The Role of Expressiveness and Attention in Human-Robot Interaction

Allison Bruce, Illah Nourbakhsh, Reid Simmons

Carnegie Mellon University, Robotics Institute
5000 Forbes Ave
Pittsburgh, PA 15213
abruce@ri.cmu.edu, illah@ri.cmu.edu, reids@ri.cmu.edu

Abstract

This paper presents the results of an experiment in human-robot social interaction. Its purpose was to measure the impact of certain features and behaviors on people's willingness to engage in a short interaction with a robot. The behaviors tested were the ability to convey expression with a humanoid face and the ability to indicate attention by turning towards the person that the robot is addressing. We hypothesized that these features were minimal requirements for effective social interaction between a human and a robot. We will discuss the results of the experiment (some of which were contrary to our expectations) and their implications for the design of socially interactive robots.

Motivation

This research is situated within a larger project with the ultimate goal of developing a robot that exhibits comprehensible behavior and is entertaining to interact with. Most robots today can interact only with their creators or with a small group of specially trained individuals. If we are ever to achieve the use of robots as helpmates in common, everyday activities, this restricted audience must expand. We will need robots that people who are not programmers can communicate with. Much work is being done on the side of receiving input from humans (gesture and speech recognition, etc), but relatively little has been done on how a robot should present information and give feedback to its user. Robots need a transparent interface that regular people can interpret.

We hypothesize that face-to-face interaction is the best model for that interface. People are incredibly skilled at interpreting the behavior of other humans. We want to leverage people's ability to recognize the subtleties of expression as a mechanism for feedback. This expression is conveyed through many channels: speech, facial expression, gesture, and pose. We want to take advantage of as many of these modalities as possible in order to make our communication richer and more effective. We also hope to discover in a principled way which ones are most significant and useful for human-robot interaction.

Most day-to-day human behavior is highly predictable, because it conforms to social norms that keep things running smoothly. When robots do not behave according to those norms (for example, when they move down a hallway swerving around human "obstacles" rather than keeping to the right and passing appropriately), it is unpleasant and unnerving. In order to be useful in society, robots will need to behave in ways that are socially correct, not just near optimality within some formal framework.

Following the line of reasoning above, it would be easy to say, "if making a robot more human-like makes it easier to understand, then the best thing to do would be to make an artificial human". Clearly this is not feasible, even if it were the right approach. But it does raise some useful questions. How anthropomorphic should a robot be? Can it be a disadvantage to look "too human"? If we can only support a few human-like behaviors, which are the most important for the robot to exhibit?

Related Work

There has been a significant amount of work towards making software agents that are believable characters who exhibit social competence. The projects such as the Oz Project [Bates 1994] and Virtual Theater [Hayes-Roth 1998] created software agents that exhibit emotion during their interactions with each other and with human users with the goal of creating rich, interactive experiences within a narrative context. REA [Cassell 2000] and Steve [Rickel 2001] are humanoid characters that use multimodal communication that mimics the body language and nonverbal cues that people use in face-to-face conversations. While this work shares our goal of expressive interaction with humans, the characters are situated within their own "virtual" space, which forces people to come to a computer in order to interact. We are interested in developing characters that are physically embodied, capable of moving around in the world and finding people to interact with rather than waiting for people to come to them.

Work of this nature with robots is less developed than similar work with software agents, but it is becoming more common. There have been several museum tour guide robots designed recently to interact with people for educational and entertainment purposes. Nourbakhsh and collaborators at Mobot, Inc. address many of the same issues in human-robot interaction that we do in their discussion of their design decisions, along with offering suggestions based on their experiences with several robots [Willeke 2001]. However, their primary focus was on using entertaining interaction to support their educational goals rather than conducting an in-depth study of face-to-face social interaction. Minerva, another museum social robot, used reinforcement learning to learn how to attract people to interact with it, using a reward proportional to the proximity and density of people around it [Thrun 2000]. The actions that the robot could employ for this task included head motions, facial expressions, and speech acts. Their experimental results did not show that certain actions were more successful than others with any statistical significance other than that friendly expressions were more successful at attracting people than unfriendly ones.

Kismet is a robot whose sole purpose is face-to-face social interaction [Breazeal 1999]. It uses facial expressions and vocalizations to indicate its emotions and guide people's interaction with it. Kismet is specifically designed to be childlike, engaging people in the types of exchanges that occur between an infant and its caregiver. In contrast, our goal is to engage people in a dialog similar to an interaction between peers, using expressiveness to support our communicative goals. Another major difference between this project and ours is that Kismet is a head and neck on a fixed base. Even though Kismet is a physical artifact, like the software agents mentioned above, it relies on people coming to it in order to engage in interaction. While our robot is stationary for this particular experiment, one of the goals of this project is to explore the effects of an agent's ability to move around freely on the quality of social interaction with it.

System

Our testbed is a RWI B21 equipped with a laser range finder. A pan-tilt device is mounted on top of the robot. Either a camera or a flat screen monitor can be attached to the pan-tilt device. We use the screen to display the robot's face, which is an animated 3D model. We use the Festival (http://www.cstr.ed.ac.uk/projects/festival/festival.html) text-to-speech software package to generate speech and the accompanying phonemes, which we use for lip-synching. The use of a software-generated face allows us more degrees of freedom for generating expressions than would be possible if we designed a face in hardware.

The face design that we are currently using for our robot, Vikia, is that of a young woman. This initial design was chosen because we hypothesized that a realistic humanoid face would be easier for people to interpret the expressions of, and we wanted the robot to appear non-threatening. Later we hope to use and compare a number of different facial designs.

The facial expressions that Vikia exhibits are based on Delsarte's code of facial expressions. Francois Delsarte was a 19th century French dramatist who attempted to codify the facial expressions and body movements that actors should perform to suggest emotional states [Shawn 1963]. He exhaustively sketched out physical instructions for actors on what actions to perform, ranging from posture and gesture to fine details such as head position and the degree to which one should raise their eyebrows to indicate emotion. His approach, designed for melodramatic stage acting, is well suited for our application because it is highly systematic and focused on the communication of emotional cues to an audience. We focused our attention on the portion of Delsarte's work that dealt with facial expressions and head position [Stebbins 1886]. An animator implemented facial expressions for many of the more common emotions (happiness, sadness, anger, pride, shame) that Delsarte codified on the model for Vikia's face. For each emotion, Delsarte's drawings indicate the deformations that must be made to the facial features to express that emotion at varying levels of intensity. We created facial expressions for Vikia at 3 intensity levels for each emotion we implemented. These facial expressions are used to add emotional displays to Vikia's speech acts. The robot's speech and the animation of the head and face are controlled using a scripting language that allows for the sequencing of head movements and facial expressions with or without accompanying speech. This allows new dialog with accompanying facial expressions to be developed with relative ease. The script for the experiment was created using this system.

Vikia is equipped with a laser range finder, which we use to track the location of nearby people. The tracker runs at 8 Hz and is capable of tracking an arbitrary number of people within a specified area (set to a 10ft x 10ft square directly in front of the robot for the purposes of this experiment). The tracker detects roughly 70% of people walking past the robot in a crowded hallway. Occlusion often makes detection of all people walking together in a group impossible. The tracker will always succeed in detecting a group of people as the presence of at least one person, however, which is adequate for the performance of this task.

Experiment

The task that the robot performed was that of asking a poll question. There were a number of reasons for choosing that task. From an implementation point of view, it is a short and very constrained interaction, so it can be scripted by hand relatively easily. And the feedback that the robot needs to give in order to appear that it has understood the human's response is minimal (a necessity for now, as we have not yet integrated speech recognition into our system). Also, because people are egocentric and interested in sharing their opinions, we believe that we can expect a reasonable degree of cooperation from participants. Taking a poll contains many of the elements of interaction we are interested in studying (particularly the aspect of engaging people in interaction) without having to deal with the complexity of a full two-way conversation. We think that success at this task will indicate a significant first step towards longer, more complicated, and more natural interactions.

The robot's script for the poll-taking task ran as follows. First, the robot waits to detect that someone is in its area of interest. When the robot detects someone, it greets them and begins tracking them. All other people will be ignored until this person leaves. If the person stops, the robot will ask them if they will answer a poll question. If they are still there, the robot will ask the poll question, asking them to step up to the microphone (mounted on the pan/tilt head) to answer. If the person does not step forward, they will be prompted to do so 3 times before the robot gives up. Once the person steps forward, the robot detects that they are within a threshold distance, which the robot interprets as a response to the question. Because there is currently no speech recognition onboard the robot, this is the only available cue that the person has answered. The robot waits for the person to step back outside of this threshold, and then prompts them to step back. Once the person is outside the threshold, the robot determines that the interaction is over, thanks the person, and says goodbye. The interaction is then repeated with the next nearest individual.

We measured the number of people that reached each stage of the interaction with the robot. We observed the number of people that passed by, that the robot greeted, that stopped, that responded to the poll question, and that finished the interaction. The quantity that we analyzed from this experiment was the percentage of people who stopped out of the number greeted by the robot. This number provides a measure of success at attracting people to interact, rather than of the success at completing the interaction. Few people out of the number that stopped actually completed the interaction. The two major reasons for this were that people could not understand the robot's (synthesized) speech and that people did not step in close to the robot to answer, so the robot would prompt them to step closer. They would answer more loudly from the same distance and become frustrated that the robot could not hear them.

Experiment Design

We were interested in exploring the effects of the presence of an additional level of expressiveness and attention on the interaction. Without the face or the ability to move, the robot relies solely on verbal cues to attempt to engage people in interaction. Passersby receive no feedback on whether the robot is directly addressing them if there is more than one person walking by at a given time (this feedback is provided by the robot using the tracking information to turn towards the person its addressing). The face offers an additional level of expressiveness through the accompaniment of the speech acts by facial expressions (the output of the speech synthesis package that we use is not modulated to indicate emotion) and supports people's desire to anthropomorphize the robot. Would people find interaction with a robot that had a human face more appealing than a robot with no face? Previous work on software agents suggests so [Koda 1996] [Takeuchi 1995], even indicating that people are more willing to cooperate with agents that have human faces [Kiesler 1997].

The emotions that the robot exhibited during this interaction were all based on its success at accomplishing the task of leading a person through the interaction. Vikia greeted passersby in a friendly way. If they stopped, Vikia asked the poll question in a manner that indicated good-natured interest. If the person answered, Vikia stayed happy. But if the person didn't behave appropriately according to the script (for example, if they didn't come closer to answer or stayed too close and crowded the robot) Vikia's words and facial expressions would indicate increasing levels of irritation. This proved to be fairly effective in making people comply or attempt to comply with Vikia's requests. However, people who didn't step closer to answer and spoke louder instead often seemed perplexed and offended by the robot's annoyance with them.

The experimental design was that of a 3x2 full factorial experiment, a common experimental design used to determine whether the factors (variables) chosen produce results with statistically significant means and whether there is an interaction between the effects of any of the factors [Levin 1999]. The factors that we controlled for were the presence the face, having the robot's pan/tilt head track the person's movements, and the time of day (since we hypothesized that people may be more, or less, likely to stop depending on how crowded the corridor is, or how hurried they are).

Experiment Schedule

	4/16	4/17	4/18	4/19
11:15	T F	T no F	no T F	no T no F
11:30	no T F	no T no F	T F	T no F
2:15	T no F	T F	no T no F	no T F
2:30	no T no F	no T F	T no F	T F

Table 1: Schedule for the experiment carried out over 4 days (T is tracking, F is face).

Factors

Face. The robot's face in this experiment was an animated computer model of the face of a young woman displayed on a flat screen monitor that was mounted on the pan-tilt head of the robot. When the face was not used, it was replaced with a camera mounted on the pan-tilt head to give the robot a more traditionally robotic appearance.

Tracking. The robot uses a laser range finder to locate and track the position of a person's legs. Using this information, the robot can turn its "head" (either the face or the camera) towards the person that it is interacting with and follow their motion.

Time. This factor's value indicates whether a trial was conducted in the morning or the afternoon. This experiment was conducted over a period of four consecutive days with 2 trials in the morning and two in the afternoon. The robot was placed in a busy corridor in a building on the CMU campus.

Results

First an F-test was performed in order to determine whether the differences between the mean values for the factor values were statistically significant. A p-value of below .05 indicates statistical significance at the 95% confidence level. Only the factors "face" and "time" proved to produce statistically significant differences in the mean value of the percentage of people who stopped. This result indicates that there were no interactions between the factors that we measured in this experiment (e.g., the difference between the percentage of people who stopped to interact with the robot when it had a face and when it did not was the same regardless of the time of day, even if the more people stopped overall during the afternoon). More importantly, this result shows that whether the robot tracked passersby had no impact on the number of people who stopped to interact with it. This result was surprising because it violated our hypothesis that indicating attention by tracking a person with the pan/tilt head would increase people's engagement.

Source	Sum of Sqrs	Mean Square	F-Ratio	P-Value
MAIN EFFECTS				
A:Tracking	2.30E-04	2.30E-04	0.06	0.819
B:Face	0.103	0.103	25.05	0.001
C:Time	0.0306	0.0306	7.43	0.026
INTERACTIONS				
AB	6.28E-05	6.28E-05	0.02	0.905
AC	7.60E-04	7.60E-04	0.18	0.679
BC	0.0148	0.0148	3.59	0.095
ABC	2.30E-04	2.30E-04	0.06	0.819
Residual	0.0329	0.00411		

Table 2: F-tests of factors
Response variable: Percentage of people stopped

We have several hypotheses for why tracking did not have the effect on people's interest in interacting with the robot that we believed it would have. One is that there may be problems with our implementation of the person-tracking behavior. The robot does not start to track someone until they come within 10 ft of it from the front or side. It may be that the robot needs to start reacting to an approaching person when they are at a greater distance. Another issue with our implementation is that of latency. We limit the speed at which the pan-tilt head turns in order to avoid jarring the screen when the movement starts and stops. If a person is walking by relatively quickly, sometimes the pan-tilt head has trouble keeping up with their movement. It may be that we either need to increase speeds or anticipate the person's movement in order to improve tracking performance. Another possible reason that the tracking did not have an effect is that this type of movement might not be significant for this type of task. It may be that merely following a person's movement is not sufficient, and that less passive forms of motion, such as approaching the person the robot wants to interact with, are necessary. Yet another possibility is that this type of action does not make a difference at all in attracting people to interact with an embodied agent, no matter what the task. It may be that our assumption that indicating focus of attention with "gaze" is important for establishing contact is wrong, and that there is another nonverbal behavior that is more important for initiating interaction.

Source	Mean	Lower Limit	Upper Limit
Tracking			
yes	0.132	0.0838	0.181
no	0.14	0.0914	0.188
Face			
yes	0.216	0.168	0.264
no	0.0559	0.00738	0.104
Time			
afternoon	0.18	0.131	0.228
morning	0.0925	0.0439	0.141
Tracking, Face			
yes, yes	0.211	0.142	0.279
yes, no	0.0541	0	0.122
no, yes	0.222	0.154	0.291
no, no	0.0577	0	0.126
Tracking, Time			
yes, afternoon	0.169	0.1	0.238
yes, morning	0.0956	0.0269	0.164
no, afternoon	0.191	0.122	0.259
no, morning	0.0893	0.0207	0.158
Face, Time			
yes, afternoon	0.291	0.222	0.359
yes, morning	0.142	0.0737	0.211
no, afternoon	0.0693	6.12E-04	0.138
no, morning	0.0426	0	0.111

Table 3: Means with 95.0 percent confidence intervals
Response variable: Percentage of people stopped

Future Work

This work is in its preliminary stages, and there are numerous promising directions we hope to explore. In the short term, we plan to repeat the test with person tracking that responds to people when they are further away and uses their trajectory information to predict their future position. This will hopefully give us insight into whether the results that we saw are implementation dependent.

We also intend to run the experiment on the robot using different faces (such as male, animal, or cartoon) performing the same interaction, in order to study the effects of appearance on people’s reaction to the robot. Additionally, we plan to test people's reaction to less passive forms of robot motion, such as the robot approaching people whom it is trying to interact with.

Conclusion

We have performed an experiment on the effects of a specific form of expressiveness and attention on people's interest to engage in a social interaction with a mobile robot. The results of this initial experiment were surprising. They indicate that the person-tracking behavior used to indicate the robot's attention towards a particular passerby did not increase that person's interest in interacting with the robot as we had hypothesized it would. This raises a number of questions, both about our implementation and the assumptions that motivated it. In future work, we will continue to experimentally test our theories about what features and abilities best support human-robot interaction.

Acknowledgements

We would like to thank Greg Armstrong for his work maintaining the hardware on Vikia, Sara Kiesler for her advice on the experiment design, and Fred Zeleny for his work on the script and facial animations.

References

Bates, J. 1994. The Role of Emotion in Believable Agents. *Communications of the ACM* 37 (7), 122-125.

Breazeal, C. and Scassellati, B. 1999. How to Build Robots That Make Friends and Influence People. In *Proceedings of IROS-99*, Kyonju, Korea.

Cassell, J., Bickmore, T., Vilhjálmsson, H., and Yan, H. 2000. More Than Just a Pretty Face: Affordances of Embodiment. In *Proceedings of 2000 International Conference on Intelligent User Interfaces*, New Orleans, Louisiana.

Kiesler, S. and Sproull, L. *“Social” Human Computer Interaction, Human Values and the Design of Computer Technology*. Friedman, B., ed. 1997. CSLI Publications: Stanford, CA. 191-199.

Koda, T. and Maes, P. 1996. Agents With Faces: The Effect of Personification. *In Proceedings of the 5th IEEE International Workshop on Robot and Human Communication(RO-MAN 96)*, 189-194.

Levin, Irwin P. *Relating Statistics and Experiment Design*. Thousand Oaks, California. Sage Publications: 1999.

Hayes-Roth, B. and Rousseau, D. 1998. A Social-Psychological Model for Synthetic Actors. In *Proceedings of the Second International Conference on Autonomous Agents*, 165-172.

Rickel, J., Gratch, J., Hill, R., Marsella, S. and Swartout, W. 2001. Steve Goes to Bosnia: Towards a New Generation of Virtual Humans for Interactive Experiences. In papers from the 2001 AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment, Technical Report FS-00-04. Stanford University, CA.

Takeuchi, A. and Naito, T. 1995. Situated Facial Displays: Towards Social Interaction. *Human Factors in Computing Systems: CHI'95 Conference Proceedings*, ACM Press: New York. 450-455.

Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A.B., Dellaert, F., Fox, D., Haehnel, D., Rosenberg, C., Roy, N., Schulte, J. and Schulz, D. 2000. Probabilistic Algorithms and the Interactive Museum Tour-Guide Robot Minerva *International Journal of Robotics Research* 19(11), 972--999.

Willeke, T., Kunz, C. and Nourbakhsh, I. 2001. The History of the Mobot Museum Robot Series: An Evolutionary Study. In *Proceedings of FLAIRS 2001,* Key West, Florida.