(Not) Interacting with a Robot Photographer

William D. Smart and **Cindy M. Grimm** and **Michael Dixon** and **Zachary Byers**
Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, MO 63130
United States
{wds,cmg,msd2,zcb1}@cse.wustl.edu

Abstract

We have deployed a robot "photographer" at several events. The robot, Lewis, navigates through the space, opportunistically taking photographs of people. We summarize the different types of human-robot interactions we have observed at these events, and put forth some possible explanations for the different behaviors. We also discuss potential models for human-robot interactions in this constrained setting.

Introduction

In this paper, we describe our experiences with a robot photography system, deployed at several real-world events. The original idea behind the project was to develop a robot capable of navigating in an unaltered space, occupied by humans, taking pictures of them, in the manner of an event photographer. Initially, we saw the project as a means of getting undergraduates interested in robotics, and as an excuse to develop basic code (for navigation, camera control, *etc.*) that could be used in other, "real" projects. However, after deploying the robot and observing reactions to it, we have become much more interested in the questions of long-term autonomy, and human-robot interaction.

This paper gives a high-level overview of the robot photographer system, and gives some of our observations, based on several deployments at different venues. We discuss both the performance of the system, and the public reaction to it, in a number of environments. We also suggest some general trends in reaction and behavior that can be extracted from our observation, and possibly used to enhance the quality of the human-robot interaction.

The Robot

The robot, called Lewis, is an iRobot B21r mobile robot platform (see figure 1). It is cylindrical, stands approximately 4 feet tall, and has a diameter of about 2 feet. A camera is mounted on a pan/tilt unit on top of the robot. This camera sits at a height of about 5 feet from the ground, making it about the eye-level of a short human. The only sensor used in this project, other than the vision system, was the laser range-finder. This generates a set of 180 radial distance measurements over the front 180° of the robot, at a height of approximately one foot from the floor.

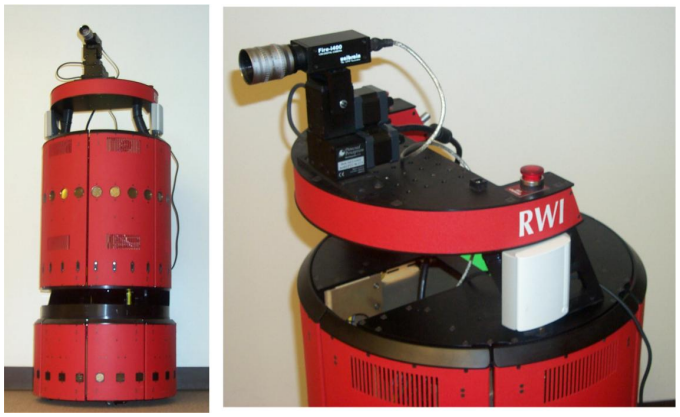

Figure 1: Lewis, the robot photographer.

The robot has a Pentium-III 800MHz processor, with 128MB of RAM, running the Linux operating system. When the system is running, the robot is completely autonomous, and all software runs on this computer. The robot communicates with the outside world over a 3Mb per second wireless ethernet link.

The robot has a set of navigation routines to support the photography application. The most basic navigation routine is obstacle avoidance. Since the robot must operate in environments where there are moving humans, this is, in some sense, the most important part of the system. Data from the laser range-finder are used to calculate a safe speed and direction of travel. If there are too many people close to the robot, then it will stop, and possibly turn around and move off in a different direction. The safety thresholds built into this layer cannot be over-ridden by any other subsystem.

The most basic navigation mode that the robot has is a random walk. The robot tends to move towards areas that appear to have fewer people in them. This mode turned out to be the most useful when the robot was operating in crowded environments. When we attempted to use more purposeful navigation, as described below, in crowded situations, the robot spent most of its time

avoiding people, and often found it difficult to get to the goal point in a reasonable time. In fact, it often took long enough to get to the goal that people had moved, and the photograph opportunity had gone. In a crowded environment, there are many potential opportunities for taking good photographs, and it seems that randomly wandering through the crowd is good enough.

For more sparsely-populated environments, however, random wandering is less likely to result in good photographs. In this situation, we use information from the photography subsystem to generate an objective function over the space that represents the expected quality of a picture taken from that point in the space. The objective function takes into account the distance from the robot to the possible photograph location, the chances that the subject will move before the robot will get there, how easily the robot can get there, and factors affecting the quality of the final picture (occlusions, background, *etc.*). The robot then drives to the highest-valued location, and takes a picture. This system worked well when there were only a few possible subjects in the environment, especially when they were clustered in small groups (which leads to only a few good positions from which to take a photograph). However, in most of the deployments, there were too many people to use this approach, and we relied on the random walk strategy.

In addition to the navigation routines, the robot had a localization subsystem, which used a large colored landmark hung from the ceiling. The robot localized by finding this landmark, and using its apparent size and elevation to estimate its distance. This localization was used at the SIGGRAPH deployment (see below) to avoid the robot traveling too far away from out booth. It was not needed at the other two deployments, since they were in closed rooms. More details of the navigation and localization strategies are available in the paper by Dixon, Grimm, and Smart (2003).

Robot Photography

The full details of the photography system are beyond the scope of this paper. Instead, we focus on the effective behavior of the system, and refer the interested reader to the paper by Dixon, Grimm, and Smart (2003). The robot is continually evaluating the images from the video camera, looking for possible skin blobs that might be faces. We coordinate this data with the laser scanner. Given a potential face blob in the image we use the laser scanner to determine how far way the "legs" under the face are. Using basic triangulation we can estimate a height for the face blob. Anything between four and seven feet tall is labeled as "human", and therefore a face.

Our system is designed to return more false positives than false negatives — we prefer not to miss faces, even if it means taking pictures of flowers. In well-lit areas where the walls are not "skin colored" (in the reddish range) we have little difficulty distinguishing faces, although there are usually non-people elements in the environment that the robot photographs fairly regularly.

The robot has a system in place to frame faces using some simple rules from photography. This system moves the camera and adjusts the zoom until a "good" picture is found, or until enough time has passed and it just snaps a picture. This last behavior was put in to make sure that the robot takes a picture, even a bad one.

The failure modes of the photography system are twofold. First, a person may be "invisible" to the robot if they are standing in the dark or in front of an area that is skin-colored.[1] Second, because of vagaries in the system, the robot ends up zooming in to frame a shot, loosing the "face" in the image, then zooming back out, and so on.

Deployments

We have deployed the robot photographer system at a number of events. In this section, we briefly describe the more important deployments, and discuss the key factors that were different between them. These differences help explain both the level of performance at each event, and also the types of interactions observed between the public and the robot. At the time of writing, the three most significant deployments of the robot photographer system are at a major computer graphics conference, at a science journalist meeting, and at a wedding reception.

SIGGRAPH 2002 The first major deployment of the system was at the Emerging Technologies exhibit at SIGGRAPH 2002, in San Antonio, TX. The robot ran for a total of more than 40 hours over a period of five days during the conference, interacted with over 5,000 people, and took 3,008 pictures. Of these 3,008 pictures, 1,053 (35%) were either printed out or emailed to someone.

The robot was located in the corner of the exhibit space, in an open area of approximately 700 square feet. The area was surrounded by a tall curtain, with an entrance approximately eight feet wide. Other than a small number of technical posters, the robot was the only object of interest in the space. Light was supplied by overhead spotlights, and three large standing spotlights in the enclosed area.

The people interacting with the robot were mostly computer graphics professionals, with a small number of local residents unaffiliated with the conference.

CASW Meeting The second major deployment was at a meeting of the Council for the Advancement of Science Writing (CASW), which took place in the dining room of the Ritz-Carlton hotel, in St. Louis, MO. The robot operated in an unaltered area of about 1,500

[1]Our robot is "color-blind" in the politically-correct sense. In the color space we operate in, YUV space, all races lie in the same reddish area and differ only in intensity.

Event	Photos per Person	%age Requested
SIGGRAPH	0.60	35%
CASW	1.47	11%
Wedding	1.17	2%

Table 1: Number of photographs taken per attendee, and percentage of photographs printed or emailed, for each of the three deployments.

square feet, as an evening reception took place. The robot shared the space with the usual furnishings, such as tables and chairs, in addition to approximately 150 guests, mostly science journalists. The robot operated for two hours, and took a total of 220 pictures. Only 11 (5%) of these were printed out or emailed by the reception guests, although several more were printed and displayed in a small gallery.

An Actual Wedding The system was deployed at the wedding reception of one of the support staff in our department. At this event, it ran for slightly over two hours and took 82 pictures, of which only 2 (2%) were printed or emailed. The robot shared a space of approximately 2,000 square feet with 70 reception guests, some of whom were dancing. Most of the people at the reception had never seen a mobile robot before, and many had only limited direct experience with any kind of computer technology.

Observations

The project is still very much ongoing, so we do not have many concrete results yet. However, we have made some preliminary observations of how people interact with the system, based on a number of real-world deployments. Although most of our evidence from these deployments is anecdotal, we believe that there are some general trends that can be extracted.

Interest in the System

The number of photographs taken per attendee, and the percentage requested (actually emailed or printed out), is summarized in table 1. It is clear that at events where the robot is the center of attention, more of the pictures are either printed out or emailed. Although this seems fairly obvious, it does show that the robot is largely ignored at events where it is only a peripheral character, such as the wedding reception. There was also a strong correlation between how much people stood and watched the robot at work and how many pictures were requested.

Interest in the output of the system also seems to be correlated with the technical sophistication of the people interacting with it. It seems that more technically savvy people are more interested in both the pictures taken by the robot, and also by the underlying technology. In all three deployments, a team of four students were present, to run the system, and to field questions. At SIGGRAPH they were constantly giving technical explanations of the system, while at the wedding reception, they only had to field a few, generally non-technical questions. The CASW meeting fell somewhere in between these two extremes.

At the wedding and the CASW meeting people were primarily interested in other activities, mostly talking to each other, and after a few questions went back to talking to each other. We noticed this phenomenon at SIGGRAPH as well, which was surprising — people who arrived in groups would, after asking questions, often stand around and talk to each other for awhile before moving on to the next exhibit.

Bimodal Interactions

One of the goals of the project was for the robot to take candid pictures of people, not just front-on shots. This requires that the subjects of the photograph not be attending to the robot. We were initially concerned that, since the robot is large and bright red, there would be a problem with it being the center of attention. However, this turned out not to be a problem. At SIGGRAPH, when a person first entered the space they tended to either deliberately stay away from the robot (reading posters or standing with their friends), or walk up to the robot and try to get its attention (by waving at it, or purposefully standing in front of it). Once the initial reaction wore off, people tended to form small social groups, and ignore the robot except for brief glances when the robot approached them.

We also observed this behavior at the CASW meeting, and at the wedding. At both of these events, however, the robot was not the reason that the people were at the event, so it is natural that they would pay less attention to it. However, it is somewhat surprising that people with little previous exposure to robots would be so blasé about one moving about close to them.

From our observations, this task seems to have a bimodal interaction pattern. Either the robot should be directly attending to someone who is trying to get its attention, or no one is paying attention to it and it should try to blend into the background. This implies that the interaction mode that the robot is in should be driven by the humans in the environment. If a human tries to engage the robot (see below), it should interrupt what it is doing and attend to them. Otherwise, it should try to be inconspicuous.

We hypothesize that people tended to ignore the robot partly because it made no attempt to externalize its state, and partly because it moves relatively slowly and smoothly. People are more likely to attend to things in the environment that are more "human-like", that is, those things that give behavioral cues similar to those that people do. For example, making eye contact or a rapid hand movement to get attention. The robot does not engage in this behavior, or respond to it, and is classified as a (moving) household appliance, and largely ignored.

Figure 2: Example photos the robot took. Images on the left are examples of poor composition, images on the right are some of the better ones. The camera used at SIGGRAPH was a 640x480 video camera, the others were taken with a Kodak digital camera, mounted on top of the video camera.

In the environments in which the robot was meant to be inconspicuous it blended into the background surprisingly well. As a result, many more candid shots of people talking and interacting with each other were taken, which was one of the goals of the project. At SIGGRAPH, when people were attending to the robot much more strongly, the great majority of shots were frontal, semi-posed shots, often with a "deer in the headlights" quality. Most of the candid shots occurred when the subject was being distracted by one of the students who was explaining how the robot worked. Figure 2 shows some example pictures taken at the different events.

Externalizing the Robot State

We found that when people are actively attending to the robot, they are much more comfortable when they have some idea of what it is doing. The most obvious example of this is when they are posing for a picture. In the system fielded at SIGGRAPH, there was an audible signal, sounding like a film camera shutter triggering, when a picture was taken. However, this turned out not to be loud enough for most people to hear over the background noise. The resulting lack of feedback led to some confusion about whether or not a picture had been taken. When a human photographer takes a picture, there is an obvious cue: they take the camera from in front of their face, and make eye contact with the subject. The robot cannot do this, because the camera is in a fixed position. This means that we must rely on some other signal that something has happened. We thought that adding in a more readily identifiable signal would alleviate this problem.

We added a professional flash to the camera for the second deployment at the CASW meeting. After some experimentation, we decided that the flash was not needed for the event, and set it up to trigger slightly *after* the picture had actually been taken. Although it was not used to take the photograph, it did provide a concrete signal to the subject that something had happened. This worked very well for shots where the subjects were attending to the robot, but proved to be distracting for the more candid shots. People involved in a conversation are less happy when a flash goes off near them unexpectedly. Again, this is an example of the bimodal nature of the interactions people had with the robot. In some cases, we want to externalize the state, since people are paying attention, and want to know what the robot is doing. In other cases, there is an advantage in not externalizing the state, since we want the robot to be inconspicuous.

Another problem subjects had was knowing when the robot was lining up a shot, when it was simply navigating through the crowd, and when it was actively trying to localize. Some of these behaviors (such as localization) could not be interrupted. This often led to frustration, especially at SIGGRAPH, when subjects were trying to get their pictures taken by standing in front of the robot when it was trying to localize. The robot could not see the landmark, and was unable to break out of the localization mode.

There were multiple suggestions from attendees about adding in feedback about the state of the robot. This was probably the single most common comment — people wanted the robot to say "cheese", or show a "birdie" picture when it was about to take a picture.

We deliberately designed the robot to take pictures relatively frequently if it suspected there was a human in the shot, even if it currently had no "good" composition. This alleviated some of the frustration experienced by participants, because the robot almost always took their picture eventually.

One problem with externalizing the robot state is that the robot often needs to do something that a human would not. For example, a human does not need to localize using landmarks on the ceiling (at least for this sort of task). Therefore, we do not believe that there is an easy way to externalize this to subjects, without having a speech output that says "Please stand aside. I need to look at the glowing ball hanging from the ceiling." This, in itself, may raise questions from the subjects. Why does the robot need to do this? Although this raises the possibility of engaging the subject in a dialog (and all of the hard research problems that implies), we do not believe that it is a completely satisfying solution. In particular, such an utterance begs the question "What happens if I stop the robot from doing that?" In our experience, people attending closely to

the robot tend to be technically curious about it, and willing to interfere with its operations (usually in a mild way) to see what happens.

Expectations and Intelligence

When directly interacting with the robot, several people tried to catch its attention by waving at it. The algorithms that we currently have in place have no way of detecting this, and this has often led to some frustration. It seems that when interacting with the robot, people have the expectation that it will react in a "human-like" way to stimuli.

Subjects seemed to apply a mental model of behavior to the robot in an attempt to explain its behavior. Again, this seemed to be bimodal. Those engaged by the robot seemed to use a human model, and attempted to interact with the robot in very human-like ways, waving and talking to it. When they were ignored, some subjects just gave up and left, while others continued to signal to the robot, often resorting to "baby talk", coordinating hand waving with simple, emphasized words. Eventually, they gave up too, but seemed somewhat more confused by the lack of response than those who gave up quickly.

Those subject who were not engaged by the robot initially seemed to quickly classify it as an appliance, or "another broken machine", and had no trouble in walking away.

When the robot did react as expected, due to some lucky coincidence, people tended to regard it as being "more intelligent" in some sense. This seems especially true of the camera motion. In the cases where someone waved at the robot, and the pan/tilt unit pointed the camera in their general direction, this seemed to deepen the interaction. It seems reasonable to suppose that, since eye contact is so important in human-human interactions, eye-camera contact will be similarly engaging in human-robot interactions.

In fact, one of the most compelling interactions was the result of our lack of code optimizations. When looking for candidate faces, the camera typically pans over a scene. Since we did not optimize our code, the process of detecting faces is slower than it could be. The result of this is that faces are generally detected after the camera has panned past them. This necessitates panning back to the position from which the image with the faces in it was taken. This produces a "double-take" sort of motion, which is surprising engaging. Again, this is the sort of reaction one expects from a human, and seeing it on a robot seems to imply intelligence.

There were a few robot behaviors that, while sensible in the context of robot navigation, communicated a distinct lack of intelligence. Due to the random walk used for navigation, for example, the robot would often spend time pointed at a wall, or moving along the wall, not "seeing" a group of people behind it. When the robot was localizing itself it appeared to be staring at the ceiling and unresponsive. Subjects often tried to help the robot by calling out to it, or by waving at it.

It seems that, once a subject believes that the robot is intelligent, they are willing to treat as if it was a (slow) human, or a pet. By assigning a known behavioral model to the robot, subjects seem to be more comfortable, perhaps because they believe that they can predict what the robot will do. The assignment of these models and the difference in interaction quality is very interesting, and is something we propose to study in a more rigorous way in the future.

Expectations

People have expectations about the behavior of the system, even if they have never seem a real robot before. Are these based on their previous experiences, or are they formed similarly, regardless of technical sophistication? A number of interactions happened in all deployments, and seemed independent of the venue or the technical experience of the subject.

It Should Respond Like A Human

Most of our interactions are with other humans, so it seems natural that subjects should use the same cues when trying to communicate with the robot. Waving to attract attention, standing still in front of the robot and smiling, and speaking are all things that most people expect to attract the attention of the robot. The robot is not currently programmed to respond to any of these cues, and this has led to some frustration on the part of the subjects.

We noticed that the robot does not need to comply with the subject's requests, just acknowledge them. Sometimes, the robot seems to notice the subject, due to a coincidental reading from one of its sensors, causing the robot to pause or turn. This "recognition" seemed to make most subjects happier than being ignored, and they seemed to interpret it as the robot saying "I see you, but I'm busy." Even if we do not have mechanisms to respond to human gestures and speech, the quality of the interaction might be improved by simply acknowledging all such attempts at communication in some way. Looking at the source of movement or sound, or pausing for a second might even be enough to convey the correct message.

Eye Contact is a Strong Cue

One of the strongest indicators of intelligence seems to be the ability to make eye contact. The movements of the camera and the pan/tilt head made the behavior of the robot seem much more intelligent than it actually was. The main examples were the "double-take" motion, described above, and visually tracking a subject as the robot is moving.

Simply keeping the camera trained on the subject while the robot body is moving increases the apparent awareness of the system. This seems particularly true when the body and the camera are moving independently. Again, this perception seemed to be shared by all subjects.

Given that camera movement seems to correlate, at least somewhat, to gaze and head direction, we might make the robot behavior more "life-like" by adding camera movements that are not strictly needed. For example, we might add small pauses before, during, and after actions. We believe that this is a general feature of robot-human interaction. It is often necessary to perform actions that are not strictly useful in accomplishing the task to make the human feel more comfortable, even if such actions hurt the overall performance of the system, as judged by strict computational metrics.

Some Robot Actions are Dumb

If we establish the robot as being intelligent, using eye-camera contact and appropriate pauses, we do not want it doing anything that could be considered stupid. Some robot actions, such as localization using landmarks, take a significant amount of time, and often have no human corollary. For example, finding landmarks might look like the robot is just staring off into the middle distance. If the robot is going to do these things, then it needs to maintain its air of intelligence. We can do this by making the actions subtle, or by explaining what it is doing. However, this explanation might be complex, and not totally compelling to some subjects. We might, alternatively, give the impression that the robot is doing something entirely different, which would explain its behavior, even if it is not the "correct" explanation.

Conclusions, Open Questions and Future Work

This project is still a work-in-progress. At the time of writing, we have the basic system in place, and have deployed it in a number of real-world environments. We are currently evaluating the results of these deployments, and deciding on the research directions to follow in the future. The robot photographer provides a good framework for research into long-term autonomy, human-robot interaction, navigation, and sensor fusion.

At this point, based on our observations, we have some tentative conclusions, and some questions that we think are important. We plan to perform experiments to verify or refute these conclusions, and investigate the questions in the near future.

The robot needs speech output. Some of the things that the robot does are complex, and have no real human analog. In order to externalize these behaviors easily, we believe that some form of speech output, even canned phrases, is essential.

We need to support bimodal interactions. For this application, at least, the robot needs to have a bimodal interaction mode. At some times it should be inconspicuous, and blend into the background. At other times it should actively engage subjects with camera-eye contact and movement. Changing between these two modes should be triggered by human interaction.

We need to support attention-getting actions. Even if we cannot completely understand or use such gestures, it seems important to respond to actions that that humans naturally use, such as waving, and speaking.

What effect does user sophistication have? Some of the people that the robot has interacted with so far are technically sophisticated, and familiar with the limitations of computer interaction, while some are not. Does the quality of the interaction change with naïve users? Do we have to change the nature of the interaction with the sophistication of the subjects?

Can we use cues to shape the interaction? Can we leverage off of cues that humans use, such as gaze-direction, speed and smoothness of movement, the notion of personal space, and (simulated) emotional state in order to direct the interaction?

Acknowledgements

This work was supported in part by NSF REU award #0139576, and NSF award #0196213. The help of Michal Bryc, Jacob Cynamon, Kevin Goodier, and Patrick Vaillancourt was invaluable in the implementation, testing, and tweaking of the photographer system.

References

Dixon, M.; Grimm, C. M.; and Smart, W. D. 2003. Picture composition for a robot photographer. Under review. Available from the authors on request.