega^{RF}) - (\omega^{LU} + \omega^{LF})], \quad (1)$$

Table 1. Movements and emotions in each projection plane.

Horizontal plane		Vertical plane		Wheel plane	
Movement	Emotion	Movement	Emotion	Movement	Emotion
 Spreading	Hospitable	 Ascending	Encouraged	 Advancing	Active
 Enclosing	Impassive	 Descending	Distressed	 Retreating	Scared

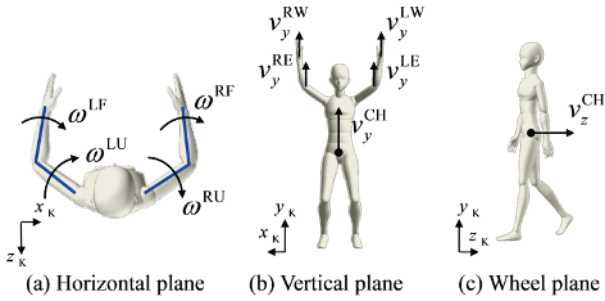


Fig. 3. Angular and linear velocities involved in movement emotion quantification.

$$E_V = S_V[\rho v_y^{CH} + (1 - \rho) \times (v_y^{LE} + v_y^{LW} + v_y^{RE} + v_y^{RW} - 4v_y^{CH})], \quad (2)$$

$$E_W = S_W(-v_z^{CH}), \quad (3)$$

where

$$S_u(x) = \frac{2}{1 + e^{-a_u x}}, \quad (4)$$

$$\rho = \begin{cases} 0, & |v_y^{CH}| < V_1^T, \\ y = 0.5, & V_1^T \leq |v_y^{CH}| \leq V_2^T, \\ z = 1, & |v_y^{CH}| > V_2^T. \end{cases} \quad (5)$$

$S_u(x)$ for $u = H, V$ are the normalized functions to make E_H , E_V , and E_W all lie within $[-1, 1]$. ρ is a factor to determine the influence of the torso velocity (represented by v_y^{CH}). Formula (5) indicates that ρ can take 0, 0.5, or 1 according to the value of v_y^{CH} .

The following parameters for movement emotion quantification can be used for reference: $a_H = 0.6 \text{ rad}^{-1}$, $a_V = 7.2 \text{ (m/s)}^{-1}$, $a_W = 3.8 \text{ (m/s)}^{-1}$, $V_1^T = 0.5 \text{ m/s}$, and $V_2^T = 1 \text{ m/s}$.

Global movement emotion: E_H , E_V , and E_W are quantified based on the projected movements in three orthogonal planes. In order to evaluate the emotion of the whole

body movement comprehensively, the values of E_H , E_V , and E_W should be integrated to generate a global movement emotion, which is denoted by E_M . In this work, fuzzy inference is used to achieve E_M .

The inputs to the fuzzy inference are E_H , E_V , and E_W . Their values are categorized into three groups: *Positive* (+), *Normal* (0), and *Negative* (-). The fuzzy membership functions of E_H , E_V , and E_W are based on S-shaped and bell-shaped curves:

$$S^-(E_u) = \begin{cases} 1 - 2(1 + E_u)^2, & -1 \leq E_u < -0.5, \\ 2E_u^2, & -0.5 \leq E_u \leq 0, \\ 0, & 0 < E_u \leq 1, \end{cases} \quad (6)$$

$$S^0(E_u) = \begin{cases} 0, & -1 < |E_u| \leq 1, \\ 2(2E_u^2 + 1)^2, & -0.5 \leq E_u \leq -0.25, \\ 1 - 8E_u^2, & -0.25 \leq E_u \leq 0.25, \\ 2(2E_u^2 - 1)^2, & 0.25 < E_u \leq 0.5, \end{cases} \quad (7)$$

$$S^+(E_u) = \begin{cases} 0, & -1 \leq |E_u| < 0, \\ 2E_u^2, & 0 \leq E_u \leq 0.5, \\ 1 - 2(1 - E_u)^2, & 0.5 < E_u \leq 1, \end{cases} \quad (8)$$

where $E_u \in \{E_H, E_V, E_W\}$.







The value of the output E_M is also categorized into *Positive*(+), *Normal*(0), and *Negative*(-). The membership functions of E_M are based on sigma-shaped and triangular-shaped curves:

$$S^-(E_M) = \begin{cases} 1, & -1 \leq E_M < -0.4, \\ -2.5E_M, & -0.4 \leq E_M \leq 0, \\ 0, & 0 < E_M \leq 1, \end{cases} \quad (9)$$

$$S^0(E_M) = \begin{cases} 0, & 0.4 < |E_M| < 1, \\ 2.5E_M + 1, & -0.4 \leq E_M < 0, \\ -2.5E_M + 1, & 0 \leq E_M < 0.4, \end{cases} \quad (10)$$

$$S^+(E_M) = \begin{cases} 0, & -1 \leq |E_M| < 0, \\ 2.5E_M, & 0 \leq E_M \leq 0.4, \\ 1, & 0.4 < E_M \leq 1. \end{cases} \quad (11)$$

Table 2. Torso poses and emotions.

Spine	Upright		Lean Forward		Lean Backward	
Shoulder						
	Normal	Tilt	Normal	Tilt	Normal	Tilt
Emotion	Formal	Relaxed	Sad	Negligent	Optimistic	Provocative

The membership functions of the above inputs and outputs are shown in Fig. 4 and Fig. 5, respectively.

As the output of the fuzzy inference, E_M is formalized in the form of the following fuzzy IF-THEN rules:

a) IF E_H is (+) OR E_V is (+) OR E_W is (+), THEN E_M is (+).

OR

b) IF E_H is (0) OR E_V is (0) OR E_W is (0), THEN E_M is (0).

OR

c) IF E_H is (-) OR E_V is (-) OR E_W is (-), THEN E_M is (-).

The above three fuzzy rules define the basic relationships between the emotions in the individual planes and the global one. Each individual rule is inferred by the OR operation, which is equal to the maximum operation:

$$Q_- = \max\{S^-(E_H), S^-(E_V), S^-(E_W)\},$$

$$Q_0 = \max\{S^0(E_H), S^0(E_V), S^0(E_W)\},$$

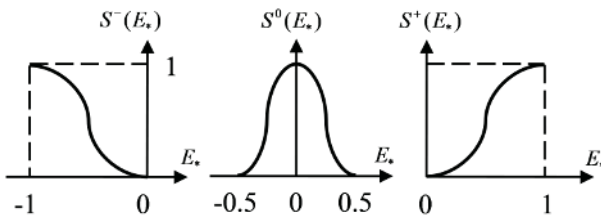


Fig. 4. Membership functions of the input movement emotions.

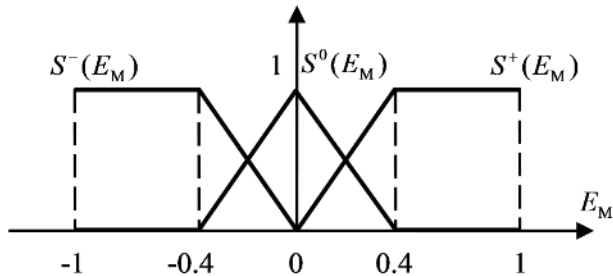


Fig. 5. Membership functions of the output global emotion.

$$Q_+ = \max\{S^+(E_H), S^+(E_V), S^+(E_W)\}, \quad (12)$$

where Q_- , Q_0 , and Q_+ denote the fuzzy outputs due to rule a), b) and c), respectively. As the three fuzzy rules are also associated by the OR operation, E_M is given by:

$$E_M = \max\{Q_-, Q_0, Q_+\}. \quad (13)$$

Defuzzification is performed using the centroid method [29]. For calculating the final output of the global movement emotion E_{M_0} , we use several discrete sampling points as follows:

$$E_{M_0} = \frac{\sum_{k=1}^{11} e_k E_M e_k}{\sum_{k=1}^{11} E_M e_k}, \quad (14)$$

where $e_k = 0.1(k-1)$, for $k = 1, 2, \dots, 11$, are the discrete sampling points for centroid method.

3.2. Torso pose emotion labeling

In the field of human motion perception, the human body is often simplified into the stick figure model, in which the torso is represented by an “I” type structure, as shown in Fig. 6(a). The poses of spine and shoulder (I_0 and I_1) dominate the performance of the upper body. And according to space harmony of LMA, Several appearances of the torso are considered to associate with specific emotions, as shown in Table 2 [28]. Table 2 indicates that the torso pose emotions are determined by the tilt angle of I_0 and I_1 . Here we propose the method of calculating these two angles and the emotion labeling regulation.

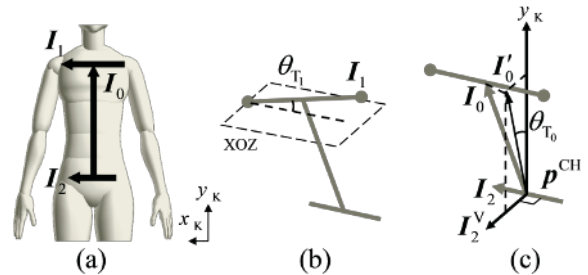


Fig. 6. The stick figure model of the torso: (a) the “I” type structure; (b) the tilt angle of the shoulder θ_{T_1} ; and (c) the tilt angle of the spine θ_{T_0} .

The tilt angle of I_1 is represented by θ_{T_1} , which is the angle between I_1 and the horizontal plane XOZ (Fig. 6(b)). The tilt angle of I_0 is related to the torso rotation, which can be calculated based on the following procedures (also illustrated in Fig. 6(c)):

1) Obtain the orientation of the body. Here, we use the vertical vector of the projection of the hipline I_2 in the horizontal plane to represent the orientation, which is denoted by I_2^V .

2) Project I_0 into the plane formed by I_2^V and the axis y_k , where the projection is I_0' .

3) The angle θ_{T_0} between I_0' and y_k is considered as the tilt angle of . θ_{T_0} is positive when the person leans forward and is negative when he/she leans backward.

The deformation of the torso is not as flexible as that of the whole body; thus, it is not necessary to estimate the torso emotion intensively. The six labels of the torso emotions are denoted by L_i ($i = 1, 2, \dots, 6$). They are matched with the combinations of θ_{T_0} and θ_{T_1} :

$$\begin{cases} L_1 = \text{"Formal"}, & \phi_{T_0}^1 \leq \theta_{T_0} \leq \phi_{T_0}^2 \text{ and } |\theta_{T_1}| \leq \phi_{T_1}, \\ L_2 = \text{"Relaxed"}, & \phi_{T_0}^1 \leq \theta_{T_0} \leq \phi_{T_0}^2 \text{ and } |\theta_{T_1}| > \phi_{T_1}, \\ L_3 = \text{"Sad"}, & \theta_{T_0} > \phi_{T_0}^2 \text{ and } |\theta_{T_1}| \leq \phi_{T_1}, \\ L_4 = \text{"Negligent"}, & \theta_{T_0} > \phi_{T_0}^2 \text{ and } |\theta_{T_1}| > \phi_{T_1}, \\ L_5 = \text{"Optimistic"}, & \theta_{T_0} < \phi_{T_0}^1 \text{ and } |\theta_{T_1}| \leq \phi_{T_1}, \\ L_6 = \text{"Provocative"}, & \theta_{T_0} < \phi_{T_0}^1 \text{ and } |\theta_{T_1}| > \phi_{T_1}, \end{cases} \quad (15)$$

where $\phi_{T_0}^1 = -0.035$ rad, $\phi_{T_0}^2 = 0.35$ rad, and $\phi_{T_1} = 0.1$ rad are the given thresholds.

4. INTERACTION STRATEGY

Although human emotion changes continuously, it has obvious gradations. To simplify the issue we categorize human emotion state into five gradations, represented by *VeryNegative* (*Neg-*), *Negative* (*Neg*), *Normal* (*Nor*), *Positive* (*Pos*), and *VeryPositive* (*Pos+*). In this section, we will devise an interaction strategy, where an FSM model is constructed to describe the transition regulations of the human emotion state based on the quantified global movement emotion E_{M_0} and the torso pose emotion label L_i , and then the appropriate interactive behavior is selected according to the inferred human emotion state.

4.1. Transition regulations of the human emotion state

An FSM model is constructed to describe the transition regulations of the human emotion state:

$$FSM_{HE} = (HE, \sum_M, F_M, HE_s), \quad (16)$$

where $HE = \{HE_p \mid p = 1, \dots, 5\}$ is the set of the five gradations of human emotion state, as shown in Fig. 7. $HE_s = HE_3$ is the initial emotion state. $\sum_M = \{e_q \mid p = 1, \dots, 5\}$

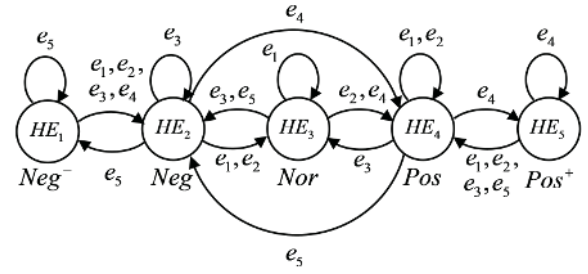


Fig. 7. The FSM of the human emotion states.

Table 3. State-transition functions of the robot emotion states.

Source State	Input	Object State	Source State	Input	Object State
HE ₁	e ₅	HE ₁	HE ₃	e ₂ , e ₄	HE ₄
HE ₁	e ₅ , e ₂ , e ₃ , e ₄	HE ₂	HE ₄	e ₅	HE ₂
HE ₂	e ₅	HE ₁	HE ₄	e ₃	HE ₃
HE ₂	e ₃	HE ₂	HE ₄	e ₁ , e ₂	HE ₄
HE ₂	e ₁ , e ₂	HE ₃	HE ₄	e ₄	HE ₅
HE ₂	e ₄	HE ₄	HE ₅	e ₁ , e ₂ , e ₃ , e ₅	HE ₄
HE ₃	e ₃ , e ₅	HE ₂	HE ₅	e ₄	HE ₅
HE ₃	e ₁	HE ₃			

Table 4. Rules of selecting e_q .

L_i	$E_{M_0} < -E_M^T$	$ E_{M_0} \leq E_M^T$	$ E_{M_0} > E_M^T$
$L_3 \& L_4$	e ₅	e ₃	—
L_1	e ₃	e ₁	e ₂
L_2	e ₁	e ₁	e ₂
L_5	—	e ₂	e ₄
L_6	—	e ₄	e ₄

is the set of input events that determines the orientation of state transition. $F_M: HE \times \sum_M \rightarrow HE$ are the state-transition functions, which finally export the current human emotion state of this FSM. The overall F_M are displayed in Table 3 and Fig. 7. We assign the input events e_q , for $q = 1, \dots, 5$, to match the combinations of incoming E_{M_0} and L_i , as shown in Table 4. E_M^T is a threshold for categorizing the global movement emotion into three extents. The value of E_M^T is decided according to the experiment (Section 5.1). Three combinations of E_{M_0} and L_i appear rarely in daily life, so they are considered to be inconsistent, and do not correspond with any e_q . If these combinations are observed indeed, the current human emotion state will remain the same as the last one.

4.2. Interactive strategy based on possible alternative behaviors

After the current human emotion state is inferred, appropriate interactive behavior including interactive movement and speech can be designed. Based on several

theoretical researches about HRI [30–32], the *safe* and *ethical restrictions* must be well considered when designing the thorough strategies. According to the *safe* and *ethical restrictions*, the interactive behaviors should be limited in a possible alternative range.

Interactive movement: During the HRI, a safe distance between the user and the robot must be ensured. The interactive movements are composed of a series of poses. According to the algorithm proposed in [33], the poses corresponding to each human emotion state can be obtained with the activation-valence value. Based on the available poses, the movements are planned by a Markov model. In order to satisfy the ethical restrictions, these generated movements are checked manually to identify any rude ones in relation to the current cultural background. Finally, several possible alternative movements are arranged to cope with each human emotion state. When the current human emotion state has been inferred, one interactive movement is picked out randomly from the possible alternative movements.

Interactive speech: The interactive speech is determined by the current human emotion state HE and torso pose emotion label L_i together. The possible alternative speeches should be polite and encouraging in all cases, and must not include vulgar language, hate speech, or slang. The speech content guidance is listed in Table 5. There are six inconsistent combinations of HE and L_i . If these combinations are observed indeed, the robot will say nothing in order to reduce misunderstandings.

5. EXPERIMENTAL RESULTS

5.1. Evaluation of the movement emotion quantification

Dataset introduction: We use the public dataset UTD-MHAD [34] to evaluate the movement emotion quantification. The primary reason for using this dataset is that many types of the included actions can match certain emotion types; e.g., “forward lunge” shows a positive emotion, whereas “sit” shows a negative one.

UTD-MHAD contains 27 actions performed by eight subjects. Each subject repeated an action four times.

Based on our assessment, we manually assigned 12 actions a positive (+) or negative (-) emotion label as the ground truth. We assume that these actions are not recognized, and the proposed SoHRI scheme is implemented to analyze E_{M_0} and L_i of the skeleton joint position sequence.

Results of SoHRI: The results are shown in Table 6. \bar{E}_M is the average global movement emotion of one type of action in the dataset, and L_i is the ratio of the torso pose emotion labels of the result. The signs of all \bar{E}_M values correspond to the ground truths, and the values are reasonable. As the actions in UTD-MHAD do not contain exaggerated leaning backward, there are no actions labeled “optimistic” or “provocative”. Most actions are “formal” or “relaxed”, while L_i of the “stand”, “sit”, and “squat” actions have a large ratio of “sad”. Understandably, the subjects bow forward during these actions. In conclusion, the body emotion analysis of SoHRI quantifies E_{M_0} and L_i of each action type precisely.

According to the experimental result, $|E_{M_0}|$ of the movements that express fairly positive or negative emotions are larger than 0.25 when using the parameter settings in Section 3. So $E_M^T = 0.25$ is an available value in our interaction strategy.

Method comparison: To the best of our knowledge, the proposed SoHRI quantifies the human movement emotion for the first time. Although several previous researches have explored how to represent the emotion of the human movement, the main measures are focused on classification. Here we use two methods, the Dynamic Time Wrapping with k-Nearest Neighbors (DTW+kNN) [35] and the Support Vector Machine (SVM) [36], to recognize the movement emotion types of the actions in UTD-MHAD for comparison.

The movement emotion types include “positive” and “negative” that correspond to the ground truth. Three experimental groups are arranged, in which the data to be analyzed are the joint locations, velocities, and accelerations, respectively. As each subject repeated an action four times, four-fold cross-validation is used to estimate the accuracies.

As for DTW+kNN, the number of the nearest neigh-

Table 5. Speech content guidance for SoHRI.

L_i	<i>Neg-</i>	<i>Neg</i>	<i>Nor</i>	<i>Pos</i>	<i>Pos+</i>
L_3 & L_4	encouraging, comforting, ask if the user feels unwell	encouraging, ask if the user wants to rest	caring	–	–
L_1	caring, comforting	caring	polite	polite, hospitable	happy, laughing
L_2	ask if the user feels unwell	ask if the user wants to rest	polite	happy	reminding the user to keep safety
L_5	–	–	happy	laughing, witty	joking, witty
L_6	–	–	happy, witty	joking, witty, remind the user to keep safety	joking, witty, reminding the user to keep safety

Table 6. Speech content guidance for SoHRI.

Action	Ground Truth	\bar{E}_M	$L_i(\%)$			
			Formal	Relaxed	Sad	Negligent
A01: right arm swipe to the left	–	–0.2354	87.50%	12.50%	0%	0%
A02: right arm swipe to the right	+	0.2238	90.63%	9.37%	0%	0%
A03: right hand wave	+	0.1934	75.00%	25.00%	0%	0%
A04: clap hands	+	0.1711	100.00%	0%	0%	0%
A05: two arms curl	+	0.3107	100.00%	0%	0%	0%
A06: two hands push	+	0.3657	93.75%	6.25%	0%	0%
A07: hand catch	–	–0.3526	78.13%	21.87%	0%	0%
A08: jogging	+	0.2745	100.00%	0%	0%	0%
A09: stand	+	0.2706	50.00%	0%	50%	0%
A10: sit	–	–0.2788	12.50%	0%	87.50%	0%
A11: forward lunge	+	0.3344	81.25%	0%	18.75%	0%
A12: squat	–	–0.3237	0%	0%	90.63%	9.37%

Table 7. Accuracies of the emotion analysis by DTW+kNN and SVM on UTD-MHAD.

Action	Ground Truth	Location		Velocity		Acceleration		\bar{E}_M
		DTW+kNN	SVM	DTW+kNN	SVM	DTW+kNN	SVM	
A01	–	60%	97%	85%	94%	50%	88%	–0.2354
A02	+	100%	100%	100%	100%	45%	100%	0.2238
A03	+	60%	97%	85%	94%	55%	91%	0.1934
A04	+	100%	100%	100%	100%	55%	97%	0.1711
A05	+	100%	100%	100%	100%	35%	97%	0.3107
A06	+	100%	100%	100%	100%	50%	91%	0.3657
A07	–	60%	94%	100%	91%	80%	66%	–0.3526
A08	+	95%	100%	100%	100%	35%	100%	0.2745
A09	+	100%	100%	100%	100%	75%	94%	0.2706
A10	–	100%	97%	100%	97%	95%	84%	–0.2788
A11	+	100%	100%	100%	100%	80%	81%	0.3344
A12	–	100%	100%	100%	97%	100%	50%	–0.3237
Average		89.58%	98.75%	97.50%	97.75%	62.92%	86.58%	

bors is set to five, and the accuracy is the proportion of the neighbors whose predicted emotion types match the ground truth. The input data need to be preprocessed as for SVM, including standardization and dimensionality reduction. The parameters of SVM are tuned by grid research. Finally the accuracy is the proportion of actions whose movement emotion types are predicted correctly.

The accuracies of the two methods in the three experiments are shown in Table 7. A01-A12 are the action marks given in Table 6. The accuracies in the acceleration group are much lower than those in the other two groups. The DTW+kNN method achieves the best accuracies in the velocity group, while the SVM method achieves the best accuracies in the location group. In general, the two methods have relatively good performances in the velocity group. This result indicates that velocity can overcome the individual differences among different people. Similarly, all the movement variables used in the body movement anal-

ysis of SoHRI are the velocities.

The global movement emotions analyzed by SoHRI are also listed in the last column of Table 7. All the signs of \bar{E}_M values correspond to the recognition results of DTW+kNN and SVM in the velocity group. The ability of obtaining quantitative emotion value of human movement is the advantage of SoHRI.

5.2. Evaluation of the whole SoHRI scheme

Experiment settings: We wrote a program to implement an HRI system embedded by the proposed SoHRI scheme. The humanoid robot NAO [23] is adopted as the interactive robot. The visual interface of the program is shown in Fig. 8. The human's skeleton, analysis result of the movement emotion, torso pose label, and the robot's speech and appearance can be obtained from the interface.

We used a questionnaire to obtain the subjective evaluation. The questionnaire is given in Fig. 9, which is de-

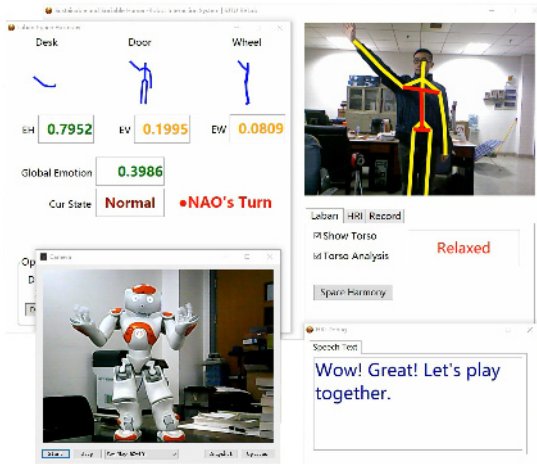


Fig. 8. Visual interface of the SoHRI-embedded system.

SoHRI Questionnaire

Your role: Participant Spectator

Q1. The participant can perform movements with different emotions:
Stiffly 1 2 3 4 5 6 7 Naturally

Q2. The emotion analysis results match your expectation:
Not at all 1 2 3 4 5 6 7 Very much

Q3. The interactive movements of the robot match your expectation:
Not at all 1 2 3 4 5 6 7 Very much

Q4. The interactive speech of the robot matches your expectation:
Not at all 1 2 3 4 5 6 7 Very much

Q5. More comments:

Q6. Your age: _____

Q7. Your gender: Male Female

Fig. 9. Questionnaire for evaluating the whole SoHRI scheme.

signed based on that in [19]. The main questions are the first four, which are presented on a 7-point Likert scale. They cover the evaluation from the user experiences to the multimodal performances of the robot. Besides the participants completed the questionnaire after their experiments, we invited some of them to watch others' experiment replays and then complete the questionnaires again. They evaluated the system as spectators.

The participants included 18 members (eleven males, seven females; nine 18-25 years old, four 26-30 years old, four 30-40 years old, one over 40 years old; twelve students, six in work), and the spectators included seven members (six males, one females; five 18-25 years old, two 26-30 years old; six students, one in work). All the 18 participants used SoHRI for the first time. Before the experiment, we introduced the typical movements specified in LMA to the participants, and showed them the questionnaire to clarify what they needed to observe and evaluate. Then the participants made several movements in front of

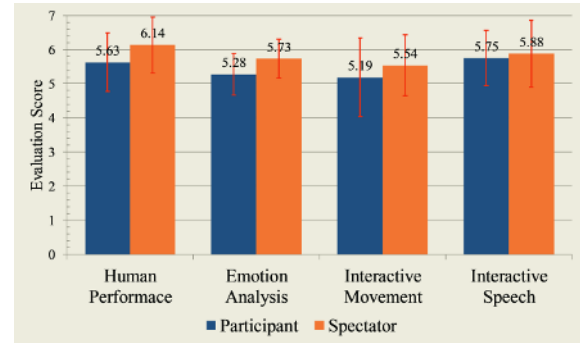


Fig. 10. Questionnaire result from participants and spectators. Red line: standard deviation.

Kinect and NAO. The experiments were recorded by capturing the videos of the robot and the program interface synchronously. The spectators watched these videos afterwards.

Evaluation results: The evaluations from the participants are illustrated by the dark blue bars in Fig. 10. Most participants were satisfied with their performances. They gave themselves 5.63/7 on average. The evaluations of emotion analysis and interactive movement were fairly positive, which are 5.28/7 and 5.19/7, respectively. In addition, the standard deviation line for the interactive movement shows that different participants had inconsistent opinions. The interactive speech received 5.75/7, a relatively favorable score.

The evaluations from the spectators are illustrated by the orange bars in Fig. 10. The performance of the participants is evaluated as 6.14/7 by the spectators on average. The emotion analysis and interactive movement received 5.73/7 and 5.54/7, respectively. The speech still achieved a high score (5.88/7).

Discussion: Based on the evaluations described above, the SoHRI scheme analyzes the body emotion effectively, and the robot reactions are reasonable. The subjects showed great interests in this novel HRI system and had very high expectations.

The evaluations of the interactive movement from the participants are a bit low. This is because the interactive movements are generated automatically, some of which are not natural enough. Furthermore, the interactive speech is outstanding, thus, the interactive movement appears worse in comparison.

During the experiments, we observed that quite a few participants performed stiffly, although we knew that he/she is a lively and expressive person. As our scheme does not serve for traditional movement tracking or action recognition, the participants need to sufficiently understand what they are going to do. This phenomenon indicates that the experiment tutorial should be more clear and encouraging.

The evaluations from the spectators are higher than

those from the participants in general. This fact suggests the users can gradually become familiar with the robot's behaviors based on SoHRI scheme after they interact with the robot many times.

6. CONCLUSION

In this work, a SoHRI scheme based on body emotion analysis was developed to achieve reasonable and natural interaction while human actions were not recognized. The SoHRI scheme thoroughly makes use of the visual movement information, and compensates for the inadequacy of the traditional HRI that neglects the case when actions are not recognized.

Movement emotion quantification and torso pose emotion labeling were proposed to quantify the body emotions included in the dynamic movement and static pose. The interaction strategy was designed, in which an FSM model was constructed to describe the transition regulations of the human emotion state, and then appropriate behavior was selected according to the current human emotion state. The quantification effect of SoHRI was verified using the dataset UTD-MHAD, and the whole SoHRI scheme was tested using questionnaires filled out by the participants and spectators. The experimental results have shown that the SoHRI scheme can analyze the body emotion precisely, and help the robot to make reasonable interactive behaviors.

According to the experimental result, we plan to further develop the SoHRI scheme in the following aspects: improving the fusion method of the global movement emotion; recording certain human movements and make the robot play for a better performance; designing better tutorial to guide the user "warms up" with the robot quickly; and customizing the interaction scheme for individual users. More participants will be invited to try out the SoHRI scheme, and provide valuable feedback for making greater improvement.

REFERENCES

- [1] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971-981, June 2013.
- [2] M. M. Ullah and I. Laptev, "Actlets: A novel local representation for human action recognition in video," *Proc. of 19th IEEE International Conference on Image Processing*, pp. 777-780, 2012.
- [3] F. Alonso Martín, A. Ramey, and M. A. Salichs, "Speaker identification using three signal voice domains during human-robot interaction," *Proc. of the ACM/IEEE International Conference on Human-robot Interaction*, pp. 114-115, 2014.
- [4] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *Expert Systems with Applications*, vol. 41, no. 3, pp. 786-794, February 2014.
- [5] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3D human action recognition," *Human Action Recognition with Depth Cameras*, Springer, pp. 11-40, January 2014.
- [6] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of Real-time Image Processing*, vol. 12, no. 1, pp. 155-163, June 2016.
- [7] V. Venkataraman, P. Turaga, N. Lehrer, M. Baran, T. Rikakis, and S. L. Wolf, "Attractor-shape for dynamical analysis of human movement: applications in stroke rehabilitation and action recognition," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 514-520, 2013.
- [8] F. G. Da Silva, and E. Galeazzo, "Accelerometer based intelligent system for human movement recognition," *Proc. of 5th IEEE International Workshop on Advances in Sensors and Interfaces (IWASI)*, pp. 20-24, 2013.
- [9] M. H. Siddiqi, R. Ali, A. M. Khan, Y. T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1386-1398, February 2015.
- [10] I. B. Yildiz, K. Von Kriegstein, and S. J. Kiebel, "From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems," *PLoS Comput Biol*, vol. 9, no. 9, pp. e1003219, September 2013.
- [11] M. Chatterjee and S.-C. Peng, "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," *Hearing Research*, vol. 235, no. 1, pp. 143-156, January 2008.
- [12] M. Lichtenstern, M. Frassl, B. Perun, and M. Angermann, "A prototyping environment for interaction between a human and a robotic multi-agent system," *Proc. of 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 185-186, 2012.
- [13] T. Yamada, S. Murata, H. Arie, and T. Ogata, "Dynamical Integration of Language and Behavior in a Recurrent Neural Network for Human-Robot Interaction," *Frontiers in Neurorobotics*, vol. 10, no. 11, pp. 6014-17, July 2016.
- [14] M. Farhad, S. N. Hossain, A. S. Khan, and A. Islam, "An efficient optical character recognition algorithm using artificial neural network by curvature properties of characters," *Proc. of International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 1-5, 2014.
- [15] R. Palm, R. Chadalavada, and A. Lilienthal, "Fuzzy modeling and control for intention recognition in human-robot systems," *Proc. of 8th International Conference on Computational Intelligence IJCCI 2016, FCTA*, Porto, Portugal, pp. 67-74, 2016.
- [16] C. R. Guerrero, J. C. F. Marinero, J. P. Turiel, and V. Muñoz, "Using 'human state aware' robots to enhance physical human-robot interaction in a cooperative scenario," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 2, pp. 250-259, November 2013.

- [17] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Data-driven HRI: learning social behaviors by example from human-human interaction," *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 988-1008, August 2016.
- [18] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 2-9, 2014.
- [19] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, pp. 325-332, 2013.
- [20] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW'08*, pp. 1-6, 2008.
- [21] Z. Liu, M. Wu, D. Li, L. Chen, F. Dong, Y. Yamazaki, and K. Hirota, "Communication atmosphere in humans and robots interaction based on the concept of fuzzy atmosphere generated by emotional states of humans and robots," *Journal of Automation Mobile Robotics and Intelligent Systems*, vol. 7, no. 2, pp. 52-63, June 2013.
- [22] W. H. Kim, J. W. Park, W. H. Lee, H. S. Lee, and M. J. Chung, "LMA based emotional motion representation using RGB-D camera," *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, pp. 163-164, 2013.
- [23] A. Robotics, "Nao robot: characteristics - Aldebaran," <https://www.aldebaran.com/en/cool-robots/nao/find-out-more-about-nao>.
- [24] R. Laban, *The Language of Movement: A Guidebook to Choreutics*, Plays Inc, Boston, 1974.
- [25] Y. Cheng, *A Study on Semantic and Emotional Messages in Robot Movements*, Department of Multimedia Design, National Taichung Institute of Technology, Taichung, 2010.
- [26] Y. Juan, *Motion Style Synthesis Based on Laban Movement Analysis*, Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, 2004.
- [27] C. Hsieh and Y. Wang, "Digitalize emotions to improve the quality life-analyzing movement for emotion application," *Journal of Aesthetic Education*, vol. 168, pp. 64-69, 2009.
- [28] M. S. Ku and Y. Chen, "From movement to emotion - a basic research of upper body (analysis foundation of body movement in the digital world 3 of 3)," *Journal of Aesthetic Education*, vol. 164, pp. 38-43, 2008.
- [29] R. C. Gonzalez and R. E. Wood, "Using fuzzy techniques for intensity," *Digital Image Processing*, 3 ed., Prentice Hall, pp. 128, 2008.
- [30] I. Asimov, "Runaround," *Astounding Science Fiction*, vol. 29, no. 1, pp. 94-103, March 1942.
- [31] E. Fosch Villaronga, A. Barco, B. Zcan, and J. Shukla, "An interdisciplinary approach to improving cognitive human-robot interaction-a novel emotion-based model," *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016*, pp. 195-205, October 2016.
- [32] M. Giuliani, C. Lenz, T. Müller, M. Rickert, and A. Knoll, "Design principles for safety in human-robot interaction," *International Journal of Social Robotics*, vol. 2, no. 3, pp. 253-274, March 2010.
- [33] G. Xia, J. Tay, R. Dannenberg, and M. Veloso, "Autonomous robot dancing driven by beats and emotions of music," *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 205-212, 2012.
- [34] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 168-172, 2015.
- [35] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," *Proc. of International Conference on Affective Computing and Intelligent Interaction*, pp. 71-82, 2007.
- [36] B. Kikhia, M. Gomez, L. L. Jiménez, J. Hallberg, N. Karvonen, and K. Synnes, "Analyzing body movements within the laban effort framework using a single accelerometer," *Sensors*, vol. 14, no. 3, pp. 5725-5741, March 2014.



Tehao Zhu received the B.S. degree in automation from the Northwest Polytechnical University, Xi'an, China, in 2009, and the M.S. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 2012. He is currently pursuing a Ph.D. degree at Shanghai Jiao Tong University, Shanghai, China. His current research interests include human-robot interaction, machine learning, and image processing.



Zeyang Xia received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2002, and the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2008. He is currently a Professor at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, and is the director of Medical Robotics and Biomechanics Laboratory (<http://www.bigsmilelab.ac.cn>). His research interests include biped humanoid robotics, medical robotics, and dental biomechanics. He has published over 80 peer reviewed papers, and applied over 40 patents. He is the vice chairman of Guangzhou Branch of the Youth Innovation Promotion Association, Chinese Academy of Sciences, and the co-chair of Guangdong Chapter of IEEE Robotics and Automation Society. He served as the Program Co-Chair of IEEE RCAR 2016 and ICVS 2017, and will be the General Chair of IEEE RCAR 2019.



Jiaqi Dong received the B.S. degree in automation from Shanghai Jiao Tong University, Shanghai, China, in 2014. She is currently pursuing a Ph.D. degree at Shanghai Jiao Tong University, Shanghai, China. Her current research interests include human-robot interaction and pattern recognition.



Qunfei Zhao received the B.S.E.E. degree from Xi'an Jiao Tong University, Xi'an, China, in 1982, and the Sc.D. degree in system science from Tokyo Institute of Technology, Tokyo, Japan, in 1988. He is currently a Professor at the School of Electronic Information and Electric Engineering, Shanghai Jiao Tong University, China. His research interests include robotics, machine vision, and optimal control of complex mechatronic systems.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.