ANALYSIS OF AFFECTIVE CUES IN HUMAN-ROBOT INTERACTION: A MULTI-LEVEL APPROACH

Ginevra Castellano and Peter W. McOwan

Department of Computer Science,
School of Electronic Engineering and Computer Science,
Queen Mary University of London, UK

ABSTRACT

This paper reviews some of the key challenges in affect recognition research for the purpose of designing affect sensitive social robots. An important requirement for a social robot is to be endowed with recognition abilities that vary according to the context of interaction. This paper presents an approach for the analysis of different affective cues depending on the distance at which user and robot interact.

1. INTRODUCTION

An important aspect of the design of social robots [1] is the ability to infer the user's affective and mental states, so as to be able to engage in and act in an appropriate way during social interactions, for example, to ensure the user is interested in maintaining the interaction or to behave empatically.

The design of an affect recognition module based on the interpretation of the user's behaviour is the first step towards the generation of a proper social behaviour. While researchers have been increasingly investigating affect recognition [2], the design of such a module to be integrated in a human-robot interaction framework has not been extensively addressed yet.

Robot companions are an example of robots which may benefit from the integration of such a "social perception module". Robot companions can be useful in many applications: they can be employed as personal assistants in smart environments, as interactive toys for therapy and rehabilitation purposes, they can provide additional functionalities to assist carers, healthcare workers, etc. For robot companions to be able to cover these roles it is necessary that they are endowed with social capabilities [3] and are sensitive to what happens in the external world, with a special attention to what the user feels or communicates.

In order to establish a truly natural interaction with the user, recognition abilities must be designed according to the context of interaction. This paper presents an approach to affect recognition in which different affective cues are analysed depending on the distance at which user and robot interact, with a focus on non-verbal cues. In the following, we review some of the issues arising in the process of endowing robots with affect sensitivity. Challenges in the design of an affective perceptual framework sensitive to different types of cues are also discussed.

2. CHALLENGES IN AFFECT RECOGNITION RESEARCH

The design of robots that behave socially requires research on affect recognition to be taken beyond the state of the art. First of all, it is necessary that robots are provided with the ability to detect more subtle states than just prototypical emotions and their affect recognition system is trained with spontaneous, real-life expressions. They need to be sensitive to multiple modalities of expression, as affective messages are multimodal by their nature. Finally, affect recognition systems need to work in unconstrained environments. In the following we review some of the challenges and relevant issues for the design of social robots from the perspective of affect recognition.

2.1. Beyond prototypical affective states

Most of the research on affect recognition mainly addressed the issue of recognition of prototypical affective states such as basic emotions (e.g., joy, sadness, disgust, surprise, fear, anger, etc.). Social robots, instead, should be endowed with affect recognition abilities which go beyond the detection of prototypical emotions and are also sensitive to application-specific affective states, such as interest, boredom, frustration, willingness to interact, etc.

Some efforts reported in the literature include the work by Kapoor et al. [4], who proposed an approach to predict frustration in users interacting with a learning companion using multimodal non-verbal cues and the system developed by el Kaliouby and Robinson [5], that allows for the detection in real-time of complex mental states such as *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure* from head movement and facial expressions.

This work was supported by the EU FP7 ICT-215554 project LIREC (LIving with Robots and intEractive Companions).

The inclusion of affect representation into a framework for affect recognition represents a pressing need, while at the same time it is still challenging. In this respect, incorporating theories from psychology [6] would allow for models for automatic affect recognition to be improved to make the detection of more complex affective and mental states feasible.

2.2. Spontaneous affective expressions

One of the primary requirements for socially intelligent robots, such as robot companions, is the sensitivity to the affective states and expressions displayed by the user in everyday's life. This requires robots to be provided with affect recognition systems that are trained with spontaneous, real-life and, possibly, application-specific expressions. One of the issues in affect recognition research is that most of the studies reported in the literature are based on acted affective expressions. These can be defined precisely and include several expressions for the same individual; they can be characterised by very high quality, but they often reflect stereotypes and exaggerated rather than spontaneous expressions.

In the affect recognition community, there have been some attempts to build systems based on spontaneous, real-life expressions. Ioannou et al. [7], for example, designed a neurofuzzy system for emotion recognition which allows for the learning and adaptation to a specific user's naturalistic facial expression. Kapoor et al. [4] proposed a system that detects and responds to a user's naturalistic non-verbal expressions that precede frustration. Devillers and Vasilescu [8] investigated the detection of affective states in a corpus of real-life dialogs collected in call centers using linguistic and paralinguistic features. Other researchers reported results on the automatic discrimination between posed and spontaneous facial expressions: see, for example, the work by Valstar et al. [9], who proposed an approach to distinguish between posed and spontaneous smiles by fusing multiple modalities of expression.

2.3. Sensitivity to multiple modalities

When considering affect recognition, it is expected that social robots are capable of integrating and interpreting multiple signals of different nature. Fusing multimodal affective cues can increase the chances that a better understanding of the affective message communicated by the user is achieved.

While multimodal affect recognition represents an important requirement of affect sensitive systems, the majority of the studies reported in the literature focus on affect recognition based on one single modality [2]. Nevertheless, studies on multimodal affect recognition are currently gaining ground. Some studies addressed bimodal affect recognition based on the fusion of facial expressions and body gesture, facial expressions and head gesture, head and body gesture, facial expressions and speech, physiological signals and speech. See [2] for an extensive overview. Some attempts of using multiple modalities have also been reported in the literature. In [4] frustration is predicted using several non-verbal affective expressions. Another study shows how facial expressions, body gesture and speech information is fused together at different levels to infer eight emotions in speech-based interaction [10]. In [9] multimodal information conveyed by facial expressions, head and shoulders movement is combined to discriminate between posed and spontaneous smiles.

A key issue in multimodal affect recognition is represented by the fusion of different modalities. Results from studies in psychology and neurology [11] report that the integration of different perceptual signals occur at an early stage of human processing of stimuli. This seems to suggest that different affective signals should be processed in a joint feature space rather than combined with a late fusion. Moreover, often features from different modalities are incompatible and their relationship is unknown. New methods for multimodal fusion should take into consideration what are the underlying relationships and correlation between different modalities [12].

The integration of multimodal information should not focus only on signals generated by the user, but also on contextual features. The appearance and the behaviour displayed by a social robot can definitely influence the perception of the user and her affective feelings towards it. Other contextual information such as the personality of the user, the type of interaction with the robot (e.g., short-term vs long-term), the environment in which the interaction takes place can also play an important role in the process of understanding how the user feels. An example reported in the literature is the work of Kapoor and Picard [13], who proposed an approach for the detection of interest in a learning environment by combining non-verbal cues and information about the learner's task.

2.4. Robustness in everyday settings

Socially intelligent robots must be able to work in the user's everyday settings. This means that affect recognition systems must be designed so as to be robust in real-world conditions: face detectors and body and facial features tracking systems which are robust to occlusions, noisy background [14], rigid head motions are examples of important requirements.

From the perspective of affect recognition, an issue that cannot be neglected is the type of platform that an affect recognition system runs on. Running complex algorithms for image processing and high-level interpretation on mobile platforms can be more problematic than for fixed robots, where distributed architectures can be designed to overcome issues related to limited processing power. To overcome this problem a possible approach can consist of having an additional laptop computer mounted on the robot. On the other hand, detection of simple low-level cues is also imporant, for example to endow a robot with an attention system [15].

Working in the user's settings, a robot must be able to infer the user's state in real-time. In this respect, the segmentation and the analysis of the temporal dynamics of affective expressions represent a key issue, since a user's affective state can start at any time [2]. The dynamics of affective expressions is a factor of primary importance in the interpretation of human behavior. Affective expressions vary over time, together with their underlying affective content. Analysis of static affect displays cannot account for temporal changes. Detection of temporal segments of affective expressions and analysis of their temporal evolution are then issues to be considered in the design of an affect recognition system for social robots. Some efforts towards a dynamic account for affective expressions that have been reported in the literature include [14], [9] and [16].

3. A MULTI-LEVEL APPROACH

As described in the previous Sections, it is expected that social robots are endowed with the ability to analyse different types of affective cues. These directly depend on the specific scenario of interaction with the user.

In a given scenario, the distance between user and robot defines the level of interaction between them. This means that the distance impacts the ability and the need for the robot to perceive and interpret different affective cues and states. In the following Sections we propose an approach in which different types of affective expressions analysis are performed depending on the distance existing between user and robot.

3.1. Short-range interaction

We refer to short-range interaction as the condition in which user and robot are face-to-face. Under these conditions, the cues that robots can base their prediction of affect on are those emerging in "face-to-face interaction". With specific reference to non-verbal cues, examples are facial expressions, eye gaze, head gestures and orientation, posture, body expressivity [17]. An examplar work is the study by Kapoor et al. [4], who proposed a method to detect frustation in users interacting with a learning companion agent based on analysis of multimodal non-verbal behaviours including facial expressions, head gestures, posture, skin conductance and mouse pressure. In the robotics domain, one of the most famous examples is the work by Breazeal and colleagues, who designed an attention system based on low-level perceptual stimuli for the Kismet robot [15]. An overview of other studies investigating analysis of affective cues in face-to-face interaction is provided in Section 2.

3.2. Medium-range interaction

In case of medium-range interaction, robot and user do not interact face-to-face, but the user is the range of the robot. When users are at this distance from the robot, not too far and not too close, two examples of affective states/events that may be important for the robot to detect, are the predisposition to interaction and the interaction initiation.

The assessment of such states/events do not necessarily require the high-level interpretation of complex cues and expressions. The main focus, at this level, is on global indicators, such as full-body movements and their qualities, such as the quantity of motion and the degree of contraction/expansion of movements and gestures, that are reported to be effective cues for affect discrimination (see, for example, [16] and [18]). Recognition of simple gestures and actions, such as waving, approach, or avoidance may also be relevant to infer some information about the user in interaction scenarios in which the user is not interacting face-to-face with the robot but at the same time is still in its proximity. A combination of low and high-level cues may be of help when the robot is required to assess an "interaction initiation" condition. As in such a condition it is expected that the user approaches the robot from a certain distance and then gets close to it, the combination of full-body movement analysis and recognition (e.g., amount of movement, motion direction, approach) and analysis of face-to-face cues, such as face direction and eye gaze (see the work by Peters [19] for an example of modelling an interaction initiation scenario in a multi-agent environment based on gaze) is envisioned to be included in a framework for affective and affect-related states recognition.

3.3. Long-range interaction

This Section refers to interaction scenarios in which the user is in the same environment, but not in the range of the robot. Presence of people in the room, predisposition to interact, group affect are some of the affect-related states and expressions that a social robot may be designed to detect. Coarse cues such as the amount of people present in the room and the frequency with which each person moves in proximity of the robot during an interaction session may be of help to the robot in determining whether it is required or not and from which user. Global indicators of movement are the main focus also for this type of scenario. Gross actions (such as walking towards the robot), expressivity of the single user [10] and group expressivity [20] are some of the indicators that a social robot can analyse in order to assess an overall propension to interaction.

4. SUMMARY AND CONCLUSION

This paper presented an overview of some of the challenges in affect recognition from the perspective of human-robot interaction. As a requirement for social robots to establish a social interaction with humans, affect sensitivity is discussed with respect to issues such as the ability to perceive spontaneous and application-dependent affective states, the ability to analyse different modalities, the robustness in everyday settings.

An approach for the analysis of different affective cues depending on the distance existing between user and robot in a given interaction scenario was proposed. We claim that in order to establish a truly natural and engaging interaction with the user, social robots should be endowed with recognition abilities that vary according to the context of interaction.

5. REFERENCES

[1] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 119–155, July 2003.

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, January 2009.

[3] K. Dautenhahn, "Socially intelligent robots: Dimensions of human-robot interaction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1480, pp. 679–704, 2007.

[4] A. Kapoor, W. Burleson, and R.W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[5] R. el Kaliouby and P. Robinson, "Generalization of a vision-based computational model of mind-reading," in *First International Conference on Affective Computing and Intelligent Interaction (ACII 2005)*, J. Tao, T. Tan, and R. W. Picard, Eds. 2005, vol. 3784 of *LNCS*, pp. 582–589, Berlin: Springer-Verlag.

[6] K. R. Scherer, "Psychological models of emotion," in *The neuropsychology of emotion*, J. Borod, Ed., pp. 137–162. Oxford/New York: Oxford University Press, 2000.

[7] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy method," *Neural Networks*, vol. 18, pp. 423–435, 2005.

[8] L. Devillers and I. Vasilescu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *International Conference on Spoken Language Processing*, 2006.

[9] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *ACM International Conference on Multimodal Interfaces (ICMI'07)*, Nagoya, Japan, 2007, pp. 38–45.

[10] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale, Eds., vol. 4868 of *LNCS*. Springer, Heidelberg, 2007.

[11] B. Stein and M. A. Meredith, *The Merging of Senses*, MIT Press, Cambridge, USA, 1993.

[12] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T.S. Huang, "Training combination strategy of multi-stream fused Hidden Markov Model for audio-visual affect recognition," in *ACM International Conference on Multimedia*, 2006, pp. 65–68.

[13] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *ACM International Conference on Multimedia*, 2005, pp. 677–682.

[14] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man and Cybernetics-Part B*, vol. 36, no. 2, pp. 433–449, 2006.

[15] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *IEEE Transactions on Systems, Man and Cybernetics-Part A*, vol. 31, no. 5, pp. 443–453, 2001.

[16] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer, "Automated Analysis of Body Movement in Emotionally Expressive Piano Performances," *Music Perception*, vol. 26, no. 2, pp. 103–119, 2008.

[17] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," in *ACM International Conference on Multimedia (MM'08)*, Vancouver, Canada, 2008, pp. 1061–1070.

[18] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 213–225, 2003.

[19] C. Peters, "A perceptually-based theory of mind model for agent interaction initiation," *International Journal of Humanoid Robotics (IJHR), Special issue on Achieving Human-Like Qualities in Interactive Virtual and Physical Humanoids*, vol. 3, no. 3, pp. 321–340, 2006.

[20] C. A. Bartel and R. Saavedra, "The collective construction of work group moods," *Administrative Science Quarterly*, vol. 45, pp. 197–231, 2000.