Intention Reading from a Fuzzy-based Human Engagement Model and Behavioural Features

Regular Paper

Sang-Seok Yun¹,², Mun-Taek Choi², Munsang Kim² and Jae-Bok Song¹*

¹ Department of Mechanical Engineering, Korea University, Korea
² Center for Intelligent Robotics at Korea Institute of Science and Technology (KIST), Korea
* Corresponding author E-mail: jbsong@korea.ac.kr

Received 29 May 2012; Accepted 12 Jun 2012

DOI: 10.5772/50648

© 2012 Yun et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract This paper presents a novel approach for a quantitative appraisal model to identify human intent so as to interact with a robot and determine an engagement level. To efficiently select an attention target for communication in multi-person interactions, we propose a fuzzy-based classification algorithm which is developed by an incremental learning procedure and which facilitates a multi-dimensional pattern analysis for ambiguous human behaviours. From acquired participants’ non-verbal behaviour patterns, we extract the dominant feature data, analyse the generality of the model and verify the effectiveness for proper and prompt gaze behaviour. The proposed model works successfully in multiple people interactions.

Keywords Fuzzy min-max neural networks (FMMNN), gaze behaviour, focus of attention, human-robot interaction, intention reading, multi-modal sensors

1. Introduction

It is of great importance for a robot to understand the complicated social behaviours of humans and properly respond to them. Social factors, such as information usability, dialogue preference, social acceptance, societal impact, intimacy level and user experience, affect human-robot interaction (HRI). To appropriately manage context awareness in the current situation, an intelligent robot system is required to evaluate external perception and internal inference, as well as the determination of holistic information for social factors. As a communication partner, a situation associated with interaction can be represented by a user’s mood or intention, the spatial distance between humans and robots, the correlation with other users and current conversation states. In particular, human intention in human-human communication refers to a human agent’s mental determination to act in a certain way and a psychophysical approach with functional neuroimaging reviewed evidence that human agents possess the ability to infer other people’s intentions by observing their actions [1].

For understanding a human’s intention from communication signals, some studies report that it is necessary to know the following information: human verbal and non-verbal behaviour [2, 3], spatial
relationships [4] and prior knowledge of the other participant as well as communication skills [5]. We can see the importance that gaze direction in human face-to-face communication serves in shifting roles during turn-exchanges [6].

As one of the most salient signals in HRI, gaze cues can also be effective in shaping the participant roles of conversational partners [7] and the selection of a conversational attention target to exhibit gaze behaviour is surely needed in the presence of multi-person interactions. Other studies propose that computational methods managing the engagement process in dynamic interactions with multiple people are required for detecting whether or not the engaged participants are actively maintaining conversational engagement [8, 9]. An emerging egocentric interaction paradigm based on perception, action, intention and attention capabilities, as well as a lack of human agents, also supports these essentialities [10].

On the other hand, multi-modal interfaces have the potential to improve the usability and accessibility of communication flows, including human activities, and supplement single sensory limitations. The use of multiple HRI modalities in the engagement mechanism with the levels of interaction engagement is proposed in [11]. The multi-modal anchoring approach also shows a schematic example for anchoring a person through anchoring component symbols to their associated percepts with the symbolic level and sensory level [12]. Accordingly, the data acquisition of human activities using multiple sensors plays an important role in the delivery of efficient human communication.

In this paper, we would like to focus on establishing a solution that judges the engagement levels of multiple people by analysing a fuzzy-based human engagement model and multi-dimensional human behavioural features. The empirical evaluation of the proposed fuzzy model, which is developed by a learning procedure of the dominant human behavioural patterns, is utilized to quantitatively classify the uncertainty of human intent so as to interact with the robot. For analysing the generality of the model with individual peculiarities, we perform a quantitative fuzzy region analysis for each participant’s activity through the communicative process and verify its effectiveness from the selection of the communicative attention target from among multiple people.

This paper is organized as follows. Section 2 introduces human activity analysis, Section 3 describes the FMMN algorithm, Section 4 deals with the evaluation of a fuzzy-based human engagement model, Section 5 presents gaze behaviour among multiple people and, finally, in Section 6, conclusions are drawn and future work is outlined.

2. Human activity analysis

2.1 System configuration

Humans may understand the other party’s intention through interpreting various forms of non-verbal behaviours during communication. To observe these behavioural characteristics, we configure the following robot system.

![Figure 1. System configuration of an interactive robot.](image)

An intelligent robot platform - namely MERO - is mainly designed to provide cognitive services to mentally disordered people. It currently provides healthcare for the elderly, educational assistance at elementary schools, and interactive contents in the reception and entertainment events. With respect to HRI, we have the following robot system configuration: a webcam on the robot head, four microphones on the base, a 3D depth sensor located at a certain distance from the robot along the Z-axis in order to detect multiple people and track each person’s trajectory easily through a large sensing range, a 4-DOF neck and 17-DOF face modules (eyebrow, eyelid, pupil and lip). Fig. 1 illustrates the configuration of the robot.

2.2 Multi-modal cues

To acquire human behavioural features, each attached sensor has the following characteristics (see Fig. 2). Four microphones were used in the auditory system, which combined the sound source localization (SSL) and a multichannel Wiener filter (MWF). The SSL is an algorithm which estimates the direction of a speaking person using time delay of arrival (TDOA) among the microphones [13]. We also used a DSP-based embedded system. The MWF and SSL are incorporated in this DSP system. To detect sound direction reliably, a post-process using a histogram is applied to achieve robust sound localization [14].

A vision sensor easily estimates face location and face recognition for multiple people through a variety of vision algorithms in complex environments. To make a unique ID - such as a person’s name - for each human model, we applied an illumination-robust face
recognition using the statistical and non-statistical methods proposed in [15]. For verifying users’ identities, it compares the feature histogram of the current facial image and the pre-registered information for each person (namely, a gallery). The recognizer can guarantee an identification accuracy of around 82%, even in a situation in which a robot head-camera dynamically moves to express its behaviours or track the human. Further detail on face recognition can be found in [15] and [16].

We also recognize the current emotional state of the people by pre-processing a facial image with a neural network and obtain four facial expression states (smiling, surprised, neutral and angry states) with recognition reliability from the current user’s facial expression through the pre-trained weighting factor.

As a spatial modality, the 3D depth sensor provides valid data to measure the 3-dimensional distance of labelled moving objects by creating a depth map clustering, and interpret specific gestures by major joint angles in the kinematic model. To get the human position and posture data, we use the OpenNI library [17]. The library provides a moving object’s location and 24 elements of skeleton position data. From the library, we developed the following perception information: human’s movement patterns (approach, depart, sideways and stop) are analysed from the velocity and direction of moving objects with a naive Bayesian filter, while human gestures (standing, waving and nod) and physical information (human height, shoulder width, limb length, etc.) were easily estimated using the velocity of each skeleton point and the linking vector between skeleton positions.

2.3 Feature extraction

Human behaviours are divided into five categories: oculesic, kinesic, proxemic, vocalic and tactile behavioural features [18]. Each behaviour has dominant features that are commonly observed in human-human non-verbal communication. These features are utilized in human-robot communication to estimate a human’s intimacy [19].

As shown in Table 1, we defined 8 dominant features in the 4 behavioural characteristics to inquire as to the dominant cues for analysing the engagement level. Each feature represents the perception information extracted by the multi-modal cues. These features do not represent all human senses and they have performance restrictions. Nevertheless, we believe that they can be significantly utilized to estimate the interaction intent of humans.

| Behaviour | Feature | Intention attribute |
|-----------|---------|---------------------|
| Oculesic  | Eye movement, $x_e$ | Gazing |
| Kinesic   | Facial expression, $x_f$ | Smiling |
|          | Body gesture, $x_g$ | Waving |
|          | Movement, $x_m$ | Moving forward |
| Proxemic  | Proximity, $x_p$ | Distance |
|          | Body orientation, $x_o$ | Confronting |
|          | Body leaning, $x_l$ | Leaning forward |
| Vocalic   | User call, $x_u$ | Calling |

Table 1. Non-verbal features based on behavioural factors

2.3.1 Oculesic feature

Oculesic features are measured by averaging the mutual gazing time of each eye movement between a human and a robot during the perception time. To perform human eye tracking and eye contact with the robot is not easy. Instead, we estimate the human’s gazing time by using face detection while a robot looks at the user as follows.

$$x_e = \frac{\sum_{s=1}^{n} e(s)}{T},$$

where $e(s) \in \{gazing, not gazing\}$ is the frontal face detection and $T$ is the perception time.

2.3.2 Kinesic feature

Kinesic features are measured by recognizing facial expressions, human movement and body gestures. Facial expressions consist of four states, as mentioned before, and each reliability value is obtained from the neural net output. To make a normalized value between 0 and 1, smiling adds a positive value and the rest of the expression adds negative values.

$$x_f = \frac{\sum_{t=1}^{T} f(I_p) - f(I_a)}{T},$$

where $f(I_p)$ and $f(I_a)$ represent the positive and negative reliability value.
As shown in Fig. 3, the human information about the 15 body components in terms of 3D coordinates allows us to locate various structural feature points around the body. We assume that a waving feature can be easily estimated by tracking the skeleton data of the human body, as follows:

\[
\begin{aligned}
\tilde{t}_{\text{hand},z} &= \tilde{P}_{\text{hand},z,t} - \tilde{P}_{\text{elbow},z,t}, \\
v_{\text{hand}} &= \sqrt{d^2 + d^2}, \\
g(s) &= \begin{cases} 
1, & \text{if } \tilde{t}_{\text{hand},z} > 0, v_{\text{hand}} > v_{th,h} \\
0, & \text{otherwise}
\end{cases}, \\
x_g &= \frac{\sum_{i=1}^{T} g(t)}{T}
\end{aligned}
\]

where \(\tilde{t}_{\text{hand},z}\) is the vector of the hand on the Z-axis, \(v_{\text{hand}}\) is the velocity of the hand on the X-axis, \(i\) is the sampling time, \(v_{th,h}\) is the threshold value of the hand moving velocity, \(P\) represents the 3D joint position for each elbow and hand point, and \(g(s)\) is two gesture states.

From the spatial trajectories of the human, we can also calculate the human movement features by applying the moving velocity and direction of the people:

\[
\begin{aligned}
dx &= \tilde{P}_{\text{torso},x,t} - \tilde{P}_{\text{torso},x,t-1}, \\
dy &= \tilde{P}_{\text{torso},y,t} - \tilde{P}_{\text{torso},y,t-1}, \\
v_m &= \sqrt{dx^2 + dy^2}, \\
m(s) &= \begin{cases} 
1, & \text{if } v_m > v_{th,m}, \theta_m < \theta_{th,m} \\
0, & \text{otherwise}
\end{cases}
\end{aligned}
\]

\[
x_m = \frac{\sum_{i=1}^{T} m(t)}{T}
\]

where \(\tilde{P}_{\text{torso}}\) is the human body position, \(v_m\) is the moving velocity, \(\theta_m\) is the moving direction, \(v_{th,m}\) and \(\theta_{th,m}\) are the threshold values of \(v_m\) and \(\theta_m\).

2.3.3 Proxemic feature

Proxemic features are easily measured by averaging the spatial distance between a person and a robot, body orientation and the angle of body leaning. To describe an appropriate spatial relationship between intimate people and in order to perform effective communication, each personal (\(\leq 1.2m\)), social (\(\leq 2.5m\)) and public (\(\leq 3.5m\)) space proposed in [20] is assigned according to the radius. As a feature value, the average distance for the perception time is compared with each predefined space:

\[
x_p = \frac{\sum d_p}{D_{\text{personal}}}
\]

where \(d_p\) is the spatial distance, \(D_{\text{personal}}\) is the radius of personal space, and \(D_{\text{social}}\) is the radius of social space.

As shown in Fig. 4, we can also estimate body orientation and the angle of body leaning by tracking the 3D skeleton point vector of the neck \(\tilde{P}_{\text{neck}}\), torso \(\tilde{P}_{\text{torso}}\) and left shoulder \(\tilde{P}_{\text{shoulder}}\):

\[
\begin{aligned}
\tilde{t}_{\text{human}} &= \tilde{P}_{\text{torso}} - \tilde{P}_{\text{robot}}, \\
\tilde{t}_{\text{body}} &= \tilde{P}_{\text{neck}} - \tilde{P}_{\text{shoulder}}, \\
\tilde{t}_{\text{head}} &= \tilde{P}_{\text{torso}} - \tilde{P}_{\text{robot}}, \\
\tilde{u}_1 &= [1 0 0]^T, \\
\theta_R &= \cos^{-1}\left(\frac{\tilde{t}_{\text{human}} \cdot \tilde{u}_1}{\|\tilde{t}_{\text{human}}\| \|\tilde{u}_1\|}\right)
\end{aligned}
\]
\[ \theta_H = \cos^{-1}\left( \frac{l_{\text{body}} \cdot v_1}{\|l_{\text{body}}\| \|v_1\|} \right) \]  

(13)

\[ \theta_o = \theta_R - \theta_H \]  

(14)

\[ x_o = \frac{1}{T} \sum_{t} \theta_o \]  

(15)

\[ \bar{t}_{\text{lean}} = \bar{j}_{\text{neck}} - \bar{j}_{\text{torso}}, \quad \bar{u}_2 = [0 \ 0 \ 1]^T \]  

(16)

\[ \theta_i = \cos^{-1}\left( \frac{\bar{t}_{\text{lean}} \cdot v_2}{\|\bar{t}_{\text{lean}}\| \|v_2\|} \right) \]  

(17)

\[ x_i = \frac{1}{T} \sum_{t} \theta_i \]  

(18)

where \( \bar{t}_{\text{human}} \) is the human vector between a human and a robot, \( l_{\text{body}} \) is the body posture vector, \( \theta_R \) represents the angle of a human position from the robot along the X-axis, \( \theta_H \) represents the human body orientation, \( \theta_o \) is the confronting angle of the body plane, \( \bar{t}_{\text{lean}} \) is the vertical body posture vector and \( \theta_i \) is the forward leaning angle based on the location of the robot body.

### 2.3.4 Vocalic feature

Vocalic features are measured by estimating the direction of a speaking person with TDOA and their spatial location. We assumed a designated person to be a speaker if a sound signal (\( s_{\text{TDOA}} \)) is detected in azimuth for the human location:

\[ \theta_{\text{human}} = \tan^{-1}\left( \frac{P_{\text{torso,x}}}{P_{\text{torso,y}}} \right) \]  

(19)

\[ v(s) = \begin{cases} 
1, & \text{if } s_{\text{TDOA}} \geq s_{\text{th}} \chi_{\text{human}} - \theta_{\text{ind}} \leq \theta_{\text{ind}, s} \\
0, & \text{otherwise}
\end{cases} \]  

(20)

\[ x_i = \frac{1}{T} \sum_{t} v_i(s) \]  

(21)

where \( \theta_{\text{human}} \) is the human relative angle for a robot, \( s_{\text{th}} \) is the threshold value of a TDOA signal, \( \theta_{\text{ind}} \) is the sound azimuth and \( \theta_{\text{ind}, s} \) is the threshold value of the azimuth.

### 3. FMMNN

In this section, we propose a method to assess engagement level using the fuzzy min-max neural networks (FMMNN), which is characterised by a nonparametric classification, a simple learning method and incremental parameter tuning [21, 22]. We applied the proposed model for the following reason: it explicitly provides quantitative design criteria with fuzzy regions so as to determine the human’s interaction intent, and facilitates the multi-dimensional analysis of the extracted numeric feature data as a powerful pattern classifier. The feature values are transformed to the quantifying membership value by fuzzy composition rules.

#### 3.1 Algorithm description

A fuzzy hyperbox membership function plays an important role in the fuzzy min-max classification. As a fuzzy set, hyperboxes are defined by pairs of min-max points. Following [21], let the \( j \)th hyperbox fuzzy set (\( B_j \)) be formally defined by the ordered set:

\[ B_j = \{X_h, V_j, W_j, b_j(X_h, V_j, W_j) \mid \forall X_h \in \mathbb{R}^n \} \]  

(22)

for all \( h = 1, 2, \cdots, m \), where \( X_h \) is the \( h \)th input feature, \( V_j \) is the min point and \( W_j \) is the max point for the \( j \)th hyperbox, and the membership function for the \( j \)th hyperbox is \( 0 \leq b_j(X_h, V_j, W_j) \leq 1 \). Following [22], to meet the required criteria, a membership function is defined as:

\[ b_j(X_h, V_j, W_j) = \min_{i=1, \cdots, a} \left( \min([1 - f(x_{hi} - w_{ji}, \gamma_i)], [1 - f(v_{ji} - x_{hi}, \gamma_i)]) \right) \]  

(23)

where \( f(r, \gamma) = \begin{cases} 
1, & \text{if } r \gamma > 1 \\
r \gamma, & \text{if } 0 \leq r \gamma \leq 1 \\
0, & \text{if } r \gamma < 0
\end{cases} \)

ramp threshold function and \( \gamma \) is the sensitivity parameter regulating how fast the membership values decrease.

**Figure 5.** The three-layer neural network which implements a FMMNN classifier.
An implementation of the neural network is shown in Fig. 5. It has a three layer feedforward neural network: input, hyperbox and class nodes. The connections between the second- and third-layer nodes are binary values stored in the matrix \( U \). The equation for assigning the values of matrix \( u_{jk} \) is:

\[
u_{jk} = \begin{cases} 1, & \text{if } b_j \text{ is a hyperbox for class } c_k \\ 0, & \text{otherwise} \end{cases}
\] (24)

where \( b_j \) is the \( j \)th second-layer node and \( c_k \) is the \( k \)th third-layer node. Each third-layer node represents a class. The output of the third-layer node represents the degree to which the input feature \( X_k \) fits within the class \( k \). The transfer function for each of the third-layer nodes is defined as:

\[ c_k = \max_{j=1}^m b_j u_{jk} \] (25)

### 3.2 FMMNN learning procedure

To create a fuzzy hyperspace for reading human intention, the following input feature vectors should be trained by the FMMNN. The training data \( D \) consists of a set of \( M \) ordered pairs, namely:

\[
\{X_k, d_h\}, \quad h = 1, 2, \cdots, M
\] (26)

\[ X_k = \{x_k(h), x_f(h), x_g(h), x_y(h), x_p(h), x_o(h), x_f(h)\} \] (27)

where \( x_i(h) \) is the normalized value of the \( i \)th feature vector at a perception time \( h \) and \( d_h \in [c_0, c_1] \) is the index of one of the class layer nodes.

We set a total of 7 feature vectors as input nodes. The outputs of the class layer nodes are defined by an engagement term with two classes of an inattention (\( c_0 \)) and an attention (\( c_1 \)) attitude and were classified by the following learning algorithm. Fuzzy min-max learning has an expansion/contraction process with four steps and repeats steps 1 to 4 for the two classes by turns:

**Step 1:** Extract the behavioural feature values from the perception information during a perception time \( T \).

**Step 2:** Search the hyperbox that can expand and then expand it when it has expanded. If there is no such hyperbox, add a new one.

**Step 3:** Find overlaps between hyperboxes from different classes.

**Step 4:** Adjust between overlapped hyperboxes to eliminate it.

Two classes can be fuzzified by hyperbox membership functions to quantify fuzzy regions for behavioural features, and we can determine the class for the engagement level by winner-take-all rules, as shown in (25). As can be seen in Fig. 6, a fuzzy hyperbox membership function can also represent these fuzzy classes on a one-dimensional feature axis.

![Figure 6. Fuzzy hyperbox membership function with the inattention and attention classes.](image)

The membership values for the class are de-fuzzified to the appropriate engagement values by fuzzy min-max composition rules from multi-dimensional features, as described in:

\[
I = F(\max(\min_{i=1-N}(b(x_i, c_0)), \min_{i=1-N}(b(x_i, c_1)))
\] (28)

where \( F(a) = \begin{cases} 1 & \text{if } \arg \max(a) = c_1 \\ 0 & \text{if } \arg \max(a) = c_0 \end{cases} \), \( I \) is the engagement value, \( i \) is the number of \( i \)th feature, \( N \) is the feature dimension size and \( b(x_i, c) \) is the membership function of class \( c \) for the \( i \)th feature. The engagement values are acquired from the membership function having maximum values by the min operation. Finally, we can evaluate a quantitative human intention interacting with the robot by numerical values.

### 3.3 Learning human model from the HRI

To acquire the training data set of \( M \) ordered pairs \( \{X_k, d_h\} \) for two classes, we conduct an interaction scenario with a turn-taking procedure. As shown in Fig. 7, the main communicative process of the HRI consists of six steps, as follows:

**Step 1:** A robot starts from idling mode when no one is present.

**Step 2:** Once a user appears, the robot detects a spatial location for the participant, begins the behavioural percepts from multi-modal sensors and induces the participant to come up to the robot.

**Step 3:** While analysing the dominant behavioural features for the participants who are not identifiable, face recognition is used to determine whether or not to register once.
Step 4: Once checking user registration, the robot simply makes an acknowledgement and greets based on the personal information for each identified person.

Step 5: After finishing the user identification for the participant, the robot establishes a simple dialogue strategy, including the participant’s vocal response, in order to perform turn-taking with ten consecutive questions.

Step 6: For the whole communicative process through the two attention and inattention classes, the robot observes human activities in each turn-taking procedure and identifies each user’s intention from the behavioural features of the human.

The turn-taking procedure in the dialogue context consists of a robot’s question, the user’s response and the robot’s interjection flow. Questions include general knowledge, personal information and the user’s current feelings through facial expression recognition. According to the user’s response for the question, a robot can be represented by positive or negative interjections.

Fig. 8 shows a sample scene of the different learning procedures of a fuzzy-based human engagement model for participants’ behavioural patterns.

4. Evaluation of a human model

Ten users participated in the learning experiment of the communicative process with 7 behavioural features. Each participant has a dialogue process with a greeting, identification and a questions game. To make a decision criteria for robot gaze behaviour, a total of 1400 elements of training data (2 classes for an engagement level, 10 participants, 10 dialog process, and 7 dominant behavioural features) were acquired in the experiment. Fig. 9 shows two fuzzy-based human engagement models with different aspects in relation to the 10 participants.

Subject 6 in Fig. 9(a) actively expresses his intention for the attention and inattention attitude. With each dialogue in an attention class, he variously changes a gaze, a facial expression and a gesture. In the opposite class is a similar trend. In particular, for the body orientation and leaning part, he shows a positive approach by expressing frontal and forward behaviours in both classes compared with the other participants.

Subject 9 in Fig. 9(b) expresses an inactive behaviour when compared relatively with the other participants. Low activities, such as significantly less gesture, a non-frontal body leaning pose and a relatively distant location, are expressed during the conversation. Regardless of the two classes, he gazes at the robot and poses his body orientation with the frontal side.

Fig. 9(c) shows the generalized fuzzy-based human engagement model made by all the training data sets. The fuzzy boundaries of most features converse in two classes. A gesture feature shows fully dominant characteristics in order to get the robot’s attention. However, movement and proximity features effects are relatively small in relation to the behavioural factors of the model.

From this result, the analysis of behavioural features in FMMNN shows clearly that the hyperboxes of each feature should be condensed and spaced between two classes and that membership values placed in fuzzy boundaries numerically represent the quantity of the interaction intent of individuals.
To validate the effectiveness of a fuzzy-based human engagement model from the feature data acquired from the 10 participants, we generate the trained model by randomly selecting 7 training sample sets while the rest are tested as a testing sample. The result demonstrates that the space is sufficient to classify the engagement level with 86.2% accuracy between two opposite classes in the scenario. Fig. 10 shows a fuzzy-based trained human model with 3 representative features of 7 participants.

![Figure 10. Fuzzy-based human engagement model trained by seven participants.](image)

We agree with the fact that the boundaries within many hyperboxes neither allow the clear description of the human’s overall behaviours nor cover the individual peculiarities of all people. However, we assert that this fact is still useful for evaluating people’s engagement levels and conversational attention, and that it forms the robot’s gaze behaviour from the human behavioural analysis.

5. Gaze behaviour in multiple people

In order to appropriately perform a robot’s gaze behaviour for multiple users, we assume that a robot can spontaneously select the person who stands closer than do other people from the robot if no one has interaction intent. To handle this situation as well as with multiple person interactions, a proximity ranking should be calculated according to the relative location of the people from the robot. We created the performance test of the robot’s gaze behaviour for 3 participants and set it so that each participant could express their behavioural activities with discretion, as shown in Fig. 11.

![Figure 11. Three participants’ activities in the event timeline; P1 = participant 1, P2 = participant 2, P3 = participant 3, R = proximity ranking.](image)

According to the event timeline in Fig. 11, three participants come to the robot sequentially and individually express human activities so as to show interaction intent with time intervals, and then go away from the robot in turn. We ordered the scenario so that nobody would perform an action in interacting with the robot for any moment in the event timelines 1 and 4 in order to verify gaze behaviour based upon the proximity ranking in an inattention class of the human engagement model.
Fig. 12 shows the participants’ moving trajectories in accordance with the event timeline, where three people employ a number of non-verbal signals to show their engagement level. Fig. 13(a) shows the analysis of the proximity ranking and each participant’s numerical feature patterns in multi-person interactions.

All of the participants as a whole tried to pay more attention to that robot which talked to them rather than to another person because the robot verifies user’s identity and asks questions for the corresponding personal information in a communicative process. Each participant has been selected by a fuzzy-based human engagement model analysis as a conversational attention target on three separate occasions and the robot performed a gaze behaviour with tracking of the designated person (see Fig. 13(b)).

6. Conclusion

We propose a novel approach in order to efficiently select a communicative attention target from among multiple people. Human activities are modelled from 8 dominant features in the four categories (kinesic, proxemic, oculsic and vocalic characteristics). The dominant features are extracted by multi-modal cues during the communicative process. The evaluation of a fuzzy-based human model with two engagement classes comes from the extraction of multi-dimensional human behavioural features and identifies the user’s intent to interact with the robot according to the empirical evaluation of two quantitative classes. To verify the quantity for each participant’s intent, we tested an engagement level by assessing the human’s numerical values and performed the robot’s gaze behaviour in an interaction situation with multiple people.

7. Acknowledgments

This research was performed for the Intelligent Robotics Development Programme, one of the Frontier R&D Programs funded by the Ministry of Knowledge and Economy (MKE) of the Republic of Korea.

8. References

[1] S.-J. Blakemore and J. Decety, “From the perception of action to the understanding of intention,” Nature Rev., Neurosci., vol. 2, pp. 561-567, 2001.
[2] E. Goffman, Behavior in public places: notes on the social order of gatherings. The Free Press, New York, 1963.
[3] T. Sato, Y. Nishida, J. Ichikawa, Y. Hatamura and H. Mizoguchi, “Active understanding of human intention by a robot through monitoring of human behavior,” in Proceedings of IROS’94, pp. 405-414.
[4] E. T. Hall, The hidden dimension: man’s use of space in public and private. New York: Doubleday, 1966.
[5] M. R. Dimatteo, A. Taranta, H. S. Friedman and L. M. Prince, Predicting patient satisfaction from physician’s nonverbal communication skills, Medical Care, 1980, vol. 18, pp. 376-387.

[6] C. Goodwin, “Restart, pauses, and the achievement of mutual gaze at turn-beginning,” Sociological Inquiry, vol. 50, pp. 272-302, 1980.

[7] M. Rehm and E. Andre, “Where do they look? Gaze behaviors of multiple users interacting with an embodied conversational agent,” in Proceedings of Int. Workshop Intell. Virtual Agents (Lecture Notes in Computer Science), T. Panayiopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier and T. Rist (Eds.), vol. 3661, pp. 241-252, 2005.

[8] D. Bohus and E. Horvitz, “On the challenges and opportunities of physically situated dialog,” in AAAI Fall Symp. Dialog with Robots, Arlington, VA, 2010.

[9] D. Bohus and E. Horvitz, “Models for multiparty engagement in open-world dialog,” in Proceedings of SIGeal’09, London, UK, Sept. 2009, pp. 225-234.

[10] D. Surie, T. Pederson and L.-E. Janlert, “Human cognition as a foundation for the emerging egocentric interaction paradigm,” in Proceedings of Human-Computer Interaction, M. Zacarias and J. V. de Oliveira (Eds.), vol. 396, pp. 349-374, 2012.

[11] T. E. Rogers, A. S. Sekmen and J. Peng, “Attention mechanisms for social engagements of robots with multiple people,” in Proceedings of ROMAN’06, Sept. 2006, pp. 605-610.

[12] J. Fritsch et al., “Multi-modal anchoring for human-robot interaction,” in Proceedings of Int. Conf. Robot. Auton. Syst., vol. 43, pp. 133-147, 2003.

[13] B. Lee, H. Kim, J. Choi, S. Kim and N. I. Cho, “DSP integration of sound source localization and multi-channel wiener filter,” in Proceedings of Int. Conf. Robot. Autom., May, 2010, pp. 4830-4835.

[14] H. D. Kim, J. S. Choi, M. Kim and C. H. Lee, “Reliable detection of sound’s direction for human robot interaction,” in Proceedings of IEEE Int. Conf. Intell. Robots Syst., Sep. 2004, pp. 2411-2416.

[15] B. Jun, J. Lee and D. Kim, “A novel illumination-robust face recognition using statistical and non-statistical method,” Pattern Recog. Letters, vol. 32, no. 2, pp. 329-336, 2011.

[16] S. Yun et al., “Proactive human search for the designated person with prior context information in an undiscovered environments,” J. Intell. Robot Syst., DOI 10.1007/s10846-011-9641-x, 2011.

[17] OpenNI Framework [Online]. Available: http://www.openni.org

[18] M. T. Palmer and G. A. Barnett, Progress in Communication Sciences, vol. 14, G. A. Barnett (Ed.), Norwood, NJ: Ablex, 1998.

[19] Y.-M. Kim and D.-S. Kwon, “A fuzzy intimacy space model to develop human-robot affective relationship,” in Proceedings of World Automation Congress, 2010, pp. 1-6.

[20] E. Hall, The Hidden Dimension, Anchor Books/Doubleday, 1966.

[21] P. K. Simpson, “Fuzzy min-max neural networks-Part 1: Classification,” IEEE Trans. Neural Netw., vol. 3, no. 5, pp. 776-786, Sep. 1992.

[22] B. Gabrys and A. Brugiela, “General fuzzy min-max neural network for clustering and classification,” IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 769-783, May 2000.