Research Article

Implementing a gaze control system on a social robot in multi-person interactions

Pourya Aliasghari[1] · Alireza Taheri[1] · Ali Meghdari[1] · Ehsan Maghsoodi[1]

Received: 15 March 2020 / Accepted: 15 May 2020

Abstract
Empowering a robot to direct its attention to the most appropriate target at all times during multi-party interactions is an interesting and useful challenge to establish natural communication between the robot and users. In this paper, implementing a social gaze control system suitable for multi-person interactions with a RASA social robot is discussed. This system takes some important verbal and non-verbal social cues into account, and at each moment enables the robot to decide socially at which human it should direct its gaze. The algorithm for the target selection has been enhanced, compared to past studies, by quantitating the effects of distance and orientation on grabbing humans' attention in addition to the inherent importance of each cue in communications based on the gaze behavior of a group of human participants. After this was completed, another group of volunteers were employed to examining the performance of the RASA robot equipped with this system. Their average gaze pattern was compared with the targets selected by the robot in a real situation, and their opinions on the sociability and intelligence of the system were recorded. We indicated that the gaze generated by the robotic system matched the average gaze pattern of the participants 76.9% in an 80-s real-life scenario. Moreover, the results of the questionnaire showed us that ~ 90% of the subjects felt that at times RASA was really looked at them with a quite high average score of 4.33 out of 5.

Keywords Human–robot interaction · Social robot · Attention modeling · Gaze control · Non-verbal behaviors · Multi-person interactions

1 Introduction

With the rapid advancements occurring in robotic technologies, social robots will play a prominent part in the future of our world [1]. These robots are designed to interact closely with people and have recently been utilized for variety of applications including education, therapy, and industry [2–4]. Accordingly, they should include autonomous capabilities to display socially acceptable behaviors so that human users will feel comfortable while interacting with them [1]. Among these features, the ability to direct their attention to the right target at the proper moment in multi-person interactions is of great importance to establish natural communication between the robot and human users. A mechanism called the gaze control system (GCS) is needed to identify humans in front of the robot, select the attention target, and continuously adjust the robot's head and eyes position.

In a follow up to our previous studies in developing and utilizing a cognitive architecture for teaching-assistant social robots [5], in this paper, we present the design, calibration and evaluation process of a gaze control system. This system which is inspired by the work done in [6] considers some important non-verbal cues as well as the location of the speaker in multi-person interactions to select automatically a single gaze behavior response. Before the

✉ Alireza Taheri, artaheri@sharif.edu | [1]Social and Cognitive Robotics Laboratory, Center of Excellence in Design, Robotics and Automation (CEDRA), School of Mechanical Engineering, Sharif University of Technology, Tehran, Iran.

Published online: 28 May 2020

potential targets compete with each other for that selection, the importance of the existing cues is modified based on the users' relative positions to the robot. To this end, the necessary coefficients which control the used rule-based GCS algorithm are extracted through a field study on virtual attention of a group of human participants. In order to improve the proposed algorithms in [6–8], we not only modified the core target-selection formulations, but also quantified the effect of proxemics and orientation of the users in the GCS, again based on the data collected from a group of human participants. This approach has not been explicitly used in the similar researches about gaze generation systems in the literature.

Finally, the gaze control system was implemented on RASA humanoid robot and was also tested in action. With this aim, another group of participants were recruited to evaluate the system performance by: (1) filling in a questionnaire and (2) comparing their own average gaze behaviors with the robotics system output in a similar social scenario. After all, by accomplishing this study, the head section of an under-development social robot converted to a lively and interactive tool, which was identified that promotes the attractiveness of the robot.

2 Background and related work

The term "Gaze" is often used as the head and eyes movements through which the center of human attention is moved to a specific target. Eye gaze as a component of social interaction is an important nonverbal cue in social interactions because humans can infer other's intentions from eye gaze [9]. Studies have shown that social robots that take advantage of a gaze control system are evaluated more positively by people [10]. For instance, Mutlu et al. [11] assessed how human gaze behavior implemented on a humanoid robot can create a natural and human-like manner of storytelling. They found that people recall the story told by the robot more effective when the robot established numerous mutual gazes with them. It has been also determined that by adding the ability of gaze shifting, the persuasiveness of the robot during storytelling will be promoted, while showing gestures without any gazing effects oppositely [12]. Another study has indicated that using eye gaze helps robot to improve the fluency and subjective experience during robot-to-human handover interactions [13]. Monitoring and maintaining user engagement is very important also in social robots used for teaching applications. If the students stop paying attention to the teacher, they will learn less [14]. Thus, in many cases, robots and virtual agents use eye gaze to maintain student engagement during teaching [15–17]. Furthermore, when a set of instructions is given by a robot to people, robot's non-verbal behaviors including gazing has shown to be so helpful on boosting the recall accuracy, especially when the task is complicated [18]. The robotic platform in our study is going to be used for teaching applications. Thus, these findings on the importance of a decent gaze control system working beside other modules of the robot have a big importance in this regard.

Researchers have built several computational tools for generating natural and acceptable robot eye gaze. These works generally focus on mathematical or technical aspects rather than the effects of the system on the interaction. However, these technologies may be evaluated by human users during the interaction. According to Admoni and Scassellati [14], one approach to developing gaze technology is to employ creativity to achieve an appropriate-looking behavior, regardless of the actual biological function in humans. Researchers have been able to directly design gaze behaviors using an understanding of psychology, and these behaviors are neither tied to underlying biological functions nor requiring large amounts of observational data.

Along with the development of new generation perception devices, such as Kinect, gaze control systems that work with 3D data have become more and more widespread. Zaraki et al. built a context-dependent social gaze control system implemented as part of the FACE humanoid social robot. Their system enables the robot to direct its gaze appropriately in multi-party interactions. The attention mechanism of the gaze-control system accounts for multimodal features such as proxemics, field of view, and verbal and nonverbal cues from the environment [6]. Yun proposed a computational model for selecting a suitable interlocutor for robots in a situation interacting with multiple persons. A hybrid approach was used for combining gaze control criteria and perceptual measurements for social cues. The perception part is aware of non-verbal behaviors based on the psychological analysis of human–human interaction. In addition, two factors of, physical space and conversational intimacy, were applied to the model calculation to strengthen the social gaze control effect of the robot [7]. Yumak et al. presented a gaze behavior model for an interactive virtual character with extra attention paid to estimating which user is engaged with the virtual character. The model takes behavioral cues, such as proximity, velocity, posture, and sound takes, into account to drive the gaze behavior of the virtual character [8].

3 Methodology

3.1 Robotic system

This GCS was implemented on a RASA humanoid robot designed specifically for teaching Persian Sign Language

(PSL) to children with hearing disabilities in the Social and Cognitive Robotics Lab., Iran [19, 20]. RASA has a total of 32 degrees of freedom (DOF), including the 3 natural DOF of a human neck. The robot's face is projected on an 8-in. screen located at the front of the robot's head with the ability to adjust the eyes direction. These features enable this robot to shift its gaze by using eye and head movements much like a human.

In order to respond to the robot's requirements for its applications as a teaching assistant, a cognitive architecture has been recently proposed [5]. The architecture has four main parts: Logic, Memory, Perception, and Action Units. The presented GCS was implemented into the designed cognitive architecture of RASA as shown in Fig. 1.

At the bottom of the Perception Unit, a Microsoft Kinect Sensor for Xbox One running the Kinect for Windows SDK 2.0 [21] equipped with a built-in microphone array captures data from the robot's environment and sends it to the ROS [22] environment via a web socket. A wide-angle time-of-flight camera together with an active IR sensor enables this version of the Kinect sensor to keep track of up to 6 persons and detect the position and orientation of 25 joints of each individual at the maximum rate of 30 frames per second. More high-level "Direction", "User Recognition", "Self-Awareness" and "Gesture Detection" boxes in Fig. 1 are acquired by processing the Kinect sensor inputs in addition to the head joint encoders' values. Then, the logic unit processes this information, which is stored in the robot's short-term memory (STM), on the basis of the model and algorithms defined in the long-term memory (LTM). This unit chooses the proper gaze target for the robot continuously for the usage of the action unit. All the mentioned units are implemented as a ROS package. Lastly, the Action Unit controls the Dynamixel MX-28 servo motors at the yaw and pitch DOF of the head of the robot and adjusts the eyes positions. The technical aspects of each unit of the system have been described in Sects. 2.1.1, 2.1.2, and 2.1.3. Figure 2 shows (the new version of) the RASA robot running the GCS.

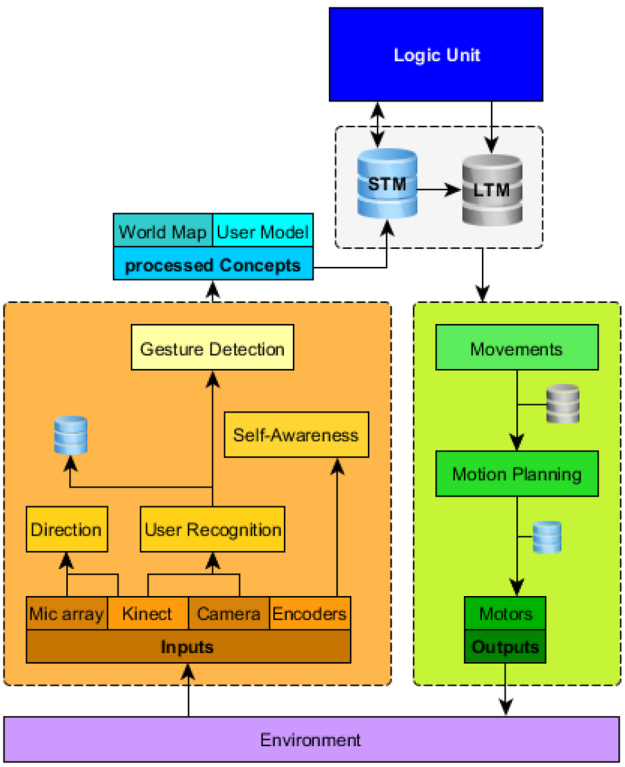

Fig. 1 The GCS incorporated into the RASA's cognitive architecture to identified the appropriate subject and shift its gaze using eye and head movements

3.1.1 Perception unit

The function of this GSC system depends on 3D visual and auditory information collected from the robot's field of view. Due to the fact that people often use a wide range of non-verbal social expressions to communicate with each other [23], the role of the Perception Unit is to detect and analyze certain non-verbal social signals that attract humans' attention. It should be also considered that people pay more attention to the individual they are listening to than to others in multiparty conversations [24]. Thus, using the skeleton tracking and sound source localization features of the Kinect sensor, the Perception Unit is aware of the presence of these social cues in each person in the robot's field of view. The social cues included: (1) speaking, (2) hand-waving, (3) pointing, (4) being engaged (paying attention to the robot), (5) entering, and (6) leaving.

The person who speaks is detected by comparing the incoming sound source angle with each person's head angle relative to the robot. The Audio Basics Kinect SDK

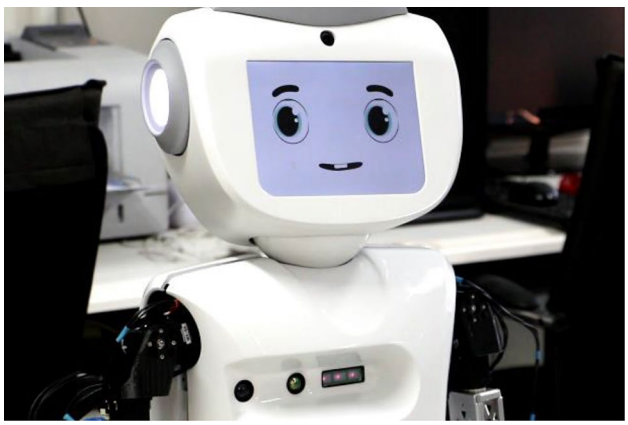

Fig. 2 RASA robot (version II) equipped with a Kinect sensor in its chest running the GCS. At this moment, the robot is looking at the person holding the camera

processes the sound signal received by the microphone array and finds its source angle using a triangulation algorithm. Environmental noises will be automatically ignored by the system after a while. It is assumed that the person located closed to the direction of the incoming sound is the speaker. Due to the limitations of the used sensor SDK, only one speaker can be detected at a time.

The Body Basics Kinect SDK recognizes the 3D body joint coordinates of up to six humans by using the information from both RGB and IR camera images. In order to determine whether a person is waving his/her hand or not, we calculate the kinematic energy of the arm, forearm, and hand links of each person whose wrist joint lays above his/her elbow joint. The coefficients as the mass of each joint to calculate this energy were estimated empirically. If the kinematic energy is higher than a pre-defined upper limit, it will be considered as hand-waving. The system can also detect if a person points to another person, by calculating the distance of the line passing through his hand to the other's heads. Also, the attention of each person to the robot (whether the user is engaged with the robot or not) is determined by comparing his head Euler angles with his orientation relative to the robot.

3.1.2 Logic unit

The Logic Unit should continuously select the robot attention target based on the Perception Unit inputs. The principles of the implemented gaze generation model are similar to the model presented by Zaraki et al. [6]. The main strategy for selecting the target that the robot should look at is to assign an elicited attention score (EA) to each person in the Perception Unit field of view. This score is calculated by considering the detected social cues and the position of every person in front of the robot as explained below.

It has been studied in social psychology, the physical distance between people during communications varies by the degree of their importance and intimacy [25, 26]. Therefore, social robots must use social spaces to establish better communication between robots and people, and to make human users feel at ease [27]. In a theory called proxemics, the spatial space surrounding persons is categorized into different zones (see Fig. 3, left semicircle). People show stronger reactions and pay more attention to others when the interaction is happening closer to them. This means that the above discussed social cues can elicit different attention if they happen closer to people. A similar phenomenon also exists concerning the angle between people while communicating. Human observers have a strong tendency to look more frequently around the center of the scene than around the periphery [28]. Social features

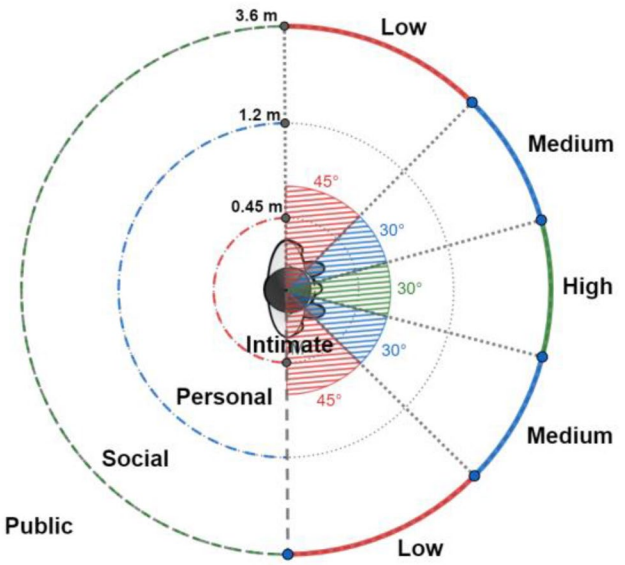

Fig. 3 Proxemics interaction zones: Intimate, Personal, Social, and Public in the left semicircle. The areas having a high, medium, or low relevance, depending on the angle from the center of human field of view, are shown in the right semicircle (inspired by the work done in [6])

collected in the more centered area elicit higher levels of visual attention. Taking these considerations into account, we suggest a different equation than [6] for calculating the elicited attention of each person (EA_i):

$$EA_i(t) = \left(\sum_{k=1}^{n} W_k\right) P(r) O(\theta) \tag{1}$$

In Eq. **(1)**, W_k is the weight factor corresponding to each social cue and n is the total number of social cues of ith person. Proxemics (P) and Orientation (O) coefficients adjust the importance of all social cues existing in each individual based on his or her distance and angle relative to the robot. $P(r)$ and $O(\theta)$ reflect the strength of elicitation of human attention for each proxemics and orientation zone as shown in Fig. 3. In this formulation, there is no need to normalize W, P and O factors, and EA_i values of all persons in the robot's field of view can be easily compared. It should be noted that when nobody is showing any social cues, a unit value is considered for the sum of W_k factors for each individual. Thus, the target selection strategy would be only on the basis of their closeness and orientation.

Next, we need to find the weight factor of each social cue relative to other cues (W), and find the $P(r)$ and $O(\theta)$ values corresponding to each area. The procedure for tuning the parameters is discussed in Sect. 2.2. Other equations governing the GCS are the same as those presented in the model in [6]. Consecutively, the Logic Unit selects

the largest value between EA_i values for all persons present in the robot's field of view and also a constant EA_V. If two or more are equal, the system selects the closer target. After passing through a moving average filter, the angle of the winner target is stored in the robot's short-term memory as the gaze angle (θ_g) for use by the Action Unit. It should be noted that EA_V is an empirical constant that is used to prevent the robot from only looking at people. In other words, by adding this virtual point to the competition, the robot is able to sometimes look at a defined point in the environment (e.g. in our case at the center) to behave more human-like. This intentional redirection away from the face of the audience is an important non-verbal cue that serves cognitive functions and it is useful for developing effective interactions between humans and robots [29].

Once a new human winner has been chosen, his/her score (*EA*) will be multiplied to a function called the habituation function ($H(\Delta t)$), and its time parameter (Δt) will be reset to zero:

$$H(\Delta t) = Peak \cdot Max\left(0, \left(1 - \frac{\Delta t}{\tau}\right)\right) \tag{2}$$

The constants *Peak* and τ in Eq. (2) are set according to [6]. This function decreases the winner's attractiveness for the robot gradually over time, allowing other people or the virtual point to attract the robot's attention. Thus, the robot does not look only at one person for a long time.

3.1.3 Action unit

Gaze shifts are accomplished by rotating the eyes and head in the same direction. The Action Unit computes the relative contributions of head ($\Delta\theta_h$) and eyes ($\Delta\theta_e$) movements towards a given gaze shift $(\Delta\theta_g)$ using the equations from [30] which are derived from statistical data.

$$\Delta\theta_g = \Delta\theta_h + \Delta\theta_e \tag{3}$$

In this model, for gaze shifts smaller than a threshold value ($\left|\Delta\theta_g\right| < \theta_t$), the head does not rotate and the robot looks at its target with only an eye movement. At any moment, θ_t is calculated as a function of the current eyes position (θ_{e_0}) as follows:

$$\theta_t = \left(-0.5\theta_{e_0} + 20\right) \times 0.56 \tag{4}$$

In Eq. (4), θ_t is positive if the eyes are initially rotated in the direction of the subsequent movement to limit the eye's rotation; otherwise, θ_t is negative. For $\Delta\theta_g$ larger than θ_t, both head and eyes rotate and the total head movement amplitude is derived from the equation discussed [6]. For our case, we found empirically that a simpler equation will result in better gaze shifts. In our system, for $\Delta\theta_g$ larger than θ_t, simply 20% of gaze shift will be performed by changing in eyes position while the remaining 80% will be carried out by a head movement. The desired head angle for performing the gaze shift is sent to a PID controller designed for the robot's neck actuators and the eyes angle sets the eyes position on the screen. On the robot's LCD face, the ability to blink was also added to help the robot show more realistic eye gazes. We used the eye-blink behavior model presented in [30]. They found that except for immediately successive blinks, the probability of a blink occurring during the interval $[t, t+1]$ (in seconds) decays exponentially with time from the last blink. They suggested a probability function of:

$$P\left(\text{blink in } [t, t+1] \text{ with last blink at } t_0\right) = 0.5e^{-0.12(t-t_0)} \tag{5}$$

The histogram of this probability in addition to blinking duration distribution is shown in Fig. 4.

As an extra capability regarding human–robot interactions, the Perception Unit is aware of hand-waving which enables the robot to detect if a person is making a "come here" gesture by waving both hands when they are completely upright. In this case, the Action Unit makes RASA move toward the person who called it using that gesture, and stops the robot at a safe distance (~ 1 m) in front of him/her making the interactions more intelligent, social, and appealing. However, it should be noted that the presented results of this study regarding the gaze control system's decisions were extracted from the robot in stationary situations.

3.2 Experimental structure

3.2.1 Calibrating the Attention's formulation

Two separate experiments were conducted to calibrate the mathematical model of Eq. (1) (i.e. find W_k, $P(r)$ and $O(\theta)$). A total of 23 volunteer Iranian students from Sharif University of Technology, consisting of 11 males and 12 females ranging in age from 19 to 29 years were asked to watch

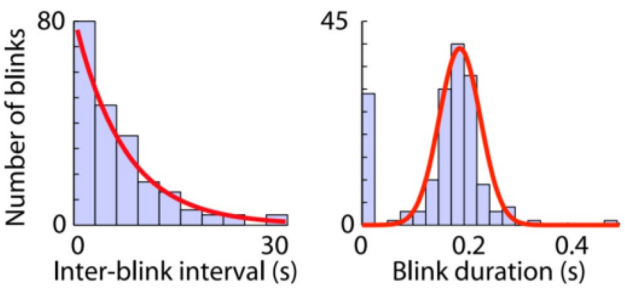

Fig. 4 Histogram of inter-blink intervals and blinking duration [30]

two separate 90-s animations. Each video showed two similar humanoid characters interacting with each other or telling a story. There was nothing in the scene except the two characters, without any colors, to keep the video as simple as possible. Thus, the viewer attention should not be distracted by anything else in the environment. This way is more similar to the robot's vision, which is only aware of the people. The animations were produced in Unity software by assigning captured body motions and speech to the characters using Cinema Mocap [31] and CrazyTalk [32] tools, respectively.

While watching the videos, participants' eye gazes were recorded using a webcam gaze recorder software called WebGazer [33]. WebGazer is an eye tracking tool that uses common webcams to record eye-gaze locations in real-time on the screen. Every volunteer was sitting at a distance of about half a meter to a screen and was asked to watch the video without any neck movement. As shown in Fig. 6a, we used a 40-in. wide-screen television instead of a regular computer monitor to extend the range of the participants' eye movements and enhance the accuracy of the captured data. The accuracy and precision of the captured visual targets while not significant were adequate for these measurements that the characters were placed as far as possible from each other and still be visible.

The first video of this experiment was used to determine the priority of each social cue to grab the participants' attention. In this animation, the characters presented every reasonable combination of the mentioned social cues. Initially, the first character enters the scene and shortly afterwards another one shows up. Then, one of them begins to give a talk while another person starts waving his hand. After that, the speaking person stops paying attention to the camera and looks away. When he looks back at the camera, he pointed to the other character. In these three samples of the happenings in the scenario, the social cue speaking was compared to hand-waving, being engaged, and pointing, respectively. The rest of the video showed the remaining rational combinations of the cues in the same manner. Figure 5a demonstrates screen-shots from the aforementioned moments of the animation. Finally, by comparing the average gaze shift of the participants in different situations, we are able to extract a score for each cue, as will be described in Sect. 3.1.

The second video was used to evaluate the effect of proxemics and orientation in humans' attention, and was played for participants after a short break. This animation consisted of five subscenes showing two characters giving the same lecture simultaneously. In the first three subscenes, one character was placed one step closer in the

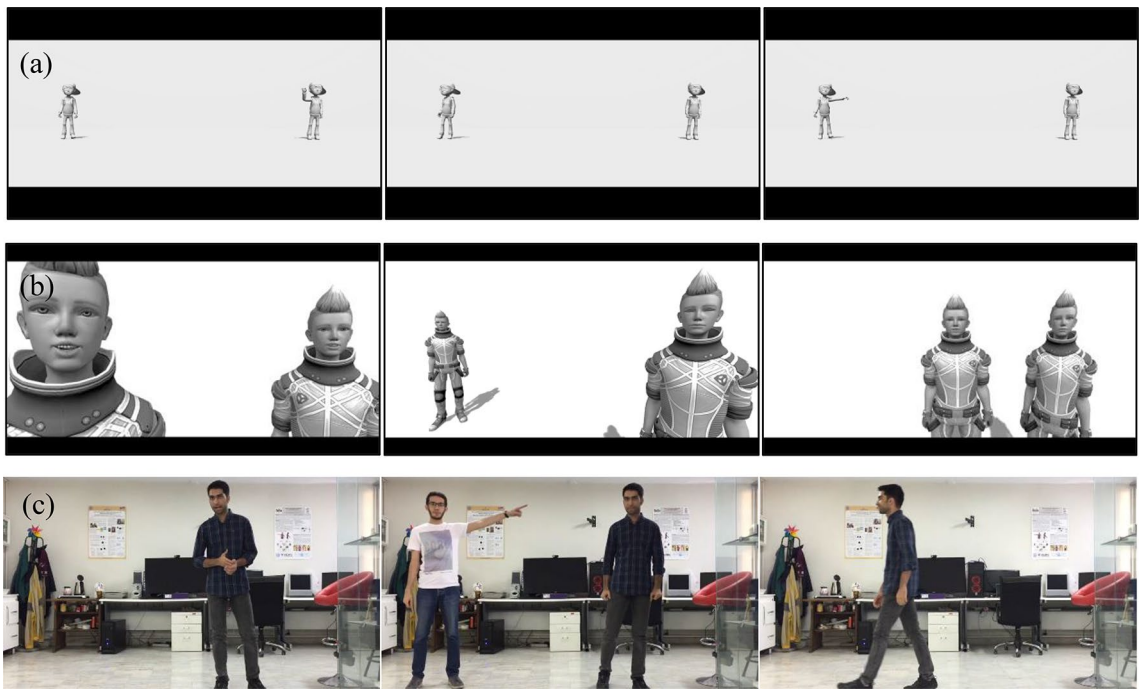

Fig. 5 **a** Screen-shots from the video used to evaluate social cues to attracting humans' attention. Speaking is compared to the other cues: (from left to right) hand-waving, engagement to the viewer, and pointing. **b** From left to right, the subscenes number 1, 2, and 4 in the animation played to measure the effect of proxemics and field of view. **c** Screen-shots from the real-life video recorded to evaluate the performance of the robotic system in gaze-shifting

Fig. 6 **a** Calibration of the Attention's formulation: a participant sitting in front of the screen and performing the calibration process in WebGazer, before the start of the gaze pattern recording. **b** Evaluation of the GCS: a volunteer interacting with the robot and one of the researchers to assess the performance of the robot

proxemics area than the other (see Fig. 5b). For instance, in the first subscene, one character is in the intimate zone and the other is in the personal zone. We calculated the average time fraction when participants were looking at the closer speaker as an indicator of their attention to this target compared to the more distant one. Since there are four proxemics areas, we had three different character placements in order to compare every two pairs of neighbor areas. The last two subscenes were devoted to measuring the effect of angle relative to the center of the view. At first, one speaker was exactly at the center and the other was at 30°. Then, each character moved 30° to the corners. Figure 5b shows three examples of these animations. It should be noted that a mirrored video was played for about half of the viewers to prevent the bias of looking at closer targets on only one side of the video. In addition, the main camera in the Unity environment was adjusted to show pictures with a field of view and perspective similar to the human eyes.

3.2.2 Evaluation of the gaze control system

To evaluate the performance of the system, in a separate experiment, we asked a new group of 21 volunteers to watch an 80-s video. The volunteers were also students of Sharif University of Technology ranging in age from 19 to 23 years old. This time, as presented in Fig. 5c, the video shows two real persons entering the room separately and then interacting freely with each other at various distances and orientations from the camera. This video was captured with the Kinect RGB camera, while the sensor served as the input unit of the GCS, simultaneously. Using the same procedure as the previous experiment, we used the participants' average gaze data to compare with the system output and assess its function in a real situation.

Following that, we let each participant interact freely with the robot and one of the researchers, while the GCS was the only active module running on the robot. The presence of one other person in the experiments had two goals. First, the participants could fully evaluate the capabilities of GCS in a multi-person interaction scenario (not a single human–robot interaction). Second, the researcher could give a list of different cues that the robot is aware of and direct the participants to only reflect their opinions on the GCS, not on any other aspects of the robot. Figure 6b shows some parts of this experiment.

Following their encounter with the robot, the participants were asked to fill in a questionnaire to rate their feelings and attitudes toward the robot with the GCS. The questionnaire was developed on the basis of standard questionnaires suggested in [34–36] and was anonymous, except for some general information such as gender, age, and academic year. First, the participants indicated their level of agreement with the four statements listed in Table 1, Part I used a five-point Likert scale to evaluate the Social Presence of the robot. The verbal anchors included in this section were: "totally disagree" (1), "disagree" (2), "neither agree nor disagree" (3), "agree" (4), and "totally agree" (5). Then, they were asked to answer four questions on the robot's sociability and intelligences, as shown in Table 1, Part II, on a 5-point scale: "very low" (1), "low" (2), "neither high nor low" (3), "high" (4), and "very high" (5). In the end, an anthropomorphism Godspeed questionnaire [34] was held to allow students to rate the human-like characteristics of the robot shown in Table 1, Part III using 5-point Likert scales.

Table 1 Questions asked of the participants in the experiment in order to evaluate the system

Part I: statements	
Q1	When interacting with the robot, I felt like I'm interacting with a real person. 1 2 3 4 5
Q2	It sometimes felt as if the robot was really looking at me. 1 2 3 4 5
Q3	I can imagine the robot to be a living creature. 1 2 3 4 5
Q4	Sometimes, the robot seems to have real feelings. 1 2 3 4 5
Part II: questions	
Q5	How well did the robot's movements adhere to human social norms? 1 2 3 4 5
Q6	How intelligent did the robot behave? 1 2 3 4 5
Q7	How well could the robot react to your actions? 1 2 3 4 5
Q8	How well could the robot understand your actions? 1 2 3 4 5
Part III: godspeed	
Q9	Fake 1 2 3 4 5 natural
Q10	Machinelike 1 2 3 4 5 humanlike
Q11	Unconscious 1 2 3 4 5 conscious
Q12	Artificial 1 2 3 4 5 lifelike
Q13	Moving rigidly 1 2 3 4 5 moving elegantly

4 Results and discussion

4.1 Calibrating the Attention's formulation

The first conducted experiment was to evaluate how intense each social cue can grab humans' attention. The final result is presented in Fig. 7. The plot shows the filtered average of the viewers' attention to the right or left character on a scale between − 1 and 1 on the vertical axis, respectively; and the time duration of the animation in seconds on the horizontal axis. The starting moment of each important occurrence in the scenario is marked separately at the side corresponding to each character in this figure. It can be seen that when the first character enters (on the left one), 100% of participant's looked at him for a while. By the arrival of the second character, 1.7 units of attention were grabbed by him, causing the line going from 0.9 at the second 6 to − 0.6 at the second 9. When the first character started to speak, he attracted 1.2 units back to himself. Using the same approach, we can interpret the whole chart.

We would like to find the W coefficients of Eq. (1). First of all, it was seen that all of the participants looked at the character for a while when he entered or left the scene. On average, people kept looking at the person who entered and left for about 3 and 3.25 s, respectively. This suggests it is more reasonable to force the GCS to look and follow the human targets when they are leaving or entering as a rule in the algorithm. Thus, we separated these two cues from the other social cues and excluded them from those we want to assign a W factor.

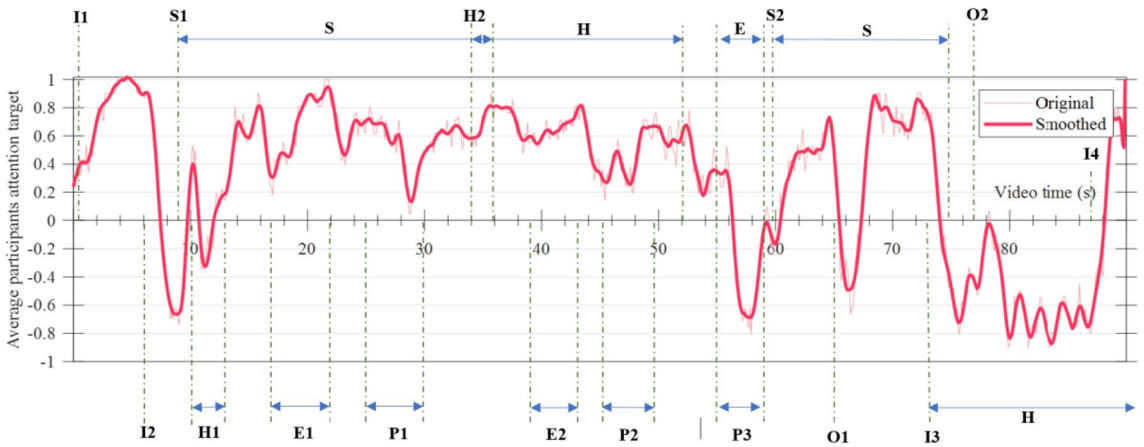

Fig. 7 Average participant attention target. The performed social cues by each character are indicated in his side with the following marks: S, Speaking; P, Being pointed; E, Being engage with the camera/robot (while the other character is not); H, Hand-waving

Table 2 presents when each social cue happened, how much it attracted attentions, and shifted the peaks of the Fig. 7. Among the four remaining social cues, the first and the second priority were "speaking" and "pointing to another person", with the average of 1.05 and 0.87 units of attention shift, respectively. The two remaining social cues ("hand-waving" and "engagement with the robot") with a factor of 0.60 and 0.55 units attracted less attention compared with the other cues. For a comparison, these results are very similar to the results found in Zaraki et al. [6]. They found that hand motion/body gesture cues have a weight equal to 0.65 of the weight factor of speaking. In the present experiment, pointing/hand-waving/engagement together can be interpreted as hand motion/body gesture, and have an average W factor of 0.67, which make the ratio of 0.64 relative to speaking. We found the leaving cue to be of the greatest important as opposed to Zaraki et al. who ranked it as the fourth priority. The small number of participants, cultural difference between the participants of the two studies, and the way the experiment was performed may be responsible for this contrast. However, entering was the most important cue in both studies.

Table 2 The results of the first experiment: social cues and the average attention shift of the participants each time the cue happened

	Attention peak shift			Avg. (W factor)
Speaking	S1: 1.2	S2: 0.9		1.05
Pointing	P1: 0.7	P2: 0.6	P3: 1.3	0.87
Hand-waving	H1: 0.8	H2: 0.4		0.60
Engagement to the robot	E1: 0.8	E2: 0.3		0.55

The second experiment determined the P and O factors as indicators of the strength of each proxemics and orientation area in attracting the humans' attention. As mentioned before, this type of quantitative analysis regarding the effect of proxemics and orientation has not been presented in similar works in the literature. Table 3 summarizes the findings of this experiment. In the subscene where one character was located in the Intimate zone and the other one in the Personal zone, participants look at the closer speaker about 61% of the time duration. In the two-remaining proxemics subscenes, almost the same distribution was observed with a small increase in the attention to the closest target when it was placed farther from the visitor. For the angle relative to the viewer, participants looked nearly 55% of the times at the character in the middle when the other one had a 30° eccentricity. When they see these characters with 30° more eccentricity, they payed attention 68% of the times to the more centered one. This observation showed that the difference in elicited attention between the High and Medium field of view zones is much higher than between the Medium and Low zones.

If we assign a unit P factor to the Public proxemics zone, the factor for the Social zone should have a value that makes a ratio of 1.6 [= 0.63/(1–0.63)] with the factor of the Social zone. Therefore, we consider the coefficients mentioned in Tables 4 and 5 as the importance of each proxemics and field of view area for Eq. (1).

4.2 Evaluation of the gaze control system

In this subsection, we would like to evaluate the GCS performance according to the participants' gazes in a defined social scenario. Moreover, the subjects' viewpoint extracted from the questionnaires are presented in the following.

Table 3 Average time fraction participants payed attention to the closer/more centered character compare to the other character

Comparison	Subscenes				
	Proxemics			Field of view	
	1 (Imitate and Personal)	2 (Personal and Social)	3 (Social and Public)	4 (High and Medium)	5 (Medium and low)
Average time fraction paying attention to closer/more centered character compare to the other	0.61	0.62	0.63	0.55	0.68

Table 4 The results of the second experiment: proxemics factors (P)

Proxemics zone	Imitate	Personal	Social	Public
	$r < 45$ cm	45 cm $< r <$ 120 cm	120 cm $< r <$ 360 cm	360 cm $< r$
$P(r)$	4.2	2.7	1.6	1

Table 5 The results of the second experiment: orientation factors (O)

Orientation zone	High $\lvert\theta\rvert < 15°$	Medium $15° \leq \lvert\theta\rvert < 45°$	Low $45° \leq \lvert\theta\rvert$
$O(\theta)$	2.5	2.1	1

First, the average gaze target of participants in watching the 80-s video is compared with the output of the GCS in Fig. 8, while the narrow blue line shows the target selected by the robotic system, the dashed blue line is the smoothed system attention pattern using a Savitzky–Goaly filter, and the red line is the filtered humans' average gaze behavior. This graph has been drawn using a procedure similar to Fig. 7 on a scale between −1 to 1, and shows that most of the time the GCS acts similar to the humans. From the 13 attention picks in the humans' gaze behavior (including the start and end points), the GCS followed 10 of them (i.e. 76.9%) which can be acceptable. The presence of the habituation function [Eq. (2)] sometimes prevents the system from shifting its gaze target rapidly, and may cause some differences between the robot's gaze and the humans' gaze, for example at time = 24 s. This function also brings the robots attention for some moments to the center, in second 61 for instance, while these random shifts do not appear in the average of multiple persons' gaze.

The second part of evaluating the performance of the robotic system was analysis of the results of the questionnaires. Table 6 presents the mean and standard deviations, minimum and maximum of the scores, and the percentages of positive (> 3) and negative (< 3) answers to each question of the survey. Cronbach's alpha internal consistency test was performed when reporting evidences based on some collections of questions.

According to Table 6, an equal number of students were positive and negative about the first statement (Q1), with the mean value of 3.00 (SD = .89) of answers

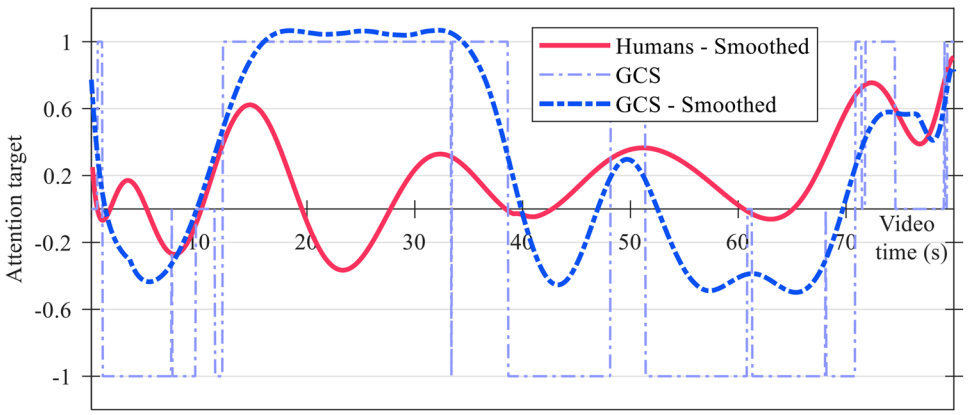

Fig. 8 Comparison of participants' average gaze with the gaze generated by the GCS

Table 6 The results of the questionnaires filled in by the 21 participants

	Mean	SD	Max	Min	POS (> 3) [%]	NEG (< 3) [%]
Part I: statements						
Q1	3.00	0.89	4	2	38	38
Q2	4.33	0.73	5	3	86	0
Q3	3.24	0.99	5	2	43	29
Q4	3.05	1.20	5	1	43	38
Part II: questions						
Q5	2.95	0.59	4	2	14	19
Q6	3.43	0.93	5	2	43	14
Q7	3.71	0.72	5	3	57	0
Q8	3.62	0.59	5	3	57	0
Part III: godspeed						
Q9	3.29	0.85	5	2	43	19
Q10	3.29	0.78	4	2	48	19
Q11	3.57	0.87	5	2	62	14
Q12	3.10	0.89	4	2	43	33
Q13	3.48	1.33	5	1	48	29

suggesting that they did not feel so much that interacting with the robot was like interacting with a real person. Participants attitude toward the robot being a living creature, having real feelings and ability to adhere to human's social norms (Q3, Q4, and Q5) were also average, at a value near 3. This could be anticipated since the robot was only performing eyes and head movements. There are lots of other features required to make the robot behave more social and alive. In the questions asking whether the interaction with robot could be perceived similar as with a real person or a living creature (Q1, Q3, and Q12) people with a realistic attitude towards the robots did not give a very high marks. These effects have caused the overall score of some questions to be not significantly high. All the mentioned questions (Q1, Q3, Q4, Q5, Q12) cumulatively (Mean = 3.10, Cronbach's $\alpha = .80$) indicate that although RASA with GCS cannot be perceived very much like a living creature, but it is not machinelike as well. As an important qualitative observation, we saw that most of the students expected the robot to react and make conversation when they said something meaningful, for example saying its name, but in our case, the robot could only detect the speaker regardless of the meaning of his or her words. Another significant observation was that when the participants were told that the robot is aware of pointing, they also pointed to objects in the laboratory and expected the robot to pay attention to them. However, the GCS was only able to detect humans and not objects.

On the other hand, with the highest mean value of 4.33 (SD = .73), 86 percent of participants gave a positive score to the statement that they felt the robot sometimes really looked at them (Q2). This shows the successful function of the Action Unit of the robot which adjusts the robot's head and eyes position, properly. We observed that starting from when people first moved (instead of standing unmoving) in front of the robot and saw how the robot followed them, they became much more excited than when they were first interacting without any movement. Students gave an average mark of 3.43 (SD = .93) to the intelligence of the robot in Q6. The Logic and Perception Units of the system were evaluated in Q7 and Q8, respectively. The average scores of answers to Q7 and Q8 are the second and third highest scores on the table with no negative answer at all. These two measures were found to have an acceptable internal reliability (Cronbach's $\alpha = .71$) to estimate the overall success of the Logic and Perception Units. When taking also Q2 and Q13 which asked about the quality and precision of movements (Action Unit) into the account, the four questions together with the average score of 3.76 (Cronbach's $\alpha = .71$) imply a near high performance of the entire system.

On the Godspeed questionnaire, the number of answers greater than 3 is more than those less than 3. However, as we mentioned about the artificial or lifelikeness of the robot (Q12), not all the participants could perceive the robot as lifelike. All in all, we observed a rather positive tendency from the participants toward aspects of the RASA robot related to the implemented gaze control system, which could be evidence of the subjects' opinions of the robot's acceptable attention behaviors. Our findings are in line with the results of other studies which also have revealed that showing a social eye gaze by the robot makes a positive impact on the interaction between the robot and human users in different tasks [12, 13, 18].

5 Limitations and future work

The small number of the participants as well as the complex non-homogenous patterns of people's gaze make it difficult to make strong claims on the generalizability of the coefficients calculated for the GCS in this paper. However, the similar trend of our findings to the other works alongside the heartwarming results observed through our questionnaires gave us an initial estimate about the acceptable performance of the modified implemented algorithm for the gaze control system. One of the serious limitations of this study (which also exists in similar previous researches) that could be investigated in future works is the lack of the robot to consider human's speech contents when evaluating the related coefficients in the algorithm/formulations. This is a phenomenon that happens quite normally in human–human interactions.

Moreover, some limitations and errors of the Kinect sensor, such as the occasionally loss of one person in the output by the sensor when the person is located near the edges of the Kinect field of view, are a distraction source for the GCS. Furthermore, in the case of some levels of uncertainty in skeleton tracking, for example due to unfavorable positioning or the body being partially covered, the output body joint coordinates will fluctuate rapidly and result in a malfunction of the Perception Unit in determining social cues. Due to the technological growth which decreases the concern of applying similar procedures in future studies, such limitations could be logically compensated for by using modern tools and algorithms [37].

6 Conclusion

In this study, we successfully implemented a gaze control system on the RASA teaching assistant social robot with the aim of making it more social and attractive. The system is an improved and modified version of a previous work done in this area. The Perception Unit of our

developed cognitive architecture is able to extract some high-level social features from the humans in front of the robot. Then the Logic Unit uses an attention control algorithm tuned by empirical data from humans' gaze pattern to find the most prominent target for the robot's gaze. Finally, the Action Unit performs the eyes and head shifts toward the people interacting with the robot in a natural manner based on the decision of the Logic Unit. Some physiological aspects of humans' verbal and non-verbal communications were considered for selecting the attention target. We also performed two extra data captures to quantify the effects of distance and field of view on attraction the humans' attention in comparison to the similar previous works.

Two different approaches were executed to evaluate the function of the GCS. First, the gaze generated by the robotic system was compared to the average gaze pattern of a group of students, and 76.9% matching between the gaze shifts of the robot and humans was observed in an 80-s real-life scenario. Second, each individual in the group was asked to fill out a questionnaire after being allowed to examine the robot by himself/herself. With an average score of 4.33/5, 86% of the participants felt like sometimes the RASA robot really looked at them. While we figured out that it is not very likely that people consider the robot a living creature, their average attitudes regarding the logic and the perception of the robot were among the highest scores in the survey.

Acknowledgements This research was partially funded by the "Iran Telecommunication Research Center (ITRC), http://en.itrc.ac.ir/". We also appreciate the Iranian National Science Foundation (INSF) for their complementary support of the Social & Cognitive Robotics Laboratory (http://en.insf.org/).

Compliance with ethical standards

Conflict of interest The authors Pourya Aliasghari, Alireza Taheri, and Ehsan Maghsoodi declare that they have no conflict of interest.

Ethical approval Ethical approval for the protocol of this study was provided by Iran University of Medical Sciences (#IR.IUMS.REC.1395.95301469).

References

1. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166
2. Zibafar A, Saffari E, Alemi M, Meghdari A, Faryan L, Pour AG, RezaSoltani A, Taheri A (2019) State-of-the-art visual merchandising using a fashionable social robot: RoMa. Int J Soc Robot. https://doi.org/10.1007/s12369-019-00566-3
3. Meghdari A, Shariati A, Alemi M, Nobaveh AA, Khamooshi M, Mozaffari B (2018) Design performance characteristics of a social robot companion "Arash" for pediatric hospitals. Int J Humanoid Rob 15(05):1850019
4. Taheri A, Meghdari A, Alemi M, Pouretemad H (2018) Clinical interventions of social humanoid robots in the treatment of a set of high-and low-functioning autistic Iranian twins. Sci Iran 25(3):1197–1214
5. Hosseini SR, Taheri A, Meghdari A, Alemi M (2018) Let there be intelligence! A novel cognitive architecture for teaching assistant social robots. In: International conference on social robotics. Springer, Cham, pp 275–285
6. Zaraki A, Mazzei D, Giuliani M, De Rossi D (2014) Designing and evaluating a social gaze-control system for a humanoid robot. IEEE Trans Hum Mach Syst 44(2):157–168
7. Yun SS (2017) A gaze control of socially interactive robots in multiple-person interaction. Robotica 35(11):2122–2138
8. van den Yumak Z, Brink B, Egges A (2017) Autonomous social gaze model for an interactive virtual character in real-life settings. Comput Anim Virtual Worlds 28(3–4):e1757
9. Emery NJ (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 24(6):581–604
10. Mutlu B, Kanda T, Forlizzi J, Hodgins J, Ishiguro H (2012) Conversational gaze mechanisms for humanlike robots. ACM Trans Interact Intell Syst (TiiS) 1(2):1–33
11. Mutlu B, Forlizzi J, Hodgins J (2006) A storytelling robot: modeling and evaluation of human-like gaze behavior. In: 2006 6th IEEE-RAS international conference on humanoid robots. IEEE, pp 518–523
12. Ham J, Cuijpers RH, Cabibihan JJ (2015) Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. Int J Soc Robot 7(4):479–487
13. Zheng M, Moon A, Croft EA, Meng MQH (2015) Impacts of robot head gaze on robot-to-human handovers. Int J Soc Robot 7(5):783–798
14. Admoni H, Scassellati B (2017) Social eye gaze in human–robot interaction: a review. J Hum Robot Interact 6(1):25–63
15. Johnson WL, Rickel JW, Lester JC (2000) Animated pedagogical agents: face-to-face interaction in interactive learning environments. Int J Artif Intell Educ 11(1):47–78
16. Szafir D, Mutlu B (2012) Pay attention! Designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 11–20
17. D'Mello S, Olney A, Williams C, Hays P (2012) Gaze tutor: a gaze-reactive intelligent tutoring system. Int J Hum Comput Stud 70(5):377–398
18. Admoni H, Weng T, Hayes B, Scassellati B (2016) Robot nonverbal behavior improves task performance in difficult collaborations. In: 2016 11th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 51–58
19. Meghdari A, Alemi M, Zakipour M, Kashanian SA (2019) Design and realization of a sign language educational humanoid robot. J Intell Robot Syst 95(1):3–17
20. Zakipour M, Meghdari A, Alemi M (2016) RASA: a low-cost upper-torso social robot acting as a sign language teaching assistant. In: International conference on social robotics. Springer, Cham, pp 630–639
21. Kinect for Windows SDK 2.0. https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn799271(v%3dieb.10). Accessed 10 Sept 2018
22. Quigley M, Conley K, Gerkey B, Faust, J, Foote T, Leibs J, Ng AY (2009) ROS: an open-source robot operating system. In: ICRA workshop on open source software, vol 3, no 3.2, p 5
23. Kees W (1972) Nonverbal communication: notes on the visual perception of human relations. University of California Press, Berkeley

24. Vertegaal R, Slagter R, Van der Veer G, Nijholt A (2001) Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 301–308
25. Hall ET, Birdwhistell RL, Bock B, Bohannan P, Diebold AR Jr, Durbin M, La Barre W (1968) Proxemics [and comments and replies]. Curr Anthropol 9(2/3):83–108
26. Hall ET (1966) The hidden dimension, vol 609. Doubleday, Garden City
27. Tapus A, Mataric MJ (2008) Socially assistive robots: the link between personality, empathy, physiological signals, and task performance. In: AAAI spring symposium: emotion, personality, and social behavior, pp 133–140
28. Tatler BW (2007) The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. J Vis 7(14):4
29. Andrist S, Tan XZ, Gleicher M, Mutlu B (2014) Conversational gaze aversion for humanlike robots. In: 2014 9th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 25–32
30. Itti L, Dhavale N, Pighin F (2006) Photorealistic attention-based gaze animation. In: 2006 IEEE international conference on multimedia and expo. IEEE, pp 521–524
31. Cinema Mocap - motion capture for Unity. https://cinema-suite.com/cinema-mo-cap/. Accessed 15 Dec 2018
32. CrazyTalk Interactive - Unity talking avatars
33. Papoutsaki A, Sangkloy P, Laskey J, Daskalova N, Huang J, Hays J (2016) Webgazer: scalable webcam eye tracking using user interactions. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence-IJCAI
34. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
35. Garrell A, Villamizar M, Moreno-Noguer F, Sanfeliu A (2017) Teaching robot's proactive behavior using human assistance. Int J Soc Robot 9(2):231–249
36. Heerink M, Kröse B, Evers V, Wielinga B (2010) Assessing acceptance of assistive social agent technology by older adults: the almere model. Int J Soc Robot 2(4):361–375
37. Ahmadi E, Meghdari A, Alemi M (2019) A socially aware SLAM technique augmented by person tracking module. J Intell Robot Syst. https://doi.org/10.1007/s10846-019-01120-z