A Case for the Design of Attention and Gesture Systems for Social Robots

Romain Maure, Erik A. Wengle(✉), Utku Norman, Daniel Carnieto Tozadore, and Barbara Bruno

Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland
{erik.wengle,barbara.bruno}@epfl.ch

Abstract. The success of social robots, even (or, especially) in use cases as simple as "manning" a booth to promote a product in a shopping mall, depends on their ability of interacting with humans in a timely, effective and enjoyable way. In this paper we present an attention system and a gesture system for use by an autonomous social robot in applications related to product promotion. Our attention system employs a modular approach and attention functions, to allow for rich run-time behaviours to arise from simple rules, while the proposed gesture system allows for tuning the gestures to convey different emotions and robot personalities. The two systems were tested in an experiment involving 790 participants, aiming to explore the attractive power of different robot behaviours.

Keywords: Social robotics · Attention system · Gesture system

1 Introduction

Imagine to enter a shopping mall. On the side, a young woman is standing next to a booth, to promote a new product among customers. Her job description is simple: 1) identify potentially interested customers and attract them towards the booth, 2) describe them the key features of the product, typically on the basis of a script learned by heart, 3) conclude the (short) interaction by giving them a sample, leaflet, or gadget, as a reminder of the product. Such a job is repetitive, tiring (often requiring people to stand still for long shifts), usually poorly paid, possibly demoralising (since most customers ignore the vendors, or treat them as nuisances) and even enforcing gender biases and stereotypes, as it typically employs young, good-looking women to attract more customers. These characteristics, which make the job unappealing for people, also make it a perfect task for social robots: repetitive, short, interactions are their *forte*, they have a natural talent for standing still and do not possess feelings that rude customers can hurt. Moreover, robots, still a rare sight in everyday life, arguably exert a strong attraction on humans, with the added benefit of not enforcing stereotypes. As an example, a recent study revealed that people who saw an advertisement

R. Maure and E.A. Wengle—Contributed equally to this work.

F. Cavallo et al. (Eds.): ICSR 2022, LNAI 13817, pp. 367–377, 2022.
https://doi.org/10.1007/978-3-031-24667-8_33

of a robot hotel service had a significantly higher purchase intention than those who watched a traditional hotel service advertisement [11].

Designing a social robot for product promotion poses a number of not-yet-fully-solved challenges, among which: 1) the robot needs to be able to identify potentially interested customers among the passers-by, 2) the robot needs to know *how* to successfully attract them to the booth and show them the product.

The first challenge requires the robot to be equipped with an *attention system* [4], allowing it to detect passers-by, identify the interested ones among them and hold their attention throughout the interaction. Early attention systems relied on psychophysical models of human attention, supported by neurobiological evidence [7]. Attempts at achieving human-like processing efficiency include VOCUS, which combines a bottom-up approach for the detection of elements of (possible) interest with a top-down extension for the identification of regions of (more) interest [3].

The second challenge requires the robot to be equipped with a *gesture* system [8], allowing it to perform gestures prior to and during the interaction, to attract a passer-by's attention and guide it towards the product on display. While it is established that the robots' non-verbal behaviour is correlated with their likeability [6], literature findings are not conclusive concerning what is the best behaviour for attracting attention. As an example, [9] suggests that multimodal behaviours draw more attention, but [10] reveals that they generate slower response times than unimodal behaviours.

In this article, we tackle the two above-listed challenges, with the overarching goal of contributing to the design and development of social robots for product promotion. We propose a fast, easily customisable *attention system* for social robots, allowing to display a natural, complex behaviour arising from the combination of simple features and a *gesture system* which allows for conveying different emotional and personality cues. Building on an experiment involving 790 passers-by, we investigate the success rate of different unimodal and multimodal behaviours in attracting the attention of passers-by.

2 System Architecture

2.1 Attention System

A key requirement for our attention system is to allow for the easy customisation of *what* is interesting and *how* to react to the detection of an element of interest. To this aim, our attention system relies on *events* and associated *attention functions*, which the system manipulates at run-time.

At the lowest level, the system relies on a portfolio of detectors of elements of interest (e.g., faces and objects), which work on the camera stream and provide in output, frame by frame, a 2D bounding box within the image space for each detected element. The goal of the attention system is to identify at all times, within the image space, the most interesting element to look at.

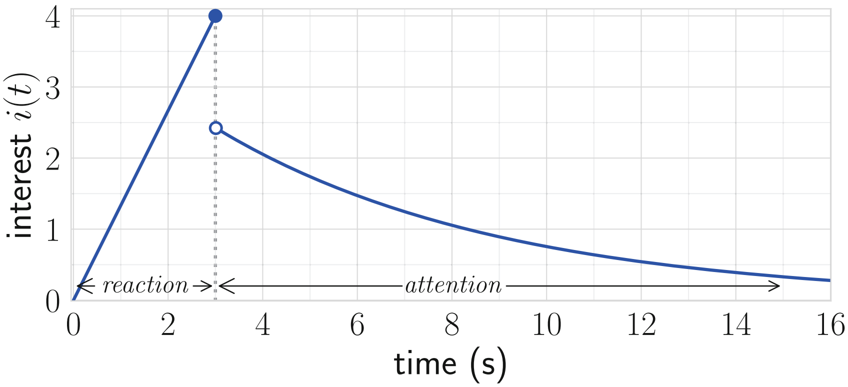

Fig. 1. The attention function, defined as $i(t)$ with $t_r = 3$, $I = 4$, $a = 6$ and $b = 0$.

At the core of the attention system there are the *events*. Each category of element of interest can be associated with one or more events, of three types: *detection* events are triggered when an element of that category first appears in the robot's field of view; *change* events are triggered when an element, already present in the robot's field of view, changes in one or more of its perceived properties; *departure* events are triggered when an element, previously present in the robot's field of view, disappears from the scene. As an example, in our experimental evaluation we consider human faces as elements of interest and define four events associated with face detection: 1) the *detection event* is raised when a new face appears in the robot's field of view; 2) the *movement event* is a change event raised when a tracked face moves within the robot's field of view; 3) the *head tilt event* is another change event, raised when a tracked face changes its orientation (e.g., to turn towards, or away from the robot); 4) the *departure event* is raised when a tracked face disappears from the robot's field of view.

Each event is supposed to raise the attention of the robot quickly, and lowering it slowly over time if no new event is triggered. An event is therefore associated with an *attention function*, which computes an interest score $i(t) \in \mathbb{R}_{\geq 0}$ for each time instant $t \in \mathbb{R}_{\geq 0}$ given as an input.

The function, shown in the graph of Fig. 1, is defined by four parameters:

- The reaction time t_r defines how quickly $i(t)$ should increase after activation and corresponds to the time for which $i(t)$ reaches its peak.
- The interest value I is the value of $i(t)$ at the peak. In our implementation, this parameter is set at run-time for tilt events, proportional with the angle between the person's gaze and the robot's orientation.
- The attention factor a defines how slowly $i(t)$ should decay after the peak.
- The base value b allows for chaining events (i.e., their attention functions) and shall be elaborated on in the following.

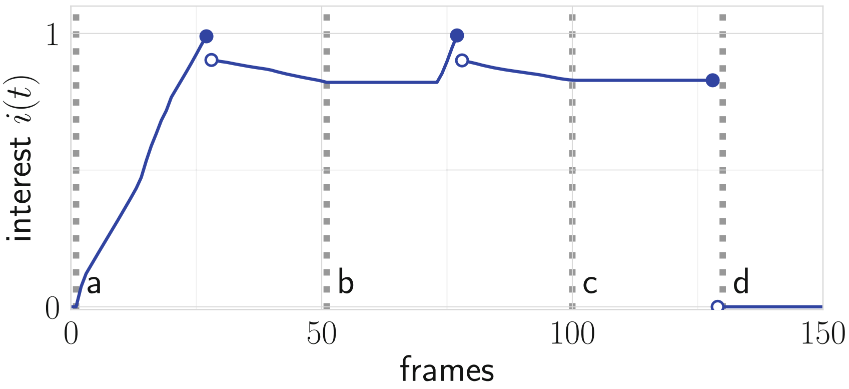

Fig. 2. Evolution of the interest score, for a person in following events: event (a) entering the robot's field of view, (b) turning to look at the robot, (c) looking away and (d) leaving. The score is computed by chaining the attention functions associated with the triggered events.

The function is then computed as follows:

$$i(t) = \begin{cases} \max\{b, I\frac{t}{t_r}\} & \text{if } t \leq t_r \\ e^{-t/a}\max\{b, I\} & \text{if } t > t_r \end{cases} \tag{1}$$

At run time, whenever an element of interest is detected, the corresponding detection event is triggered, and the associated attention function is activated and linked with the element's bounding box. The interest value $i(t)$ of the attention function thus represents the interest that the element within the bounding box has for the robot, at any given time. Later events related to the same element (e.g., if a detected face changes orientation) cause the previous and new attention functions to be merged as follows. Let us assume that at time t^* an element of interest generated by a previous detection event has an interest value $i_d(t^*) = i^*$, where $i_d(t)$ denotes the attention function associated with the detection event. If, at time t^*, the element experiences a change, the associated attention function is re-set as:

$$i(t) = i_c(t - t^*) \text{ such that } b_c = i_d(t^*) = i^* \tag{2}$$

where $i_c(t)$ denotes the new attention function associated with the change event and b_c is its base value. Note that $i_c(0) = b_c$.

As an example, the graph of Fig. 2 illustrates the evolution of the interest score associated with a face that (a) enters the robot's field of view, (b) turns to look at the robot, (c) looks away and (d) leaves.

The tracking of the elements of interests is done via the Euclidean distance metric, that first associates each bounding box detected at frame n with the closest bounding box detected at frame $n-1$ of the same type and then, by evaluating the difference between the previous and current element within the box, triggers the appropriate event, if any. To allow the aggregation of the interest score of multiple, possibly overlapping, bounding boxes, the robot's field of view is partitioned into a grid of equally sized cells. The interest score of each cell corresponds to the sum of the interest scores of all the bounding boxes it intersects with. Finally, at each frame, the output of the attention system is the coordinates of the cell with the highest interest value. If no bounding box is present, the interest score of all cells defaults to 0, and no cell is returned.

2.2 Gesture System

The gesture system endows the robot with the ability to perform *deictic*, or pointing, gestures [1], aiming to attract others' attention to a specific entity in the environment. In our use case, deictic gestures can be used to attract the passers-by's attention on the robot, as a form of attractive behaviour, as well as on the product on display, once a person has approached the booth.

The system is designed as a hierarchy of ROS2 services. At the lower level, a `goto` service relies on the robot's inverse kinematic model to allow the robot's end effectors reach any reachable pose within the workspace from their current pose. The service provides two types of motion trajectories, the *linear* trajectory and the *minimum jerk* trajectory, and allows for tuning the motion speed in terms of a percentage of the actuators' maximum velocity.

The `point_to` service relies on the `goto` service to endow the robot with the ability to point towards an entity or direction of interest. Given a point of interest $P = (x_P, y_P, z_P)$ in the Cartesian space whose origin O is located at the robot's torso, the direction of interest is defined by the vector $\overrightarrow{OP}$ and the final position P^* of the robot's end effector can be located anywhere along the $\overline{OP}$ segment. A parameter allows to specify the end effector's position along the segment, as a percentage of the farthest reachable point along the segment, defined as the intersection point between $\overrightarrow{OP}$ and the robot's workspace boundary.

The service provides several types of pointing gestures. The *basic pointing* gesture aims to mimic human's pointing gestures: the robot simply moves its end effector from its current position to P^*, oriented along the direction of interest. The *two-steps pointing* gesture is a combination of two successive movements. The robot first moves the end effector from its current position to a home position near the shoulder, then moves it along $\overrightarrow{OP}$ from this home position to P^*, keeping its orientation always aligned with $\overrightarrow{OP}$. Intuitively, this gesture type aims to attract viewers' attention on the direction of interest and make it easier for them to follow it. Lastly, the *self-pointing* gesture aims to attract viewers' attention on the robot itself. This gesture can be seen as a reversal of the *two-steps pointing* gesture, where the robot first points towards the human whose attention it intends to attract and then follows the direction of interest backwards, bringing the end effector close to itself.

The gesture system also allows to simultaneously move the robot's head during a pointing gesture, to increase the impact of the attractive behaviour. Finally, for the purposes of the experimental evaluation described in Sect. 3, a simple *waving* gesture has also been developed, to allow for a comparison between its attraction power and the one of other types of gestures and behaviours.

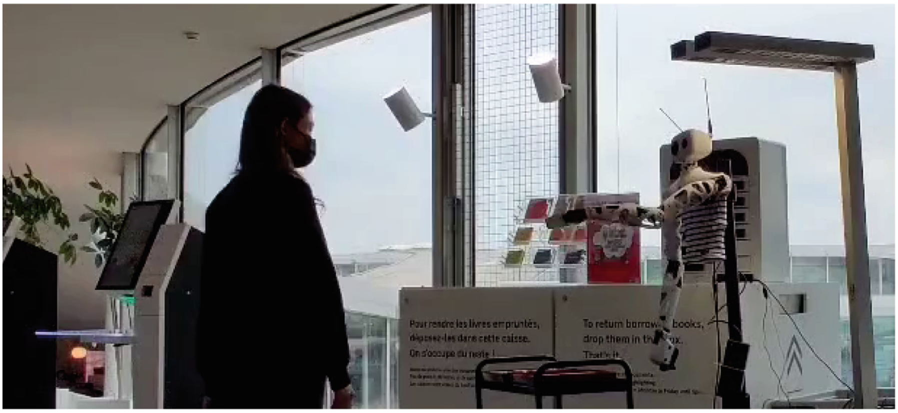

Fig. 3. Experiment setup.

3 Experimental Evaluation

3.1 Experiment Design

The primary goal of the experimental evaluation is to assess the effectiveness of various types of robot behaviours in attracting the attention of passing-by people. The experiment also allows for an evaluation of the attention system's performance in identifying and responding to events of interest, in a highly dynamic context and for events not necessarily defined a-priori.

We placed the robot at the entrance of the EPFL's library (Fig. 3), sideways with respect to the path followed by student to enter/exit the study area. This choice ensures a large variability in the type of situations presented to the attention system (people moving alone or in groups, walking in a hurry or leisurely strolling, etc.). Moreover, it guarantees that attracted passers-by visibly alter their behaviour w.r.t. not-attracted passers-by (e.g., by turning their head to the side, or by changing their path to come closer to the robot). We deem this setting to be a reasonable example of a product promotion use case, typically taking place in a crowded environment, in which people move with a purpose and where the booth is placed sideways to not block the customers' path.

The experiment is a natural experiment study [5], with random-like assignment of participants to the five conditions determined by the behaviour used by the robot to attract passers-by. The robot ran each condition for thirty consecutive minutes, changed manually by the experimenters. In the *control* condition, the robot only moves its head, as determined by the attention system. This behaviour is identical across all conditions. In the *waving* condition, the robot waves every time a new passer-by currently ignoring it (henceforth referred to as "target person") is detected by the attention system, while in the *self-pointing* condition the robot, upon detecting a new "target person", performs a self-pointing gesture. In the *speaking* condition, the robot's attractive behaviour is to utter the sentence “Hey, would you come over here?” while, lastly, in the *multimodal* condition the robot’s reaction to a new “target person” envisions the concurrent uttering of a sentence and execution of a self-pointing gesture. Our hypothesis is that the order in which conditions are listed above corresponds to the attraction effectiveness of the corresponding robot behaviours, i.e., it increases from the control to the multimodal.

The interaction unfolded as follows. The robot’s attention system continuously scans the environment at three frames per second. Once a person is detected, the attention system orients the robot’s head toward the point of maximum interest (if needed), checks if the person has already been detected or not and whether they are already moving and/or looking towards it (on the basis of their head orientation). A not-currently-tracked participant whose head is not oriented towards the robot is a “target person”. Upon identifying a new “target person”, the robot computes and stores a metric related to the distance between the human and itself and performs the attractive behaviour associated with the currently active condition. After executing the behaviour, the robot recomputes the distance to the person: if it is significantly smaller than the one computed before the attractive behaviour (concretely, if the ratio of the area of the current bounding box over the initial area of the bounding box at the detection event has increased beyond 125%), the robot invites the person to take a chocolate from a nearby box, by pointing to it, and utters a goodbye sentence. If the distance has not reduced, the robot murmurs “I blame myself”, lowers the arm used to attract the “target person” (in the conditions which imply arm movement) and resumes scanning its surroundings.

3.2 Evaluation Metrics

The simplest metric to measure the effectiveness of the different robot behaviours in attracting passers-by is to consider how many among them came close enough to the robot and picked up a chocolate from the chocolate box. Let us denote with P the number of all passers-by, and with P_{Choc} the number of people who came close enough and picked a chocolate. The *strong attraction rate* A_{strong} is thus computed as:

$$A_{strong} = \frac{P_{Choc}}{P} \tag{3}$$

Conversely, the *weak attraction rate* measures the number of times the robot elicited any type of response from a participant, even if they did not move closer. Let us denote with P_{Int} the number of passers-by who displayed at any time an interest in the robot, e.g., by coming closer, or verbally, or by hand-waving to it. The *weak attraction rate* A_{weak} is thus computed as:

$$A_{weak} = \frac{P_{Int}}{P} \tag{4}$$

Finally, we introduce the notion of *persuasiveness* of the robot to solely focus on those people that, while initially uninterested in the robot, changed their behaviour as a response to the robot's action. Let us denote with P_{Att} the number of passers-by not interested in the robot upon entering its field of view (i.e., not walking toward the robot and not looking at the robot) that the robot identified as "target people" and thus tried to attract, and with P_{React} those among them who responded in any way to the robot's action. The robot's *persuasiveness* Per is thus computed as:

$$Per = \frac{P_{React}}{P_{Att}} \tag{5}$$

3.3 Participants

The participants of the study are EPFL students and personnel, passing by the EPFL library on the day the experiment took place. No personal information was collected before, during nor after the experiment. The age of participants can be estimated to lie in the $[16, 65]$ range, with most participants likely below 30. Concerning gender, our participants' pool is likely skewed towards men, since women represent approx. 30% of all EPFL students [2]. The experiment took place on a Tuesday morning and lasted approximately 2.5 h (30 min per condition). A total of 790 people passed by the robot during the study. The average interaction time of the users which showed interest in the robot (P_{Int}) was 22.93 s (with a standard deviation of 15.12 s). A Kruskal-Wallis H test indicated no statistically significant differences between attractive behaviours for the duration of interaction of the users that showed interest ($H = 2.81$, $p = .59$).

4 Results

4.1 Attractive Behaviours Evaluation

Table 1 shows the attraction rates obtained for the various attractive behaviours.

Table 1. Attraction rates and robot liveliness.

Condition	P	P_{Choc}	A_{strong}	P_{Int}	A_{weak}	P_{Att}	P_{React}	Per
Control	104 (66)	9	8.65% (13.64%)	8	7.69% (12.12%)	0	0	0.00%
Waving	172 (86)	18	10.47% (20.93%)	26	15.11% (30.23%)	14	5	35.71%
Self-pointing	177 (84)	28	15.82% (33.33%)	25	14.12% (29.76%)	19	3	15.78%
Speaking	145 (63)	37	**25.52% (58.73%)**	22	15.17% (34.92%)	17	3	17.64%
Multimodal	192 (99)	45	23.44% (45.45%)	35	**18.23% (35.35%)**	7	4	**57.14%**

As expected, the table shows that the addition of attractive behaviours has a non-negligible impact on the number of passers-by being attracted towards the robot. In terms of the strong attraction rate A_{strong}, the best performance is achieved by the *speaking* condition (25.52%), closely followed by the *multimodal* condition (23.44%). Conversely, in terms of the weak attraction rate A_{weak}, the best performance is achieved by the *multimodal* condition (18.23%), while all other attractive behaviours display similar rates. Lastly, although the small number of people involved and the imbalance between the conditions do not allow to draw significant conclusions, the persuasiveness analysis suggests that the *multimodal* attractive behaviour (57.14%) is remarkably more persuasive than all others, with a 60% improvement over the performance of the second-best behaviour (*waving* with 35.71%).

4.2 Attention System Evaluation in the Wild

The face detection rates in the various experimental conditions are reported in Table 2: in total, 277 out of 398 (69.60%) people walking towards the robot were detected and therefore considered by the attention system.

Table 2. Number of passers-by per condition and face detection rates.

Condition	Total	Facing away	Facing towards	Detected	Detection rate
Control	104	38	66	37	56.06 %
Waving	172	86	86	61	70.93 %
Self-pointing	177	93	84	55	65.48 %
Speaking	145	82	63	40	63.49 %
Multimodal	192	93	99	84	84.85 %
TOTAL	790	392	398	277	69.60 %

We think that these detection rates are reasonably high, considering that (1) the experiment was performed over multiple hours, with different natural lighting conditions (2) the experiment took place in December 2021, when COVID-19 countermeasures still enforced the use of masks indoor. Indeed, most mis-detection were caused by the concurrent use of surgical masks, glasses, scarfs and/or hats, which left only a very small portion of the face actually visible. Lastly, on average, half of the people entering the robot's field of view were not facing the robot and thus impossible to detect. We hypothesise that, exactly as the robot had difficulties seeing people facing away from it, those people might have had difficulties in seeing the robot. To account for this possibility, in Table 1 we report, in parenthesis, the total number of participants whose moving direction was to walk towards the robot and the corresponding A_{strong} and A_{weak} rates. The considerations reported in Sect. 4.1 still apply.

The experiment generated a large amount of events (both foreseen and not foreseen at design time), where 4 scenarios are worth reporting.

In the first situation, the two people moved their gaze from the robot to each other throughout the interaction, which caused the robot to similarly shift its attention from the first person to the second, in a natural manner. Conversely, in the second situation the two people in the robot's field of view never looked away from it. This caused the robot to continuously move its gaze from one to the other, in a rather unnatural fashion. In the third situation four people entered the scene and moved towards the box of chocolate, completely oblivious to the robot. Upon noticing it, they thanked it for the chocolate and left. Throughout this interaction, the robot's gaze stayed stably on one single person, only moving towards a different person as the first one moved to take a piece of chocolate and looked away from the robot. This behaviour is deemed correct and natural. Lastly, in the fourth situation, two people entered the robot's field of view, with one way more interested in the robot than the other. In an attempt at catching the robot's attention, the first person caused the face of the second person to be repeatedly obstructed, and thus, by repeatedly appearing as a new face, be counted as more interesting for the attention system. This behaviour is deemed not correct, as the robot behaved differently from the person's expectation.

The good performance of the system even in unforeseen and "stressful" conditions (e.g., with four people surrounding the robot, or with a number of passers-by in the background) give us hope that the proposed approach could strike a good balance between speed, simplicity of setup and complexity and richness of the run-time behaviour.

5 Conclusions

In this paper we investigate the requirements posed on an autonomous social robot by the use case of manning a booth for product promotion in a shopping mall, a job which, while pervasive in our societies, has a number of drawbacks making it unattractive for humans. Such a robot requires a fast and simple-to-tune attention system, allowing it to identify potentially interested customers among the passers-by. To this end, we propose an attention system which allows for obtaining rich run-time behaviours as a combination of simple events, with associated attention functions. At the same time, this application requires the robot to be capable of gestures-based non-verbal interaction. To this end, we propose a gesture system allowing for the parameterisation of the robot's movements. In an experiment involving 790 participants, we compared the performance of five different uni-modal and multimodal behaviours in attracting the attention of passing-by people, showing that a multimodal behaviour combining an utterance with a self-pointing gesture not only elicits more responses from passers-by, but seems more persuasive (i.e., better able of attracting the interest of previously uninterested people) than all other behaviours.

Acknowledgments. This project was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 765955 (ANIMATAS) and as a part of NCCR Robotics, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_185543).

References

1. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: categories, origins, usage, and coding. Semiotica **1**(1), 49–98 (1969)
2. Equal opportunity office: gender monitoring EPFL 2019–2020. Technical Report, EPFL (2020). https://www.epfl.ch/about/equality/wp-content/uploads/2020/10/GenderMonitoring_EPFL_2020_en.pdf
3. Frintrop, S.: Computational visual attention. In: Salah, A., Gevers, T. (eds.) Computer Analysis of Human Behavior, pp. 69–101. Springer, London (2011). https://doi.org/10.1007/978-0-85729-994-9_4
4. Lanillos, P., Ferreira, J.F., Dias, J.: Designing an artificial attention system for social robots. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4171–4178. IEEE (2015). https://doi.org/10.1109/IROS.2015.7353967
5. Leatherdale, S.T.: Natural experiment methodology for research: a review of how different methods can support real-world research. Int. J. Soc. Res. Method. **22**(1), 19–35 (2019). https://doi.org/10.1080/13645579.2018.1488449
6. Lewandowski, B., et al.: Socially compliant human-robot interaction for autonomous scanning tasks in supermarket environments. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 363–370. IEEE (2020). https://doi.org/10.1109/RO-MAN47096.2020.9223568
7. Norman, D.A.: Toward a theory of memory and attention. Psychol. Rev. **75**(6), 522–536 (1968). https://doi.org/10.1037/h0026699
8. Van de Perre, G., De Beir, A., Cao, H.L., Esteban, P.G., Lefeber, D., Vanderborght, B.: Reaching and pointing gestures calculated by a generic gesture system for social robots. Robot. Autonom. Syst. **83**, 32–43 (2016). https://doi.org/10.1016/j.robot.2016.06.006
9. Saad, E., Neerincx, M.A., Hindriks, K.V.: Welcoming robot behaviors for drawing attention. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 636–637. IEEE (2019). https://doi.org/10.1109/HRI.2019.8673283
10. Torta, E., van Heumen, J., Piunti, F., Romeo, L., Cuijpers, R.: Evaluation of unimodal and multimodal communication cues for attracting attention in human–robot interaction. Int. J. Soc. Robot. **7**(1), 89–96 (2014). https://doi.org/10.1007/s12369-014-0271-x
11. Zhong, L., Sun, S., Law, R., Zhang, X.: Impact of robot hotel service on consumers' purchase intention: a control experiment. Asia Pacific J. Tourism Res. **25**(7), 780–798 (2020). https://doi.org/10.1080/10941665.2020.1726421