Long-term socially perceptive and interactive robot companions: challenges and future perspectives

Ruth Aylett
MACS, Heriot-Watt University
Riccarton
Edinburgh EH14 4AS, UK
r.s.aylett@hw.ac.uk

Ginevra Castellano
Schl of EECE
University of Birmingham
Birmingham, B15 2TT, UK
g.castellano@bham.ac.uk

Bogdan Raducanu
Computer Vision Centre
Edifici "O" – Campus UAB
08193 Bellaterra (Barcelona), Spain
bogdan@cvc.uab.es

Ana Paiva
INESC-ID and Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal
ana.paiva@inesc-id.pt

Marc Hanheide
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
m.hanheide@cs.bham.ac.uk

ABSTRACT
This paper gives a brief overview of the challenges for multi-model perception and generation applied to robot companions located in human social environments. It reviews the current position in both perception and generation and the immediate technical challenges and goes on to consider the extra issues raised by embodiment and social context. Finally, it briefly discusses the impact of systems that must function continually over months rather than just for a few hours.

Keywords
Social robotics, human-robot interaction, multi-modal interaction

1. SOCIAL ROBOTICS
In the last fifteen years or so, there has been a perceptible shift away from a narrow task-oriented view of the relationship between human users and computer-based technologies. Thus in HCI, the focus has moved from usability to the user experience [Law]; graphical Embodied Conversational Agents (ECAs) have been developed in which multi-modal interaction to support expressive behaviour and affective engagement has become a research area [7], and a whole new field known as Human-Robot Interaction, or HRI, has developed [12]. Here robots are no longer merely machines for achieving tasks but become social actors in real-world human environments. The special session "Long-term socially perceptive and interactive robot companions: Challenges and future perspectives" at the International Conference on Multimodal Interaction 2011 was a further illustration of this trend.

2. MULTI-MODAL INTERACTION
An early result with respect to graphical characters was that users would often treat them as if they were human interaction partners, even though they knew that they were not [29]. Work with early social robots such as Kismet [5] substantiated this for robots too. Users act as if these technological artefacts had their own inner life: motives, goals, beliefs, and feelings; thus applying the Intentional Stance [13] to them. We argue that the term *believability* [2], often used as an evaluation metric for such interactions, is closely related to how far a user feels able to take this intentional stance and suspend their disbelieve in the actual status of the interaction partner as machine or collection of graphical pixels.

A significant factor in believability, and thus user engagement, is the extent to which the interaction partner displays appropriate affective behaviour, thus allowing the user to track their putative inner state, and the extent to which it responds appropriately to the user's affective behaviour. This is sometimes called the *affective loop* [31] and has implications for multimodal interaction both on the side of perception and the side of generation.

2.1 Multimodal perception
The socially perceptive abilities that are a key requirement for a robot to be able to interact socially with human users [9] include: recognising people's social affective expressions and states, understanding their intentions, and accounting for the context of the situation.

These cannot be reduced to issues of speech recognition or natural language understanding, important though these modalities are. It is known that in the human-human case, a substantial proportion of the overall interaction may be carried by body language [26] and that this is at its greatest for affective aspects of communication. Overall, face and body appear to be the most important channels for communicating behavioural information, including affective states [1].

Recent work in affective computing is making progress towards the design of systems capable of perceiving social affective cues and using them to infer a person's affective state or intention [34, 37]. In addition to these natural modalities, work is also being carried out into the use of biological and physiological signal modalities. Thus the advent of sports monitoring equipment for heart-rate and pulse also facilitates the detection of states of raised arousal known to be associated with some affective states. Brain-computer interfaces make use of external sensors registering the activation of particular areas of the brain and particular patterns of brain electrical activity [19].

International Conference on Multimodal Interaction 2011 Nov 14-18, 2011, Barcelona, Spain.

However a limitation of much work so far on the automatic detection of user behaviour and states is that it is usually specialised for handling specific situations and is often much less successful under the natural and unanticipated conditions that occur in a real environment. This is why the design of socially perceptive robots requires the research community to address more comprehensively a specific set of key technical challenges [8]

1) **Spontaneous and non-prototypical affective expressions and states.** Research on automatic affect recognition needs to move towards systems sensitive to spontaneous, non-prototypical affective expressions and states. A typical issue in real environments is that the affect being expressed is likely to be fleeting and of low magnitude. Furious rage or ecstatic happiness are rare in HRI, and if they do occur – usually as the result of a deliberate experimental design – they are unlikely to last for very long. An alternative is to give a user explicit modalities with which to express their state, whether using devices such as the WiiMote to carry out specific gestures [25], or specifically designed interfaces [30].

2) **Multimodal nature of affective expressions**. Humans usually rely on several different channels of information to understand the affective messages communicated by others. Similarly, an automatic affect recognition system should also be able to analyse different types of affective expressions. Moreover, all of the modalities involved in natural human-robot interaction are difficult to process, especially outside of laboratory-based protocols or the use of actors to exaggerate the affective body language. Thus a multimodal approach, often one tailored to the specific interaction domain, is currently required to achieve acceptable results.

3) **Robustness to affect dynamics and real-world environments.** Social robots employed in real world applications require systems for the detection and interpretation of social affective cues that are robust to dynamic and noisy real world environments. This implies the ability to, in real-time, account for affective states and expressions with variable initialisation times, as well as for their temporal dynamics.

4) **Sensitivity to context**. The detection of the most subtle and complex states and intentions can only be achieved through a comprehensive analysis, not only of their effects, but also of their causes and context. An affect recognition system for a social robot should take into account the events that triggered the generation of an affective state and be context sensitive. Thus it should be able to capture, elaborate and integrate contextual information such as individual differences in affect expression, personality, preferences, goals, task, environment, etc..

2.2 Multimodal generation

Multimodal generation of social signals has been extensively investigated in the area of intelligent graphical characters, notably in those known as Embodied Conversational Agents, or ECAs [7]. Here, back-channel behaviour during conversations focused researchers on the generation of appropriate contextual use of facial expression, glance, and gesture [7], in systems that often had a pseudo-human appearance. Facial animation typically used the Facial Action Coding System (FACS) [16] which was derived from the analysis of human faces.

However, what counts as an appropriate social signal cannot be divorced from the way an interaction partner is embodied. This issue becomes far more significant in the case of robots, where embodiment must to a greater or lesser extent be physically implemented and thus is not easy to vary. While some work, typically in Japan, is based on robots intended to look as human as possible – robots sometimes referred to as *androids* [27] - this is only a small segment of the robot field. A 2003 survey [18] identifies *anthropomorphic* as only one approach to embodiment, with *zoomorphic*, *caricatured* and *functional* as other possibilities.

Human-like embodiments are seen as having advantages, for example in supporting imitation-based learning. On the other hand, the by-now well-known concept of the Uncanny Valley [28], in which close-to-human robots may provoke very negative feelings in interacting humans, has led to work studying social signals in other types of robots. There is evidence that a perceived consistency between the social role suggested by the embodiment and the range and type of social signals is a desirable aim [36]. It is clear that some researchers seek naturalism in expressive behaviour whether inspired by human or by animal behaviour. Others are more influenced by dramatic inventions, or by a more functional view of robot expressive behaviour.

Certainly there are feasibility as well as consistency issues. A functional-type robot without a face cannot generate facial expression. Here, modalities such as flashing lights may be required, or the use of movement dynamics (for example surprise could be conveyed by a short rapid movement backwards by a mobile robot). Movement dynamics are an area in which animated film offers design inspiration since these are techniques long-used there.

A robot that looks like a dog (Aibo) or a seal (Paro) cannot generate standard FACS-based facial expressions. Even in the case of caricatured robots with elements of humanoid embodiment, it is not always clear that naturalistic FACS-based facial expressions should be the aim as distinct from the exaggerated and simplified expressions of animated cartoon characters. Kismet [6], which had controllable eyebrows, ears, eyeballs, eyelids lips, and neck, had this type of expressive behaviour.

Apart from embodiment and consistency issues, research in multi-modal generation also covers the models used to determine in context what is an appropriate expressive behaviour. In intelligent graphical characters, the use of cognitive appraisal-based architectures [14] is one approach to linking perceived events to an appropriate affective state to be expressed. In robots, such symbolic-level architectures are much more complex to implement since they must depend on deriving symbolic information from sensor data that may be noisy and ambiguous. Lower level architectures have tried to produce more direct mappings [33] between sensor data and expressive behaviour and this facilitates the use of imitation learning as a way of developing such behaviours [4]. However these approaches do not integrate well with the use of natural language where expressive behaviour should relate to semantic content and are thus more appropriate in robots playing a more animal- or machine-like role.

3. FROM LAB TO REAL WORLD

The test of multi-modal interaction is the extent to which it can be embedded in real-world artifacts that can function successfully in human social environments.

3.1 Applications - Companions

Once the issue of a social environment is raised, then the social role that a robot companion is being asked to play becomes a key factor in defining its desired interactive capabilities. The very term *companion* is ambiguous in this context and may be given quite different interpretations by different groups [11]: for example a pet; a butler; a care assistant; a lab team member; a playmate; a pseudo-child. Some work suggests that more functional robots that do not attempt to masquerade as objects of affective attachment are preferred [11]; on the other hand work in graphical characters has shown that the development of a perceived personality may also improve interaction outcomes [3].

Some work focuses on developmental robotics, in which the pseudo-child role is the context for the development of learning capabilities [17] but there are also significant projects looking at specific companion applications [22, 10, 15 , 20]. As an example, the LIREC project [22] has studied the scientific issues relating to robot companions in the home, as assistants in maintaining independence for the elderly; in the work environment, as assistants to a lab-based research team or visitors to a new building; and as playmates for children in the context of a game such as chess.

The most successful companion project in terms of real-world deployment is Paro [35], a seal robot intended to provide extra interactive stimulation for elderly sufferers from dementia. In picking an animal-like embodiment , moreover an animal that few humans actually interact with, it avoids heightened expectations as well as over-demanding multi-model interaction requirements.

Where conventional robotics focused on health and safety, and often solved such issues with straightforward engineering approaches such enclosures and bumper-bars for an automatic stop, the embedding of robot companions into social environments both makes those issues much more demanding and adds concerns relating to privacy, data security, and ethical standards [Vargas]. Multi-modal interaction must then not only deal with the demands of the specific interaction, but also with these constraints on the nature and content of the interaction in the given interaction context and with the given interaction history. Much further research is needed in this area.

3.2 From days to months

The ability to perform a social role successfully in most cases also implies the ability to do so for open-ended periods of time. Yet interaction over the long-term, past the period in which a novelty effect operates, raises serious research issues that are only beginning to be explored.

Some of these issues relate to basic competence, robustness and autonomy. Thus a robot that claims human attention for recharging every couple of hours lacks a basic capability for long-term social interaction. Similarly, a robot that perpetually collides with furniture or people, or is forever asking a human interaction partner to repeat itself due to inaccurate speech recognition is not able to play any long-term social role other than that of intolerable irritant. In domestic settings, we know how to deal with a washing machine that has broken down, but not with a robot in a like case.

While these are not trivial problems to solve or to engineer around, there are also issues relating specifically to interaction. What may be engaging and pleasurable when it is a novelty, may be boring or irritating as a long-term behavioural pattern. Over-emphatic multi-modal output may perform badly over the long-term just as failures in multi-modal perception may become more obtrusive. The mutual adaptation of long-term human-human interaction remains largely to be studied in human-robot interaction, though existing studies do already demonstrate its necessity if human engagement in the interaction is to be maintained [23].

In long-term interaction, the detection of subtle differences between users, so as to adapt in a personalized way, becomes essential. This requires not only the ability to change the interaction repertoire, but also to detect and act upon the preferences of the user as well as the specific history of interaction with them. Longer periods of interaction do help with the detection problem, but work in adaptive interfaces suggests that a disconcerting inconsistency may be perceived by human interaction partners if adaptation is carried out automatically without explicit communication. This has implications for the mix of declarative and procedural components of a long and short-term companion memory: what cannot be explicitly recalled cannot be explicitly discussed either. Thus the development of acceptable human-like memory capabilities may be a central research issue for successful long-term interaction [24].

4. CONCLUSIONS

In a new field there are always many challenges. All the more so when long-term integration into human social environments is the overall aim. We have argued that both multi-model perception and generation face many as yet unsolved research issues. However, perhaps more important is the impossibility of separating the multi-modal interaction challenges from issues of embodiment on the one hand, and social context on the other. As an example, consider a simple affective loop in which a robot companion recognizes a smile and smiles back.

Certainly there are technical challenges in recognizing a smile in the first place, especially if the recognizing system is mounted on a mobile robot in a variably-lit human environment, and the user can be standing at variable distances and angles, with skin, facial and head hair of different colours, and variable height and clothing.

However, even when these are overcome, a companion still needs to know what the significance of the smile is before automatically assuming it reflects a happy interaction partner. Here contextual information both from other modalities, from a memory of previous interactions with this user, as well as the generic information about the relationship between events and emotions supplied by a cognitive appraisal system can all help. Knowing whether the user has just entered the lab and is greeting the robot or whether they have just dropped a whole pile of paper and might be embarrassed would also help.

Once the relevant expressive behaviour is determined, the robot companion still has to respond. And with: a mouth? Some lights? Happy movement? A purring noise? Raising ears? Something else? Some combination? Here in minature are some of the challenges of a fascinating research field.

5. REFERENCES

[1] Ambady, N. and Rosenthal, R. (1992). Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. Psychological Bulletin, 111(2), 256-274.

[2] Bates, J. (1994) The role of emotion in believable agents. Communications of the ACM: Special Issue on Agents 37(7) (Jul 1994) 122–125

[3] Bickmore,T.W; Shulman, D. & Yin, Langxuan (2009) Engagement vs. Deceit: Virtual Humans with Human Autobiographies. IVA 2009: 6-19

[4] Boucenna, S.; Gaussier, P.; Andry, P.; Hafemeister, L.; (2010) Imitation as a communication tool for online facial expression learning and recognition. IROS 2010 pp 5323-28

[5] Brezeal, C. (2002) Designing Sociable Robots. MIT Press.

[6] Breazeal, C. (2009) Role of expressive behaviour for robots that learn from people. Philosophical Transactions of the Royal Society B, 364:3527–3538, 2009.

[7] Cassell J. 2000. Embodied conversational interface agents. *Communications of the ACM*, 4 (April), 70–78.

[8] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A, & McOwan, P. W. (2010). Affect Recognition for Interactive Companions: Challenges and Design in Real World Scenarios. *Journal on Multimodal User Interfaces*, 3(1), 89-98, Springer.

[9] Castellano, G., & Peters, C. (2010). Socially Perceptive Robots: Challenges and Concerns. *Interaction Studies*, 11(2), John Benjamins Publishing Company.

[10] CompanionAble: http://www.companionable.net/ Accessed 5/9/11

[11] Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., Koay, K. L. & Werry, I. (2005) What is a robot companion—friend, assistant or butler? In Proc. IEEE IRS/RSJ IROS 2005

[12] Dautenhahn, K. (2007) Socially intelligent robots: Dimensions of human-robot interaction. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480):679–704, 2007.

[13] Dennett, D.C. (1989) The Intentional Stance. MIT Press

[14] Dias, J., Paiva, A.: Feeling and reasoning: A computational model for emotional agents. In: 12th Portuguese Conference on Artificial Intelligence (EPIA 2005), Portugal, Springer (2005) 127–140

[15] DOMEO: http://www.aal-domeo.org/ Accessed 5/9/11

[16] Ekman, P. & Friesen, W. (1982) Measuring facial movement with the facial action coding system. In: Emotion in the Human Face, Cambridge University Press 1982

[17] FEELIX Growing: http://www.feelix-growing.org/ Accessed 5/9/11

[18] Fong, T; Nourbaakhhsh, I. & Dautenhahn, K. (2003) A survey of socially-interactive robots. Robotics and Autonomous Systems, v42, pp143-166

[19] Gunes, H., and Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition, International Journal of Synthetic Emotions, Vol. 1, No. 1, pp. 68-99, 2010.

[20] KSERA: http://ksera.ieis.tue.nl/ Accessed 5/9/11

[21] Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., Kort, J.(2009) Understanding, Scoping and Defining User Experience: A Survey Approach. CHI'09. April 4-9, 2009

[22] LIREC: http://lirec.eu Accessed 5/9/11

[23] Leite, Y; Pereira.A; Castellano,G; Mascarenhas,S; Martinho.C & Paiva,A. (2010) Social Robots in Learning Environments: a Case study of an Empathic Chess Companion. Personalization Approaches in Learning Environments (PALE), CEUR Workshop Proceedings

[24] Mei Yii Lim, Ruth Aylett, Wan Ching Ho, Patricia Vargas and Sibylle Enz, (2009) *A Socially-Aware Memory for Companion Agents,* IVA2009 pp20-26, Springer

[25] Lim, M.Y; Aylett, R.S; Enz,S; Kriegel,M; Vannini,N; Hall, L. and Jones, S. (2009), Towards Intelligent Computer Assisted Educational Role-Play, 4th Int. Conf on E-Learning and Games, Banff, Canada, August 9-11, 2009, Springer

[26] Mehrabian, Albert (1972). *Nonverbal Communication*. Chicago, IL: Aldine-Atherton

[27] Minato, T., Shimada, M., Ishiguro, H., & Itakura, S. (2004). Development of an Android Robot for Studying Human-Robot Interaction. Innovations in Applied Artificial Intelligence, (pp. 424-434).

[28] Mori, M. (1970) Bukimi no tani [the uncanny valley]. Energy, 7:33–35

[29] Reeves,B. & Nass,C.I. (1996) The media equation: How people treat computers, television, and new media like real people and places. Univ Chicago Press

[30] Ståhl A., Höök K., Svensson M., Taylor A., Combetto M. (2008). Experiencing the affective diary. J. Pers. Ubiquit. Computing

[31] Sundström, P. (2005). *Exploring the affective loop,* Stockholm: Stockholm University.

[32] Vargas, P. A., Fernaeus, Y., Lim, M. Y., Enz, S., Ho, W. C., Jacobsson, M. and Aylett, R. (2011): Advocating an ethical memory model for artificial companions from a human-centred perspective , *Artificial Intelligence & Society*

[33] Velasquez, J.D. (1997) Modeling Emotions and Other Motivations in Synthetic Agents. Proceedings, 14th National Conference on AI, AAAI Press, pp10-16

[34] Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. Image and Vision Computing Journal, 27(12), 1743-1759.

[35] Wada,K. and Shibata,T. (2007) 'Living with seal robotsits sociopsychological and physiological influences on the elderly at a care house', IEEE Transactions on Robotics, 23(5), 972–980

[36] Walters ML;,Syrdal DS; Dautenhaun K; te Boekhorst R. & Koay KL (2008) Avoiding the uncanny valley: Robot appearance, personality and consistency of behavior in a an attention-seeking home scenario for a robot companion. Auton Robot 24:159–178

[37] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39-58.