Robotica (2017) volume 35, pp. 2122–2138. © Cambridge University Press 2016
doi:10.1017/S0263574716000722

A gaze control of socially interactive robots in multiple-person interaction

Sang-Seok Yun*

Korea Institute of Science and Technology (KIST), Seoul, Korea

(Accepted September 10, 2016. First published online: October 3, 2016)

SUMMARY

This paper proposes a computational model for selecting a suitable interlocutor of socially interactive robots in a situation interacting with multiple persons. To support this, a hybrid approach incorporating gaze control criteria and perceptual measurements for social cues is applied to the robot. For the perception part, representative non-verbal behaviors indicating human-interaction intent are designed based on the psychological analysis of human–human interaction, and these behavioral features are quantitatively measured by core perceptual components including visual, auditory, and spatial modalities. In addition, each aspect of recognition performance is improved through temporal confidence reasoning as a post-processing step. On the other hand, two factors of the physical space and conversational intimacy are tactically applied to the model calculation as a way of strengthening social gaze control effect of the robot. Interaction experiments with performance evaluation are given to verify that the proposed model is suitable to assess intended behaviors of individuals and perform gaze behavior about multiple persons. By showing a success rate of 93.3% in human decision-making criteria, it confirms a potential to establish socially acceptable gaze control in multiple-person interaction.

KEYWORDS: Gaze control, Human–robot interaction (HRI), Non-verbal behavior, Robot attention, Multi-party, Socially interactive robots.

1. Introduction

The psychology of human social interaction is such that people actually have the ability to identify thoughts, feelings, attitudes, desires, and intentions from the behavior of other people, and these human behaviors provide important cognition cues in social interaction.[1] In addition, a human understands the other party's mental determinations through interpretation of various forms of non-verbal behaviors during communication.[2] Similar to humans, to show '*human social*' attributes, social robots are actually required to have the following elements: intelligent communication, emotional exchange, learning social competences and models of other agents, distinctive personalities and characteristics, and performance of natural expressions using gaze, gestures, etc.[3] As a prerequisite for representing a socially acceptable manner in many such elements, the robot must be able to understand semantics of human social cues in a given context and cope with the situation adequately through interaction strategies in the same way a human would. In particular, as one of the most salient signals in human–robot interaction (HRI), robot gaze can be an excellent predictor of conversational attention[4] and support communicative mechanisms such as turn-taking, synchronized speech, and suitable engagement actions.[5] Furthermore, gaze decision process can be used to trigger attention shifting according to the level of interest or shape the participant roles of conversational partner in social interaction involving multiple-person interactions (MPIs).[6,7] Accordingly, gaze control of the robot (e.g., select a pertinent interlocutor) can vary widely depending on how well interaction strategy is composed,[8–10] and it has the advantage to establish a close rapport and strengthen positive relationships between humans and a robot on the interaction practice (see Fig. 1).

Present address: Hwarangno 14-gil 5, Seongbuk-gu, Seoul 02792, Republic of Korea.
* Corresponding author. E-mail: yssmecha@gmail.com

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

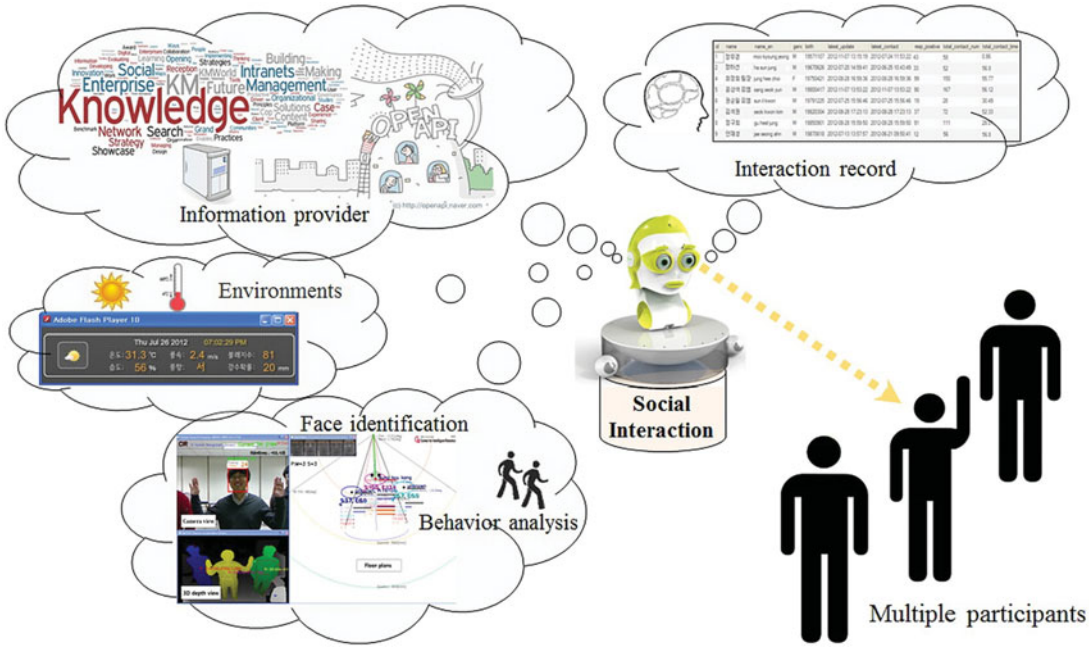


Fig. 1. A social gaze control in multiple-person interactions.

However, the issue is that the research has not been made sufficiently where the robot can perform sensible gaze decision for multiple persons by recognizing the various verbal and non-verbal behaviors and act a rational manner to participants in MPIs through gaze behavior.

Main interest of this paper is to propose a reliable gaze control model that performs gaze behavior effectively in the social interaction with multiple persons. To support this, this paper first classifies representative non-verbal behaviors based on the cognitive science research and psychological analysis of human–human interactions. In addition, to understand semantics of human intimate behaviors that happen to the robot in a multi-person setting, this paper places emphasis on the reliable processing of multimodal signals by incorporating several core perceptual features including visual, auditory, and spatial modalities. Along with the recognition module, the post-processing of temporal confidence reasoning (TCR)[11] is applied on the model so as to support the improvement of recognition performance. This paper then adopts the special criteria with physical space and communicative intimacy in order to act a humanlike way in the interaction design. In this regard, the robot easily identify the person who displays the greatest level of interaction intent by conducting socially acceptable decision-making and explicitly express attention to the designated target in a given situation by gaze control.

The rest of this paper is organized as follows. Section 2 introduces the related work for the study. Section 3 explains the robot gaze strategy for multiple persons by using a hybrid approach. Section 4 details the computational modeling for gaze decision in each part. Section 5 describes the experimental results of recognition performance and participants' evaluations. Finally, the concluding remarks are made in Section 6.

2. Related Work

For recognizing human behaviors, multi-sensor integration has the potential to improve the usability and accessibility of communication flow by complementary relations. For instance, Okuno *et al.*[12] announced dialog control to communicate with people by speech synthesis and visual information from talking persons, and Trafton *et al.*[13] introduced conversational attention system to track the person who is talking by integrating visual and auditory information in a cognitive architecture. Lang *et al.*[14] and Fritsch *et al.*[15] also showed a schematic example for tracking a speaker through anchoring component symbols to their associated percepts with sound source localization (SSL),

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

face recognition, and leg detection. However, the presented above methods were mainly focused on finding a speaker without sufficient consideration of various social signals.

In other cases, Michalowski *et al.*[16] accomplished to detect a person who is interested to the robot by the spatial formation so as to perform receptionist tasks interactively and Kondo *et al.*[10] showed how to design of multi-party interaction with the level of interaction engagement by using natural body gesturing and speech dialogue. In addition, Bohus and Horvitz[17] proposed computational methods for situated, multiparty, open-world dialog, which is required for detecting whether engaged participants are actively maintaining the conversational engagement or not. Nevertheless, its capability to determine the interlocutor is limited to only two areas—face detection and sound localization—which further curtails its ability to interact actively and dynamically. In measuring human social cues, human identification can be very important as a way to perform user-oriented tasks and other advanced interaction technologies. Gockely *et al.*[18] integrated laser range measurement and card swipe for estimating people's locations and identification in predefined zones. Bellotto *et al.*[19] also fused laser and visual data, including body feature, face, and clothes, to estimate position and identity of the people in the dynamic environment. However, swiping a card offers the discomfort to re-visitors in the HRI design, and facial recognition in the second study is solely used to identifying visitors in order to provide more information about the university without interaction scheme for multiple persons. As stated above, the importance of multimodal cues is obvious. Nonetheless, the relative contribution of each modality in terms of efficacy is still less clear. For this part, Torta *et al.*[20] introduce that a robot can conduct multiple non-verbal communication cues to attract a person's attention. Through two experiments they investigated that a robot could attract attention more effectively when using actions that combine cues with verbal and auditory actions. Unfortunately, the experimental result only draw a comparison between unimodal and multimodal stimuli targeting a single user. In addition, recognition capability of multimodal cues is not considered, even if it is of great importance for a robot to sense human behaviors in terms of interaction intent.

Along with the perception ability, the use of gaze control has been studied in the embodied conversational agent to engage in a meaningful interaction with the targeted person among participants. For example, Kirckner *et al.*[21] simply showed robot's gaze effect in the group by randomly selecting a group member without being aware of social context during a robot-group interaction, Mtulu *et al.*[22] developed conversational gaze mechanisms for humanlike robots as one of the social members, and Zaraki *et al.*[23] presented the gaze-control system capable of humanlike gaze behavior by using a humanoid robot that performs the gaze control for the most prominent target on the basis of visual feature analysis for social cues. However, although the system generates appropriate robot's attention in a multi-person setting, it does not mean the fact that the system can be successful in the social contexts of showing the attentive listening according to the communicative intimacy or expressing the interaction intent such as calling a robot's name. As another considering factor that affects robot gaze, Nakano *et al.*[24] introduced a regression-based model to show how gaze behaviors are affected by conversational dominance and participation roles in multiparty conversations. By using speech and gaze information in an empirical study, they insist that attentional behavior based on dominance estimation could make agents more socially effective in a given user group. However, they did not show gaze management capability capable of sensing user intention of engagement in multiparty conversations due to a Wizard-of-Oz setting.

In order to face this issue, this paper examines social gaze behavior of the robot in the MPIs by incorporating of multimodal perceptual features from the interaction design with gaze control criteria.

3. Robot Gaze Strategy for Multiple Persons

In the MPI design, we propose a hybrid approach that performs socially acceptable gaze control in selecting an interaction partner. As a top-down approach, differentiated interest level for all participants can be based on special criteria such as personal intimacy for the identified person, positive tendency, prior interaction experiences, physical attractiveness, environmental factors, and even behavioral events.[8,25,26] The application of these rules can be seen as expressing a proper social manner to participants in a given situation and it leads to conversational attention to participants depending on the situation.[15] As our interaction design regarding to these factors, the physical space, which is classified as personal, social, and public zones surrounding a robot according to the interpersonal distance, proposed by Hall,[27] mentioned by Patterson,[28] and applied to the HRI

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

Table I. Dominant features based on non-verbal behavior factors.

Behavior	Feature	Intention attribute
Oculesics	Eye movement x_e	Gazing
Kinesics	Facial expression x_f	Smiling
	Body gesture x_g	Waving
	Movement x_m	Moving forward
Proxemics	Proximity x_p	Distance
	Body orientation x_o	Confronting
	Body leaning x_l	Leaning forward
Vocalics	User call x_v	Call robot's name

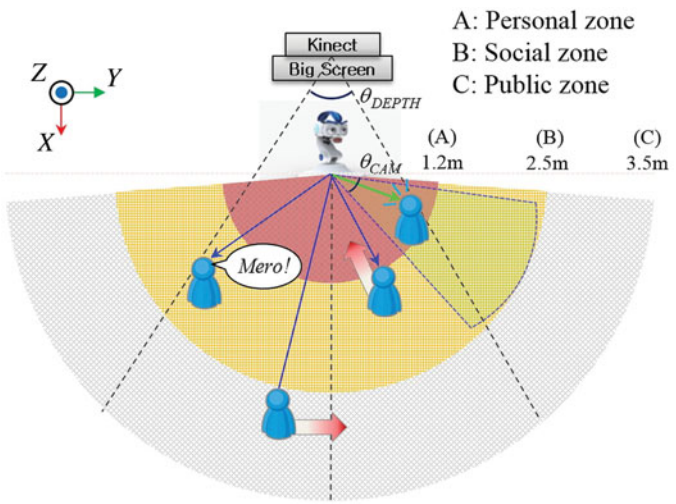


Fig. 2. Interaction zones between a robot and multiple participants according to the physical space and communicative intimacy.

by Walters,[29] was utilized to set the priorities for multiple persons. To be more specific, within the personal zone, the robot preferentially conducts intensive interaction for the private episodes to the identified person. In the social zone, it tactically shifts its gaze toward the participant who expresses the most positive behaviors in order to interact with. And then, the public zone includes the space required for detecting passersby and inducing them to close in with the robot. Another factor that this paper proposes is the communicative intimacy associated with the extent of positive response in social interaction with the targeted person. It is intended to give in ways that the robot interacts with the interlocutor in the personal zone as much as possible except the response for momentary events. If the interlocutor does not indicate any positive response during interaction, the robot can be more likely to shift its gaze to other target in order to attract the attention.[30,31] It would be vital to fulfilling proper social manners effectively (see Fig. 2).

In contrast, as a bottom-up approach, the robot evaluates individual interaction cues through the perceptual processing in order to determine a gaze target in MPIs. The non-verbal factors in social science, discussed by Mehrabian, are leading to a conclusion about a human's emotional state because total liking consist 7% of verbal liking, 38% of vocal liking, and 55% of facial liking.[32] In addition, human manner or behavior in psychological aspects can be divided into five categories with regards to non-verbal intimacy exchange between individuals: oculesic, kinesic, proxemic, vocalic, and tactile behavior.[33,34] Each behavior has dominant features that are commonly observed in human–human non-verbal communication, and it can be an essential element capable of a better understanding of human-interaction intent in the HRI.

From the four behavioral characteristics, this paper defined eight dominant features to inquire as to the salient behavioral cues for analyzing the level of human-interaction intent as shown in Table I. Each feature represents the perceptual information extracted by the multiple sensors of

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

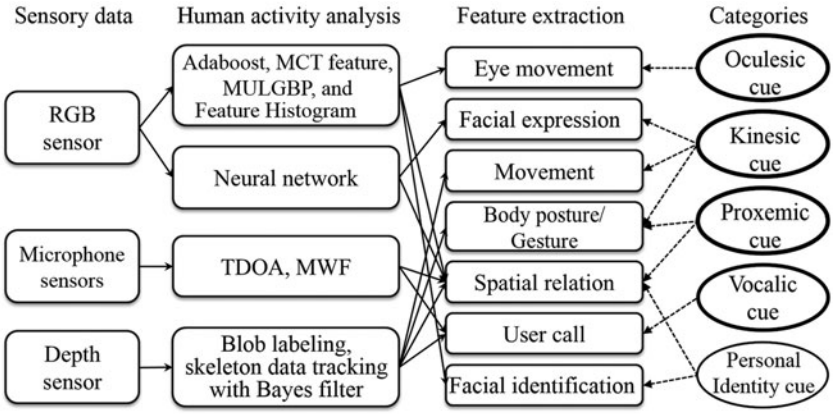


Fig. 3. Feature extraction process of intimacy intent from multimodal sensors.

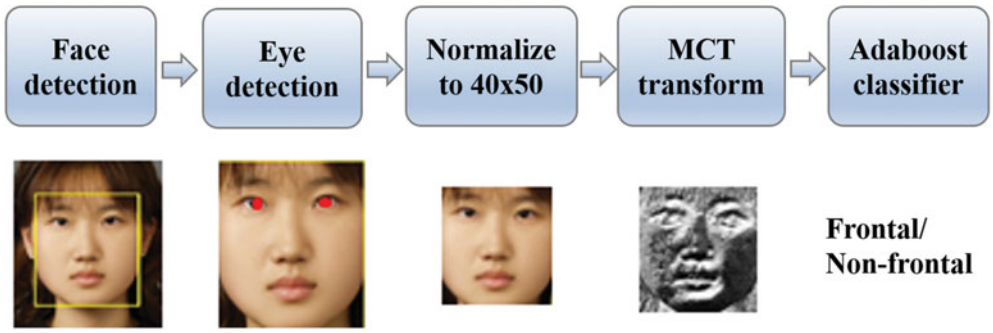


Fig. 4. Overall structure of the frontal face classifier.

visual, auditory, and spatial modality, and a quantitative appraisal for each feature is used to the foundation for performing a gaze control of the robot targeted at multiple people.

4. Computational Modeling for Gaze Decision

4.1. Dominant feature extraction

In the robot system, the proposed multimodal sensors are used to extract dominant features with the four behavioral characteristics, and each attached sensor provides integral capacities that enable robust perceptual function, as shown in Fig. 3.

Each feature extraction is as follow.

4.1.1. Oculesic cue. This is measured by averaging the mutual gazing time of each eye movement between a human and a robot during an interval. However, performing human eye tracking or eye contact is not actually easy in a wide and dynamic environment. Instead, a frontal face classifier, which was proposed in Yoon and Kim,[35] is utilized in the oculesic feature extraction. As shown in Fig. 4, it has the advantage of real-time face classification which consists of five steps including modified census transform (MCT) feature transform and the AdabBoost classifier. The classifier has been trained by 68,000 samples of frontal and non-frontal face to determine whether the face is frontal or not. Accordingly, the oculesic cue can be estimated from the gazing time of a subject while a robot looks at opponents as follows:

$$x_e = \frac{1}{T}\sum_{t=1}^{T} e_t(s), \tag{1}$$

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

where $e(s) \in e_{\text{non-gaze}}$, e_{gaze} represents the binary state provided by frontal face detection and T is the perception time.

4.1.2. Kinesic cues. These are measured by recognizing facial expression, body gesture, and human movement. Facial expression recognition can first classify the current emotional state of people by pre-processing a facial image with a neural network, and on that basis obtain the degree of similarity of facial expressions $e(s) \in e_{\text{ne}}, e_{\text{sm}}, e_{\text{su}}, e_{\text{an}}$ with a neutral, smiling, surprised, and angry state through each reliability value. To normalize this value between 0 and 1, smiling adds a positive value and the rest of the expressions add zero values in the feature extraction process.

$$I_P \in \begin{Bmatrix} e_{\text{sm}}, & \text{if } \arg\max_{\text{s}} e(s) = e_{sm} \\ 0, & \text{otherwise} \end{Bmatrix}, \tag{2}$$

$$x_f = \frac{1}{T} \sum_{t=1}^{T} f_t(I_p), \tag{3}$$

where $f(I_p)$ represents the positive value in the normalized condition.

Body posture and actions for multiple people are recognized by using each moving object's location and 15 elements of skeleton position data from the labeled 3D depth blobs of the Kinect sensor, which is provided by the OpenNI library.[36] By interpreting the linking vectors between skeleton positions and the velocity of each skeleton point, the body gestures and human movement are distinguished by whether to wave hands or come closer to the robot. And then, these can be analyzed stochastically by determining actions at some point in time based on Bayes' theorem as follows:

$$p(z\,|g_i) = \begin{Bmatrix} g_{\text{wave}}, & \text{if} l_{\text{hand_z}} > 0, v_{\text{hand}} \geq v_{\text{th_hand}} \\ g_{\text{stand}}, & \text{otherwise} \end{Bmatrix}, \tag{4}$$

$$x_g = \frac{\sum_{t=1}^{T} g_t(s)}{T}, \quad \text{g}(s) \in \begin{Bmatrix} 1, \text{if } \arg\max_{\text{g}} p(g\,|z) = g_{\text{wave}} \\ 0, \text{otherwise} \end{Bmatrix}, \tag{5}$$

where $l_{\text{hand.z}}$ is the position vector pointing from the elbow to the hand on the Z-axis, v_{hand} is the velocity of the hand on the Y-axis, t is the sampling duration, $v_{\text{th_hand}}$ is the threshold value of hand velocity, g_{wave} is the waving state, and g_{stand} is the non-waving state.

Moving state is classified as the sequence of human movement, such as $m(s) \in m_{\text{fw}}, m_{\text{bw}}, m_{\text{pass}}, m_{\text{stop}}$, based on the robot location. From the probability of each moving state, a movement feature indicating intimate intention attribute can be calculated as follows:

$$p(z\,|m_i) = \begin{Bmatrix} m_{\text{fw}}, & \text{if} v_m > v_{\text{th_torso}}, \theta_m \leq \left|\theta_{th_fw}\right| \\ m_{\text{nfw}}, & \text{otherwise} \end{Bmatrix}, \tag{6}$$

$$x_m = \frac{\sum_{t=1}^{T} m_t(s)}{T}, \quad m(s) \in \begin{Bmatrix} 1, \text{if } \arg\max_{\text{m}} p(m\,|z) = m_{\text{fw}} \\ 0, \text{otherwise} \end{Bmatrix}, \tag{7}$$

where v_m is the current moving velocity of human body position in the XY plane; $v_{\text{th_torso}}$ is the threshold value for moving velocity; θ_m is the moving direction; θ_{th_fw} is the threshold value at the boundary of moving forward; and m_{fw} is the forwarding state and the rest are the non-forwarding state (see Fig. 5).

4.1.3. Proxemic cues. These are easily measured by averaging the spatial distance between a human and a robot, body orientation, and the angle of body leaning. To implement an appropriate spatial relationship according to the physical distance proposed in Hall,[27] personal ($\leq$ 1.2m), social ($\leq$ 2.5m), and public ($\leq$ 3.5m) zones are assigned a Gaussian-typed different intention value which exponentially decreases in the social zone and gradually decreases in each personal and public zone

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

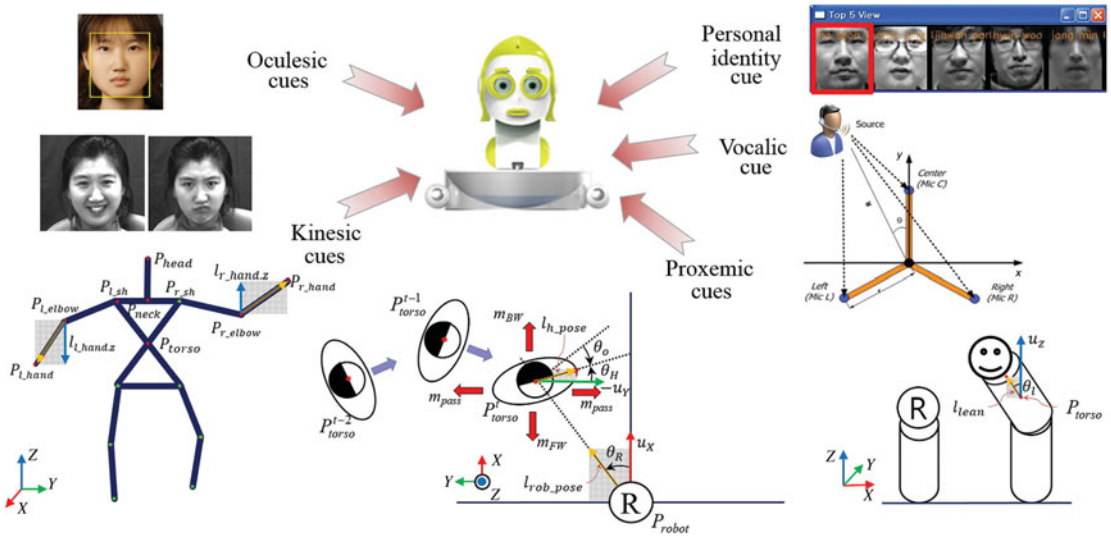


Fig. 5. A configuration of multimodal cues for five dominant features representing intimate intention attributes.

according to the radius. The average feature value for the perception time T is given by

$$x_p = \frac{1}{T}\sum_{t=1}^{T} \exp\left\{\frac{-(d_p(t))^2}{2{\sigma_r}^2}\right\}, \quad d_p = \sqrt{l^2_{\text{neck}.x} + l^2_{\text{neck}.y}}, \tag{8}$$

where d_p is the spatial distance between a robot and an individual person, and σ_d is the covariance of the distance.

The body orientation and the angle of body leaning can also be estimated by tracking the 3D skeleton point vector of the neck P_{neck}, torso P_{torso}, and each shoulder P_{l_sh}.

$$\theta_R = \cos^{-1}\left(\frac{l_{\text{rob_pose}} \times u_X}{\|l_{\text{rob_pose}}\| \ \|\text{u}_X\|}\right), \tag{9}$$

$$\theta_H = \cos^{-1}\left(\frac{l_{h_\text{pose}} \times (-u_Y)}{\|l_{h_\text{pose}}\| \ \|-\text{u}_Y\|}\right), \tag{10}$$

$$x_o = \frac{1}{T}\sum_{t=1}^{T} \exp\left\{\frac{-(\theta_o(t))^2}{2{\sigma_o}^2}\right\}, \quad \theta_o = \theta_R - \theta_H, \tag{11}$$

where θ_H represents the angle of a user's position from the robot along the X-axis, θ_B represents the user's body orientation along the Y-axis, θ_o is the confronting angle of the body plane, and σ_o is the covariance of the confronting angle.

$$x_l = \frac{1}{T}\sum_{t=1}^{T} \exp\left\{\frac{-(\theta_l(t) - \mu_l)^2}{2{\sigma_l}^2}\right\}, \quad \theta_o = \cos^{-1}\left(\frac{l_{\text{lean}} \cdot u_Z}{\|l_{\text{lean}}\| \ \|\text{u}_Z\|}\right), \tag{12}$$

where θ_l is the forward leaning angle based on the location of the robot body, μ_l is the average forward leaning angle, and σ_l is the covariance of the forward leaning angle.

4.1.4. Vocalic cue. This is measured by estimating the direction of a speaker producing a vocal tone. The auditory system consists of the SSL and a multichannel Wiener filter (MWF) with four microphones. The SSL is an algorithm which estimates the direction of a speaking person using time delay of arrival (TDOA) among the microphones.[37] DSP-based embedded equipment is also

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

implemented in the interactive robot, and the MWF and SSL are incorporated in this DSP system. To detect sound direction reliably, post-process using a histogram is applied to achieve robust sound localization. Thus, it is reasonable to assume that a designated person is a speaker if a sound signal (s_{TDOA}) is detected in the azimuth of that user's location.

$$v(s) \in \begin{Bmatrix} 1, \text{ if} s_{\text{TDOA}} \geq s_{th}, \ |\theta_{\text{torso}} - \theta_{snd}| \leq \varepsilon_a \\ 0, \text{ otherwise} \end{Bmatrix}, \tag{13}$$

$$x_v = \sum_{t}^{T} v_t(s), \tag{14}$$

where θ_{torso} is the angle of the user relative to the robot, s_{th} is the threshold value of a TDOA signal, θ_{snd} is the sound azimuth, and ε_a is the sensor tolerance.

4.1.5. Identity cue. In a variety of recognition algorithms, identity features are measured using face recognition, similar to the approach humans take. In addition, the use of this cue aims to avoid the simplicity of robot's gaze decision in MPIs and attracting the participation intent of participants by providing relational discourse for the identified interlocutor through pre-entered dialog scripts. To make a unique ID—such as a person's name—for each human model, this paper applied illumination-robust facial recognition using the statistical and non-statistical methods proposed in Jun *et al*.[38] The recognizer identified the participant accurately 86% of the time in spite of image instability caused by the robot shifting its head to track the target.

4.2. Temporal confidence reasoning

Despite efforts to measure human-interaction intent, some features may be extracted erroneously due to imperfect sensors. Accordingly, all employed logical inference methods must produce robust results even under conditions of imperfect sensing data. Because of its ability to ascertain semantic knowledge despite imperfect measurements, TCR is applied to stochastically exclude misrecognized features. It represents consecutive relationships between time intervals and statistical reasoning by estimating recognition confidence results in post-processing. In addition, this approach has been used to instantiate, update, and remove object information by determining the confidence level with which convergence can be assured from consecutive recognition results. Following Lim,[11] TCR can be statistically obtained from each average recognition rate, assuming that the interval-counter is given by experts beforehand. The interval-counter for each feature recognizer is defined following the *confidence law of inertia*. If the recognition rate of one feature A is r_A, $(1-r_A)$ is the probability that the recognition data for feature A can be false. From that, $(1 - r_A)^{\tau_A}$can be calculated to define the probability when the values of τ_{A} consecutive data are all false. If the result of $(1 - r_A)^{\tau_A}$is less than 5%, then it can be said that the data lie within 95% (1.96σ, $P = 0.05$) of the confidence level. Thus, the interval-counter can be represented as follows:

$$\tau_A = \min\{\tau \in I | (1 - r_A)^{\tau} \leq P\}, \tag{15}$$

where $P(=0.05)$ represents the confidence level.

4.3. Computational model for interlocutor selection

The weighted sum method, which is the best known and simplest multi-criteria decision analysis,[39] is used to evaluate human-interaction intent quantitatively. Individual behavioral features to satisfy the confidence level by logical thinking are first converted to an average value of the normalized measurement during the perception time. Then, the final quantitative value comes from considering weighting factor for each feature value according to their importance in situations or frequency count, which is given by experts from prior-experience.

$$\alpha = \sum_{i=1}^{N} \omega_i x_i(s), \tag{16}$$

where α is the quantitative appraisal value of each participant, and ω is the weighting factor.

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

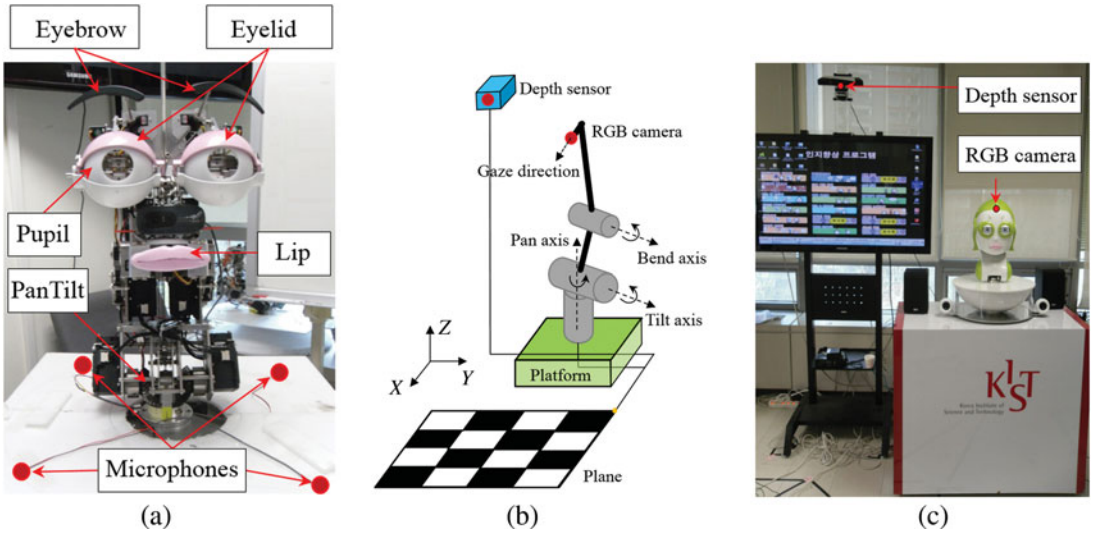


Fig. 6. System configuration of an interactive robot: (a) A humanoid head mechanism, (b) Global coordinate systems, and (c) Test environments for multi-party interaction.

To achieve attentive listening by gaze control, the factor of communicative intimacy was configured to depend on the number of interlocutors' responses, which indicate liking and disliking by providing relational questions. If the jth participant (p_j) engages in dialogue with the robot, the communicative intimacy of the interlocutor L_j is calculated as follows:

$$L_j = \sum_{m=1}^{M} \frac{\rho_m}{\lambda_m} \cdot \text{ind}(p_j, e_m), \tag{17}$$

where ρ_m is the number of positive responses and λ_m is the total number of interactions for the mth registrant in enrollment; and $\text{ind}(p_j, e_m)$ is an indicator function returns one if p_j is equal to the mth person (e_m), otherwise it returns zero.

In MPIs, the designated interlocutor is achieved through the selection of the maximum appraisal value for each participant as follows:

$$E_T = \max[pri(\alpha_1, L_1), pri(\alpha_2, L_2), \cdots, pri(\alpha_N, L_N)], \tag{18}$$

where E_T is the selected interlocutor in all participants, $pri(.)$ is the sum value of individuals based on weighted combination of dominant feature appraisal and communicative intimacy level, and N is the number of participants.

4.4. Robot system configuration

The robot's gaze decision through the computational model is implemented with an interactive robot platform, MERO. As a stationary robotic head type, MERO is mainly designed by product designers based on what a sample target group of people perceived to be an aesthetically pleasing, human-friendly external appearance. Figure 6 shows the configuration of the socially interactive robot, MERO.

A webcam on the robot head, four microphones on the base, and a Kinect sensor located at a certain distance from the robot along the Z-axis are attached on the robot system in order to have a reliable perception, and the robot is equipped to perform human tracking, shifts in gaze, head gestures, and facial expressions through 4-DOF neck and 17-DOF face modules (eyebrow, eyelid, pupil, and lip). In addition, the database for facial registration and the user profiles are built for all participants in advance. Dialog contents of the robot is entirely automated using a finite number of pre-scripted stories including each registered participant where it conducts linguistic utterances

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

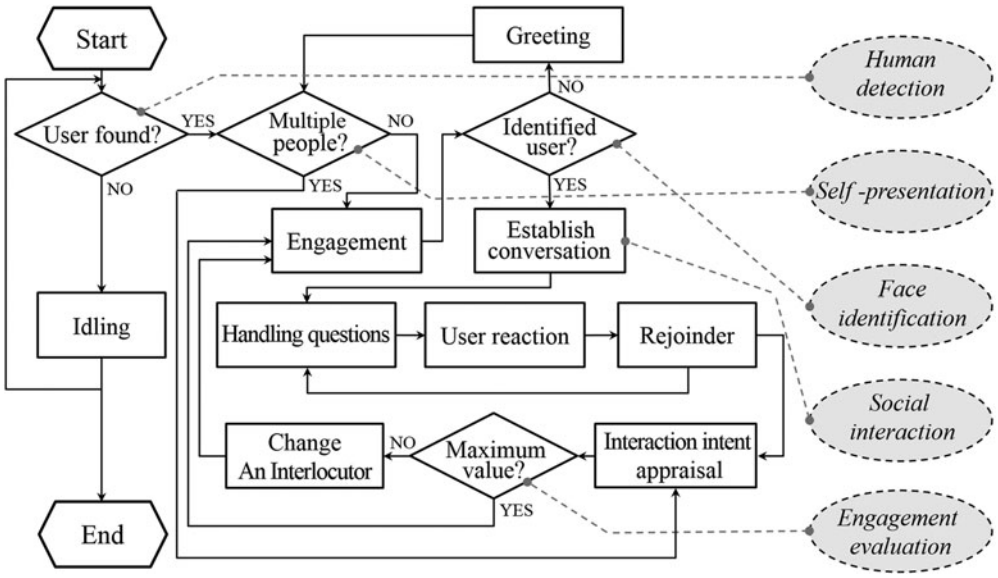


Fig. 7. Gaze decision process of the robot in multi-party interactions.

from the pre-entered 600 dialog scripts and 100 interjections via text-to-speech, lip-sync, and voice synthesizer modules.[40]

From this configuration, gaze control process for multiple persons is performed in five steps: (a) human detection, (b) self-presentation, (c) face identification, (d) social interaction, and (e) interlocutor selection (see Fig. 7).

The robot is put on standby until finding an interaction target arbitrarily. Once people appear, the robot checks the number of participants and the dominant features for interlocutor selection. And then, it greets the interlocutor so as to check face registration while expressing intimate behavior. After identifying each participant, the robot asks questions through pre-entered relational dialog script, which the interlocutor may be potentially interested in, and copes with the interlocutor's reaction. For the enrolled person, questions including calling individual's name are based on the user profile and previous visit recorded in the system. During the question and answer time, the robot quantitatively evaluates all participants' interaction intent through a comparison of the accumulated appraisal values and determine whether to continue social interaction with the interlocutor by taking the person who has the highest score.

5. Performance Evaluation

A series of experiments were designed to show how much the gaze control model with perception modules works in the interaction situation with multiple people. In the following experiments, the individual and multimodal features are evaluated by whether or not to match with intended behaviors given by the participants, and the gaze control model is evaluated by a quantitative analysis of human-interaction intent in the gaze decision process through the pre-set scenario.

5.1. Individual features

To evaluate the performance of the individual recognizer, 10 people participated in the experiment. They each repeated the behaviors indicating interaction intent, and the inverse case for human behavioral cues, 10 times.

Figure 8 shows the performance result of seven dominant features. Most perceptual capabilities mainly show the recognition performance of 80% or more. However, facial expression recognition shows a 76.2% recognition success rate. The main reason of recognition failure is mainly caused by the intensity difference between changing and maintaining facial representations. For the SSL, its success rate is 73.1% and the primary cause is due to not triggering the audio signal at a long distance correctly.

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

Table II. Comparison results of TCR process for individual recognizer.

	Misidentification rate		
Feature	Without TCR	With TCR	Improvement
x_f	23.79%	14.88%	8.91% p
x_g	11.81%	6.20%	5.61% p
x_m	14.40%	11.50%	2.90% p

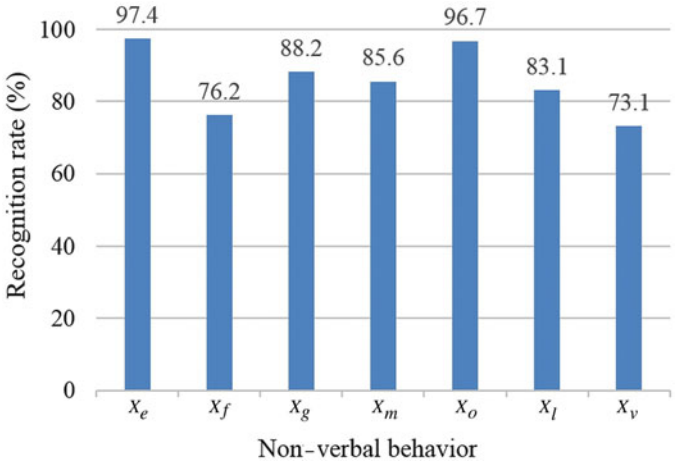


Fig. 8. Recognition results for non-verbal behaviors.

As a post-processing step, TCR was performed under conditions of imperfect sensing data in order to evaluate the reliability of recognition results by estimating confidence of each recognition result. Table II shows the improvement of recognition performance for three features, which have a variety of dynamic characteristics, through TCR.

In particular, misidentification rate of facial expression is 14.88%, which is an 8.91% p improvement. This improvement seems to be a result of raising the confidence level information from $\tau = 3$. From the consecutive relationships between time intervals, TCR removes misidentified data and allows more robust recognition by explicitly modeling the confidence level of perceptual information.

5.2. Multimodal features

The next experiment is to evaluate the suitable interlocutor selection in social interaction with multiple persons by means of multimodal features. Figure 9(a) shows the meaning of the indices in Fig. 9(b); that is, a line connecting several of the features indexed by a number in Fig. 9(a) indicates that those behaviors were combined to attract the robot's attention by participants, and that index corresponds to the same index on the X-axis of Fig. 9(b).

A total 20 people participated in the recognizable behavior combination of 20 cases in MPIs. As shown in Fig. 9(b), the designated target selected by the proposed model exhibited an average concordance rate of 85.6% in the comparison with their own evaluation of the participants. In particular, the proposed model showed strengths in personalized engagement situation by a body pose feature and interlocutor's likes and dislikes at closer quarters. However, the model occasionally exposed a limitation in complementary performance of multimodal features due to the recognition failure of momentary events such as calling name.

5.3. Intentional behaviors in MPIs

The following experiment is to check quantitative capabilities for selecting the interlocutor in a given scenario where three participants (P1–P3) express several intentional behaviors that fit their profiles according to the event timeline such that anybody could easily find which person has the highest interaction intent as assessment criteria (see Fig. 10). If the robot stared at the designated interlocutor

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

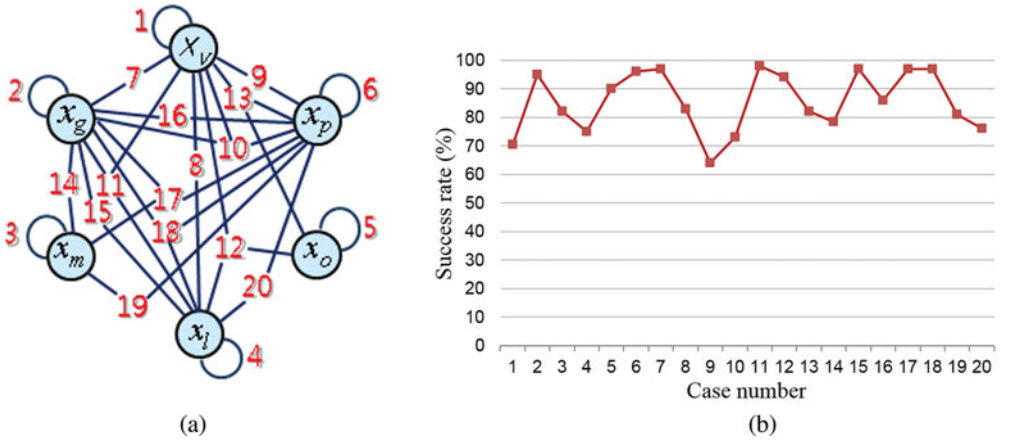


Fig. 9. Recognition results of multimodal features: (a) Feature combinations in each test case and (b) Success rate of the 20 cases.

Human	Behavior	Event timeline												
		1	2	3	4	5	6	7	8	9	10	11	12	13
*P*1	x_e													
	x_f, x_g, x_m													
	x_p, x_o, x_l													
	x_v													
*P*2	x_e													
	x_f, x_g, x_m													
	x_p, x_o, x_l													
	x_v													
*P*3	x_e													
	x_f, x_g, x_m													
	x_p, x_o, x_l													
	x_v													
Designated interlocutor		*P*1	*P*2	*P*3	*P*1	*P*2	*P*1	*P*3	*P*1	*P*3	*P*2	*P*1	*P*3	*P*1

Fig. 10. Intentional behaviors of three participants in the event timeline.

in each sequence of event timeline, the attempt would be considered a success, and its performance is evaluated by the average success rate for 13 events.

For this experiment, the 15 participants were involved in interaction with the robot after forming three-person teams. While performing the actual experiment, participants appropriately responded to the robot's questions according to their own ideas and performed the planned behaviors if the robot was not gazing. Raw feature data were acquired at a 10Hz sample rate. Figure 11(a) and (b) show the results of each participant's accumulated feature values every 1 s and response sequence in one scenario of MPI, respectively. Figure 11(c) shows the results of interlocutor selection representing highest scorer by comparison of cumulative values at the appraisal time. Figure 11(d) shows human trajectories in the test scenario and Fig. 11(e) shows the experimental result of gaze decision derived by means of face identification, target selection, and robot gaze trajectory including three participant's locations in the course of gaze decision process. Throughout the pre-set scenario, the robot appropriately conducts self-introduction (t1–t3), a short greeting (t3–t5) for each participant after identifying the face, and a gaze shift to the designated target (t6–t13) by calculating the quantitative appraisal value and communicative intimacy of participants.

Total experiments showed a success rate of 93.3%. It shows that the robot is able to reproduce socially acceptable gaze control to the person who expresses intimate behavior among multiple persons, even though the model occasionally exposed a limitation in momentary events such as

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

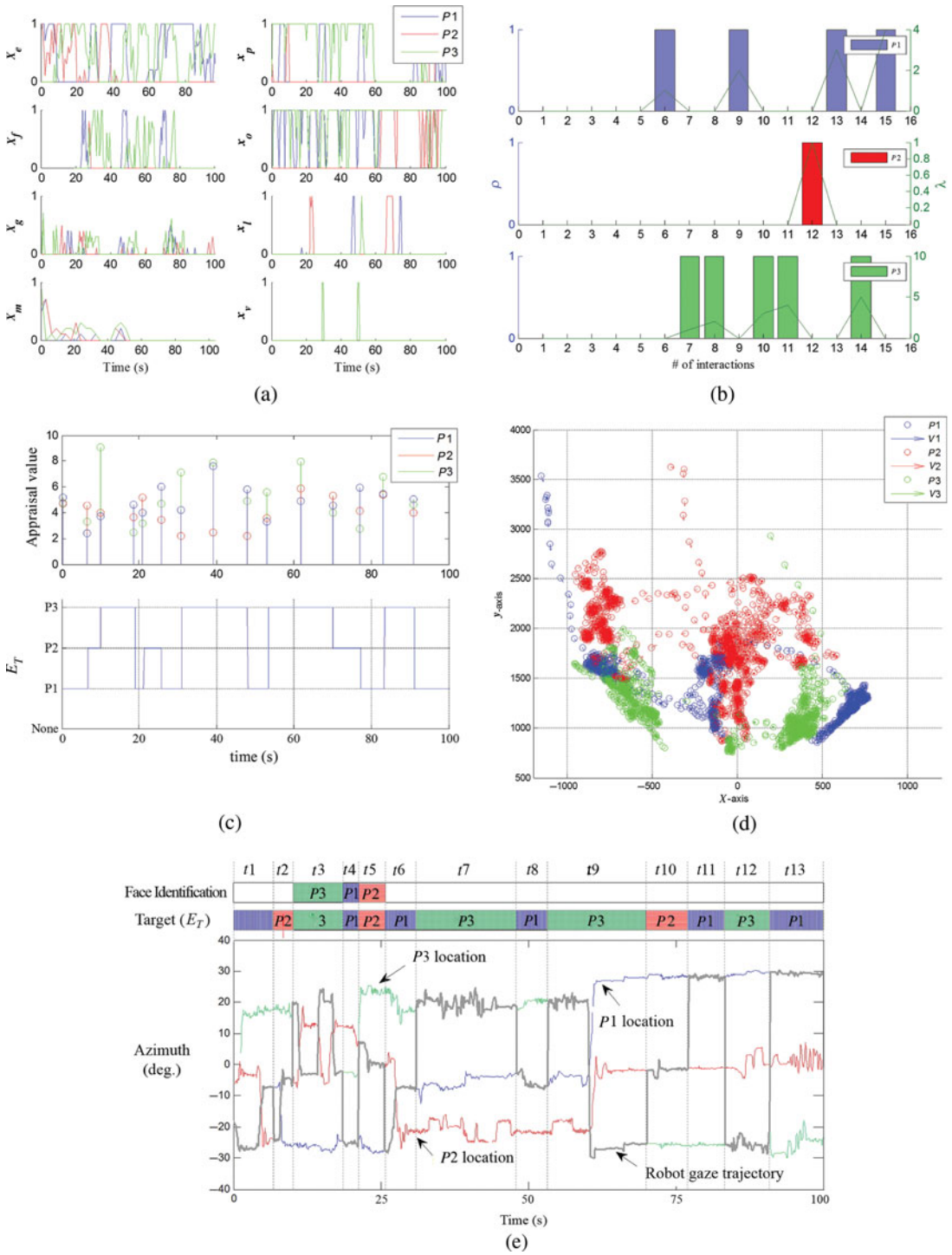


Fig. 11. Quantitative measurements for the interlocutor selection in MPIs: (a) The numerical values of each feature per participant, (b) The number of responses for robot's questions, (c) The interlocutor selection with appraisal values of each participant, (d) Trajectories for three participants in the test scenario, and (e) The face identification, target selection, and robot's gaze shifts according to the timeline.

calling robot's name. Figure 12 shows a representative scene of robot's gaze decision in MPIs to appropriately demonstrate decision-making through the quantitative appraisal model.

In order to check the influence of each feature in the MPI, participants set the priority of eight features in advance, and they re-evaluated the priority from personal opinion after experiencing social interaction with robot. Figure 13 shows the pre-post comparative results. Comparing before and after

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

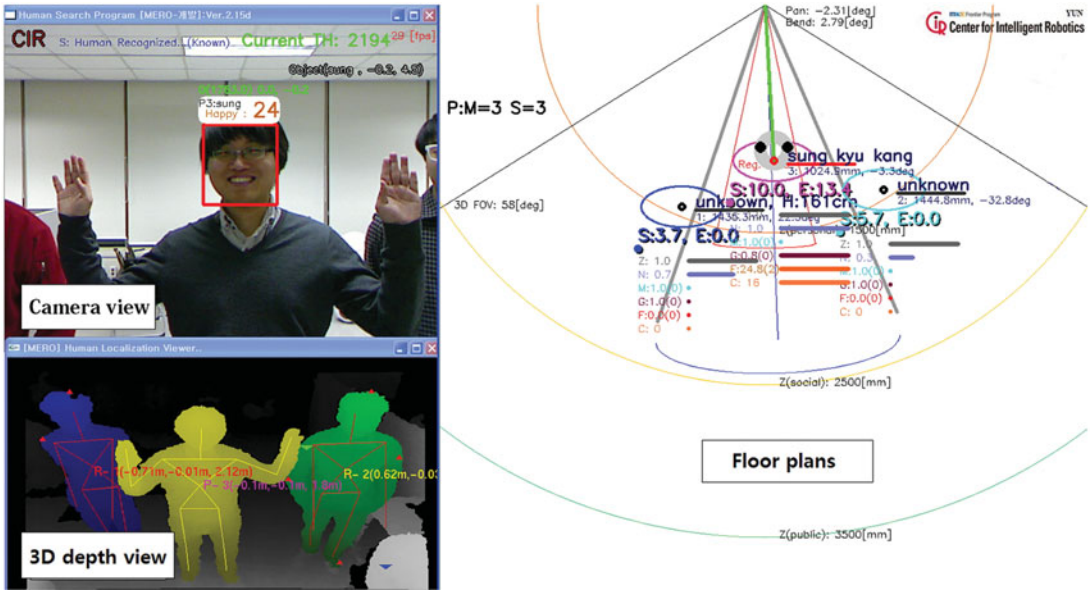


Fig. 12. A demo scene of representing intended feature by participants.

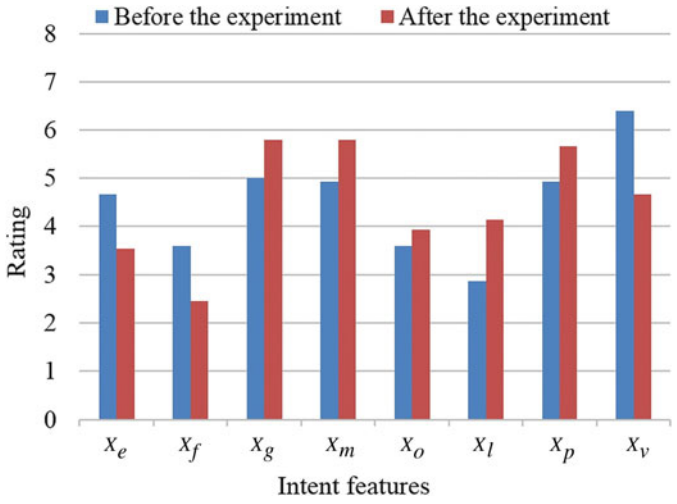


Fig. 13. Pre-post comparative approach of each intention feature by participants' evaluation.

the experiment, the features of user call x_v, eye movement x_e, and facial expression x_f were ranked lower than participants had expected in advance. On the contrary, the behaviors of body gesture x_g, movement x_m, and proximity x_p were more clearly elevated, and the percentage increase in the body leaning x_l showed the highest (44.2%) level of interaction intent. Taken together, these kinesic and proxemic factors are affecting pre-post comparative results.

Three main findings emerged from this performance evaluation. First, as a key element of the robot required for selecting a gaze target, understanding of an individual's interaction intent is accessible by the dominant feature extraction and multiple feature combination regarding representative non-verbal behaviors. In addition, TCR helps to improve the misidentification rate as a post-processing step. Second, a hybrid approach incorporating gaze control criteria and numerical computation of dominant features was designed to avoid random or simplistic perception-based gaze target selection. It has the advantage to improve the quality of ongoing dialog by the high concordance rate with human evaluation in a given test scenarios. Third, while performing a gaze behavior, the process of recognizing each participant and calling their names was mentioned as the most impressive point in the subjective opinion by participants. Therefore, it is confident that the gaze decision process could establish positive engagement with the opponent in terms of performing acceptable gaze control.

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

However, there are many difficulties to consider for coping with various participants' characteristics, diversifying interaction scenarios, and accounting for changing environments due to the limitation of the corresponding recognition techniques and computational capabilities. In particular, in the interaction design, the success rate of a proposed model was determined by whether to match opinions of most participants for the given scenarios such that anybody could easily find which person has the highest interaction intent. But humans could make a decision in selecting a different interlocutor in accordance with person and/or social context. The weighting factor of the proposed model were also configured through a heuristic method in a given experimental environment without suggesting clear basis. In addition, the proposed scheme only carried out social interaction with the interaction partner through one-sided questions about pre-configured individual dialog topics to prevent the poor response for unpredictable conversation. The application of robust speech recognition and dialog content would directly enhance the capability to make a gaze decision in the situated social interaction.

Nevertheless, it is confirmed that the proposed scheme is suitable for selecting an appropriate interaction target via socially acceptable manner.

6. Conclusion

In this paper, a computational model for gaze control scheme of socially interactive robots in an environment involving multiple possible interaction partners has been presented. To support this, reliable recognition modules were applied through multi-dimensional analysis and dominant feature extraction from human behavioral cues. In addition, two factors of physical space and communicative intimacy were applied to improve the quality ongoing dialog in a controlled manner. TCR was also considerable help to reduce misidentification rate. By performance evaluation with interaction scenarios, the proposed model conducted a reasonable selection of the interlocutor. Accordingly, it shows a potential to establish an intimate and positive social relationships with multiple people.

However, improved voice recognition and rich dialog contents remain a problem that should be done in the future.

Supplementary Material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0263574716000722.

References

1. M. Argyle, *Social Interaction* (Methuen, London, 1969).
2. S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nat. Rev. Neurosci.* **2**, 561–567 (2001).
3. T. Fong, I. Nourbakhsh and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.* **42**, 143–166 (2003).
4. N. Kirchner, A. Alempijevic and G. Dissanayake, "Nonverbal Robot-Group Interaction Using an Limited Gaze Cue," *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. (HRI 2011), Lausanne, Switzerland, IEEE Press, New York, NY (2011) pp. 497–504.
5. J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost and M. Stone, "Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents," *Proceedings of the 21st Annual Conference on Computer Graphics Interactive Techniques* (SIGGRAPH 1994), Orlando, USA, ACM, New York, NY (1994) pp. 413–420.
6. M. Rehm and E. Andre, "Where do They Look? Gaze Behaviors of Multiple Users Interacting with An Embodied Conversational Agent," *Proceedings of the International Workshop on Intelligent Virtual Agents* (IVA 2005), Kos, Greece, (2005), pp. 241–252.
7. B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro and N. Hagita, "Footing in Human-Robot Conversations: How Robot Might Shape Participant Roles Using Gaze Cues," *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (HRI 2009), San Diego, USA, IEEE Press, New York, NY (2009) pp. 61–68.
8. Y. Matsusaka, S. Fujie and T. Kobayashi, "Modeling of Conversational Strategy for the Robot Participating in the Group Conversation," *Proceedings of the European Conference on Speech Communication and Technology*. Aalborg, Denmark (INTERSPEECH 2001) pp. 2173–2176.

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

9. D. Bohus and E. Horvitz, "Facilitating Multiparty Dialog with Gaze, Gesture, and Speech," *Proceedings of the International Conference on Multimodal Interfaces* (ICMI 2010), Beijing, China, (2010), pp. 5:1–5:8.
10. Y. Kondo, K. Takemura, J. Takamatsu and T. Ogasawara, "A gesture-centric android system for multi-party human-robot interaction," *J. Human-Robot Interact.* **2**(1), 133–151 (2013).
11. G. H. Lim and I. H. Suh, "Robust robot knowledge instantiation for intelligent service robots," *Intel. Serv. Robot.* **3**, 115–123 (2010).
12. H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi and H. Kitano, "Human-Robot Interaction Through Real-Time Auditory and Visual Multiple-Talker Tracking," *Proceedings of the IEEE Internatonal Conference on Intelligent Robots and Systems* (IROS 2001), St. Louis, USA, IEEE Press, New York, NY (2001) pp. 1402–1409.
13. J. G. Trafton, M. D. Bugajska, B. R. Fransen and R. M. Ratwani, "Integrating Vision and Audition within a Cognitive Architecture to track Conversations," *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (HRI 2008), Amsterdam, Netherlands, IEEE Press, New York, NY (2008) pp. 201–208.
14. S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink and G. Sagerer, "Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot," *Proceedings of the International Conference on Multimodal Interfaces* (ICMI 2003), Vancouver, Canada, ACM, New York (2003) pp. 28–35.
15. J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink and G. Sagerer, "Multi-modal anchoring for human-robot interaction," *Robot. Auton. Syst.* **43**, 133–147 (2003).
16. M. P. Michalowski, S. Sabanovic and R. Simmons, "A Spatial Model of Engagement for a Social Robot," *Proceedings of the IEEE International Workshop on Advanced Motion Control* (ACM 2006), IEEE Press, New York, NY (2006) pp. 762–767.
17. D. Bohus and E. Horvitz, "Dialog in the Open World: Platform and Applications," *Proceedings of the International Conference on Multimodal Interfaces* (ICMI 2009), Beijing, China, ACM, New York (2009) pp. 31–38.
18. R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz and J. Wang, "Designing Robots for Long-Term Social Interaction," *Proceedings of the IEEE International Conference on Intelligent Robots and Systems* (IROS 2005), Alberta, Canada, IEEE Press, New York, NY (2005) pp. 2199–2204.
19. N. Bellotto and H. Hu, "Multimodal Perception and Recognition of Humans with a Mobile Service Robot," *Proceedings of the International Conference on Robotics and Biomimetics*, Guangxi, China, (ROBIO 2009) pp. 401–406.
20. E. Torta, J. Heumen, F. Piunti, L. Romeo and R. Cuijpers, "Evaluation of unimodal and multimodal communication cues for attracting attention in human-robot interaction," *Int. J. Soc. Robot.* **2**, 89–96 (2015).
21. N. Kirchner, A. Alempijevic and G. Dissanayake, "Nonverbal Robot-Group Interaction Using an Imitated Gaze Cue," *Proceedings of the IEEE International Conference on Human-Robot Interaction* (HRI 2011), Lausanne, Switzerland, IEEE Press, New York, NY (2011) pp. 497–504.
22. B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Trans. Interact. Intell. Syst.* **1**(2), 1–33 (2012).
23. A. Zaraki, D. Mazzei, M. Giuliani and D. D. Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Trans. Human-Machine Syst.* **44**(2), 157–168 (2014).
24. Y. I. Nakano, T. Yoshino, M. Yatsushiro and Y. Takase, "Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction," *ACM Trans. Inter. Intell. Syst.* **5**(4), 157–168 (2015).
25. S. Weis, Theory and Measurement of Social Intelligence as a Cognitive Performance Construct *Ph.D. Dissertation*, (Otto-von-Guericke University, Magdebrug, Germany, 2008).
26. W. C. Ho, K. Dautenhahn, M. Y. Lim and K. D. Casse, "Modeling Human Memory in Robotic Companions for Personalisation and Long-term Adaptation in HRI," *Proceedings of the 1st International Conference on Biologically Inspired Cognitive Architectures,* BICA, Virginia, USA (2000) pp. 64–71.
27. E. T. Hall, *The Hidden Dimension: Man's Use of Space in Public and Private* (Doubleday, New York, 1966).
28. M. L. Patterson, "Spatial factors in social interactions," *Human Relations*, **21**, 351–361 (1968).
29. M. L. Walters, K. Dautenhahn, R. Boekhorst, K. L. Koay, D. S. Syrdal and C. L. Nehaniv, "An Empirical Framework for Human-Robot Proxemics," *Proceedings of the New Frontiers in Human-Computer Interaction,* Edinburgh, Scotland, (AISB 2009) pp. 144–149.
30. R. Hartson, "Cognitive, physical, sensory, and functional affordances in interaction design," *Behav. Inf. Technol.* **22**(3), 315–338 (2003).
31. A. Bruce, I. Nourbakhsh and R. Simmons, "The Role of Expressiveness and Attention in Human-Robot Interaction," *Proceedings of the IEEE International Conference on Robots and Automation* (ICRA 2002), Washington, USA, IEEE Press, New York, NY (2002) pp. 4138–4142.
32. A. Mehrabian, *Nonverbal Communication* (Aldine-Atherton, New York, 1972).
33. D. Egolf and S. L. Chester, *The Nonverbal Factor: Exploring the Other Side of Communication* (iUniverse, New York, 2007).
34. P.A. Andersen, The cognitive valence theory of intimate communication **In** M.T. Palmer and G.A. Barnett (Eds.), *Progress in Communication Sciences*, vol. **14**: Mutual influence in interpersonal

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press

communication theory and research in cognition, affect, and behavior (pp. 39–72) (Ablex, Norwood, NJ, 1988).
35. J. Yoon and D. Kim, "Frontal Face Classifier Using Adaboost with Mct Features," *Proceedings of the International Conference on Control, Automation, Robotics and Vision*, Singapore, (ICARCV 2010) pp. 2084–2087.
36. OpenNI NITE, Available at: http://openni.ru/files/nite/.
37. B. Lee, H. Kim, J. Choi, S. Kim and N. I. Cho, "DSP integration of sound source localization and multi-channel wiener filter," *Proceedings of the International Conference on Robotics & Automation* (ICRA 2010), Alaska, USA, IEEE Press, New York, NY (2010) pp. 4830–4835.
38. B. Jun, J. Lee and D. Kim, "A novel illumination-robust face recognition using statistical and non-statistical method," *Pattern Recog. Lett.* **32**(2), 329–336 (2011).
39. E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study* (Kluwer, Norwell, MA, 2000).
40. K.-G. Oh, C.-Y. Jung, Y.-G. Lee and S.-J. Kim, "Real-Time Lip Synchronization Between Text-to-Speech (TTS) System and Robot Mouth," *Proceedings of the International Symposium on Robot and Human Interactive Communication* (RO-MAN 2010), Viareggio, Italy, IEEE press, New York (2010), pp. 620–625.

https://doi.org/10.1017/S0263574716000722 Published online by Cambridge University Press