Challenges and Opportunities in Building Socially Intelligent Machines

Laurel D. Riek and Peter Robinson

Introduction

At the recent ACM Multimedia conference in Florence, Ramesh Jain boldly and succinctly stated, "Content without context is meaningless." He presented a paper on the topic at the conference's *Brave New Ideas* session, and challenged multimedia researchers to "stop ignoring the elephant in the room" (context) and to recognize that by reducing problems to being entirely content-focused, they are doing both their research and their community a disservice [1]. Further, he writes that many intractable problems in multimedia analysis have been substantially aided by taking context into account, such as object recognition, image search, and photo management.

Multimedia is not the only field that can benefit from this mindset. Researchers working in social computing, social signal processing, human-machine interaction, robotics, computer vision, computer security, or any other field concerned with the automatic analysis of (and response to) human behavior may be greatly aided by understanding the role context plays. Indeed, some might say that understanding social context is one of the grand challenges of these fields.

But how does context affect social behavior? And, further, as researchers how can we build autonomous systems that take advantage this contextual information? This paper will broadly introduce social context, and discuss some of the challenges involved in building real-time systems that can process and respond to this contextual information. By clearly articulating these challenges, our hope is that researchers will be better equipped to confront them in their work, and can transform them into opportunities.

Defining Social Context

For the purposes of this article, we define social context to be the environment, E, where a person, P, is situated, with four factors that may influence P's behaviors. These factors include: the situational context, P's current social role in E, the cultural conventions of both E and P, and the social norms of E. (See Figure 1). P also employs social learning by observing how others in E behave, and adapting accordingly.

Situational Context

The situational context is the particular situation that P is in. This can be comprised of the physical place (e.g., a cafe), a set of actions (e.g., eating a sandwich, drinking coffee, reading the newspaper), a particular social occasion (e.g., a lunch meeting), time of day, and so on. In the fields of linguistics and pragmatics, situational context is generally defined as anything non-linguistic in the environment that can impact communication.

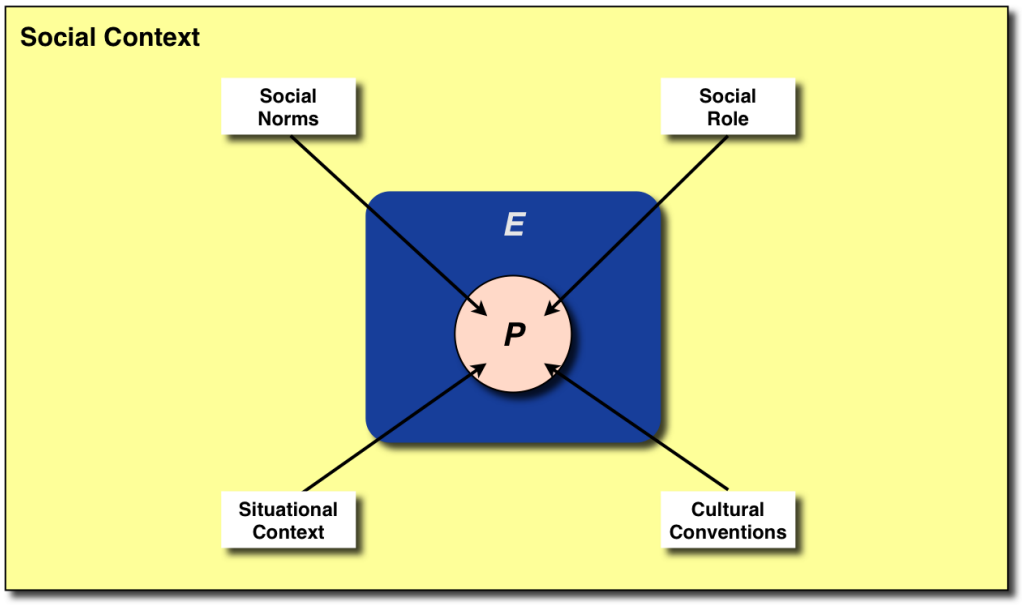

Fig. 1. **Social Context**. This figure depicts some of the ways in which social context might influence a person *P*'s behavior in environment *E*, as described by Philippot et al. [2] and Burke and Young [3].

Situational context can be particularly fluid, which can be problematic for researchers interested in building systems that automatically identify situations. For example, the situation where *P* is sitting at the company cafeteria alone eating lunch and reading the paper suddenly becomes extremely different when the company president decides to sit at the same table. *P*'s behaviors will likely become more formal, they will perhaps worry more about dining etiquette, sitting up straight, and so on.

Social psychology has been aware of the influence of situational context since the 1930s, with studies showing its effect on responsibility diffusion, conformity, and obedience to authority [4]. In recent years, social cognition and social neuroscience have shown very interesting results in this realm. Both Lieberman [4] and Philippot et al. [2] describe a number of studies in this area that show how emotional attributions, dispositional judgements, emotion recognition, and even deception can be manipulated by altering the situational context. And many of these same effects of situational context can be seen across both the behavioral literature as well as in the social neuroscience literature, which strongly suggests situational context plays a powerful role in our social behavior as humans.

Social Roles

In addition to situational context affecting *P*'s social behaviors, the particular role *P* is expected to play in environment *E* can also affect things. People generally adopt different social roles depending on the situational and cultural context they find themselves in, and the existing relationship they have with the people whom they are interacting with. Roles can have an authoritative basis (e.g., parent-child, student-teacher) or be peer-based, can vary depending on the closeness of relationship between parties (e.g., family members, strangers, work colleagues), and can vary greatly by the situation. These roles can greatly affect our behavior. For example, La France and Hecht ([2], p. 45) describe job situations where workers are required to smile for their job, such as flight attendants. These behaviors are closely linked to the influence of situational and cultural context.

Fig. 2. **Situational Context and Social Roles**. In the exact same physical environment, such as a classroom, people adopt completely different social roles depending on the situation. In the left image, a classroom is used for a lecture, where the listeners adopt very specific, student-like social behaviors (staying quiet, taking notes on what the speaker says, not gesturing very much). On the right, a classroom is used for a party, and people are able to behave more freely. *Credits: Sarvodaya Shramadana (left) and Judy Baxter (right), Creative Commons License.*

Recently, Vinciarelli [5] published an article describing several advances in building systems capable of automatically detecting social roles, such as who is the most dominant in a conversation, group formation, and conflict. These results are encouraging, and perhaps it may be possible to take what we can assess about roles and apply them to understanding other factors that affect social context.

Just as situational context presents automatic recognition challenges due to its frequently changing nature, social roles too can be very fluid and may change rapidly, even among the same group of people. Furthermore, some of these changes can be very subtle and difficult even for humans to detect, let alone machines.

Cultural Context

Kupperbusch et al. ([2], p. 38) provide a thorough review of the cross-cultural psychology literature on the role culture plays on how people express and perceive the emotions of others. Similar results have recently been shown in the nascent field of cultural neuroscience [6]; people from different cultures may process and react to images, language, and social behavior completely differently. Many of these studies compare collectivist (e.g., East Asian) to individualistic (e.g., Western) culture, and often show stark differences in studies involving perception, emotion, causal attribution, and motivation [6].

This is, however, only part of the problem of cultural context. People change their behaviors to match the cultural context they happen to be situated in. In his book *The Geography of Thought*, Nisbett [7] describes several anecdotes and studies of how people dramatically alter their behaviors after living in a different culture. For example, a Canadian psychologist who had lived in Japan for awhile went on the academic job market in North America and began his cover letter apologizing for his unworthiness for the job. Similarly, Japanese who spend more time in Western countries show a boost in self-esteem and subsequent behavior.

Fig. 3. **Cultural Context**. Culture can also play a substantial role in how we interact with one another and with technology. In the left image, a man from Oman shakes hands with an android robot in a Western fashion. On the right, he greets the robot with a traditional Bedouin greeting, by rubbing noses three times. Outside of the Gulf, this same gesture could have a completely different meaning.

For technology developers, cultural context raises all sorts of interesting "social localization" issues. A machine learning system trained on British urban cultural behaviors will likely have problems in rural Japan. Furthermore, assumptions technologists make about how their systems will be used by different cultures can also be very difficult to predict in advance. See Figure 3 for an example of an android robot one of the authors worked with in the United Arab Emirates, and how people interacted with it in unexpected ways.

Social Norms

Burke and Young [3] define a social norm as "a standard, customary, or ideal form of behavior to which individuals in a social group try to conform." The authors suggest that the key feature of social norms is that they can help promote a "positive feedback loop" in behaviors, in that the more widely people in a social group follow a norm, the more people deign to adhere to it. These norms are enforced by several internal and external factors, including coordination with others (e.g., everyone in society agrees to pay for things with cash instead of bottle caps), the threat of social disapproval should a norm be violated, and an internalization of the current accepted norms of what sorts of behavior are acceptable in a given situation (e.g., it is not acceptable to litter, even if no one will notice).

Social norms can vary widely by location, and thus people are generally very good at figuring out how to behave by observing what everyone else does. This adaptation is particularly apparent when traveling; a social norm entirely acceptable in one's hometown may be completely unacceptable (or even illegal) in a different country.

Thus, social norms are intricately tied to cultural context, and indeed may even be considered a subset.

Fig. 4. **Situational Context**. These two scenes share similar compositions but depict extremely different social meanings. On the left, the "sad" face on the man can only be understood by the use of contextual objects - the suitcase, bus stop, and bicycle taken together indicate that the couple are about to go their separate ways. On the right image, by taking into account the man and woman's proxemics (distance), noticing their opposite head and gaze orientations, one would not necessarily assume any prior relationship, just two people waiting for a bus.

Practical Challenges in Exploiting Context

Building machines that can make sense of social scenes is a very difficult problem; however, it may be possible to use contextual cues to help with clarification. For example, see Fig. 4. In the left image, face recognition alone will not explain why the man is sad - it is only by taking into account the bicycle, suitcase, and busstop that social meaning can be inferred. The couple are about to go separate ways - one is taking the bus somewhere, possibly far away given the suitcase, and the other is going somewhere locally on the bicycle. In the right image, by detecting the distance between the subjects and by noting they do not look at one another, one can probably assume that they are strangers.

This sort of reasoning about situational context likely requires *a priori* common-sense knowledge about the world. Systems like Cyc [8] and others are developing large ontologies that capture and can reason about this knowledge. Cyc is probably well-equipped to handle the first busstop problem, as it surely knows bicycles, busses, and suitcases are used for travel, people who look sad when suitcases are present are likely saying goodbye to one another, etc.

However, Cyc may not do as well with problems involving cultural context. For example, knowing the fact that some people in Gulf countries greet each other with a nose rub does not mean that a Cyc-like system could predict in advance how someone from Oman might greet someone in New York. Or, indeed, a nose rub a New Yorker gives to another New Yorker probably does not mean the same thing as the Bedouin greeting. Most automatic gestural understanding systems will probably encounter difficulty if used on people from different cultures in different countries, as gestures (or even expressivity) is not universal.

Solving this class of problem requires the ability for machines to learn social and cultural norms on the fly. The new field of Socially-Guided Machine Learning [9] is well-poised to address these problems. The idea behind this approach is to have human teachers help embodied (or virtual) agents learn new tasks and beliefs about the world through the use of situated learning.

Even by combining these approaches, a large practical problem remains - human unpredictability. In

previous experiments we have conducted on real-time social interaction between people and physical robots, we found that rarely do people move in ways we can predict in advance [10]. A lot of our initial assumptions during human-machine interaction are often unwittingly thwarted by people who stand out of view of the camera, talk too slowly or quietly for our microphones and speech recognition software to process their voices, or adopt other behaviors that make it extremely difficult to support real-time, autonomous interaction.

But just as humans are able to "fill in the gaps" and reason about social context under less-than-perfect conditions, there is no reason why autonomous systems, too, cannot ultimately be programmed to do so. This is, perhaps, the grand challenge that faces those of us who work in fields that require the ability to automatically analyze and respond to human behavior. Real-world data and real-time data is noisy, chaotic, and unpredictable. However, perhaps by using context in clever ways, we can face the elephant in the room and make some substantial progress on tackling the challenges facing our field.

ACKNOWLEDGMENTS

The authors would like to thank Andra Adams, Shazia Afzal, and Tadas Baltrustritus.

REFERENCES

[1] R. Jain and P. Sinha, "Content Without Context is Meaningless," in *In Proceedings of ACM Multimedia (ACM-MM '10)*, 2010.

[2] P. Philippot, R. Feldman, and E. Coats, *The social context of nonverbal behavior*. Cambridge Univ Pr, 1999.

[3] M. A. Burke and P. Young, "Norms, Customs, and Conventions," in *Handbook for Social Economics*, J. Benhabib, A. Bisin, and M. Jackson, Eds. Elsevier, 2010.

[4] M. D. Lieberman, "Neural bases of situational context effects on social perception," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 2, pp. 73–74, 2006.

[5] A. Vinciarelli, "Capturing order in social interactions [Social Sciences]," *Signal Processing Magazine, IEEE*, vol. 26, no. 5, pp. 133–152, 2009.

[6] E. A. R. Losin, M. Dapretto, and M. Iacoboni, "Culture and neuroscience: additive or synergistic?" *Social Cognitive and Affective Neuroscience*, 2010.

[7] R. Nisbett, *The geography of thought: How Asians and Westerners think differently... and why*. Free Pr, 2004.

[8] K. Panton, C. Matuszek, D. Lenat, D. Schneider, M. Witbrock, N. Siegel, and B. Shepard, "Common sense reasoning–from Cyc to intelligent assistant," *Ambient Intelligence in Everyday Life*, pp. 1–31, 2006.

[9] A. Thomaz, "Socially guided machine learning," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[10] L. Riek, P. Paul, and P. Robinson, "When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 99–108, 2010.

Laurel D. Riek is a PhD candidate at the University of Cambridge Computer Laboratory. She researches natural human-robot interaction, in particular, facilitating non-verbal communication with robots. Her research explores expression synthesis on android and humanoid robots using naturally evoked human data. She also explores sustainable interaction with robots by applying social signal processing techniques to the analysis of dyadic human conversations. Prior to starting her PhD, she worked for eight years as a Senior Artificial Intelligence Engineer and Roboticist at MITRE, on projects involving search and rescue robots, unmanned vehicles, and human language technology. She received her BSc in Logic and Computation from Carnegie Mellon University in 2000.

Peter Robinson is Professor of Computer Technology at the University of Cambridge Computer Laboratory, where he leads work on computer graphics and interaction. His research concerns new technologies to enhance communication between computers and their users, and new applications to exploit these technologies. Recent work has included desk-sized projected displays and inference of users' mental states from facial expressions, speech, posture and gestures. He is a Chartered Engineer and a Fellow of the British Computer Society.