Exploiting Unconscious User Signals in Multimodal Human-Computer Interaction

ELISABETH ANDRE, Augsburg University

This article presents the idea of empathic stimulation that relies on the power and potential of unconsciously conveyed attentive and emotional information to facilitate human-machine interaction. Starting from a historical review of related work presented at past ACM Multimedia conferences, we discuss challenges that arise when exploiting unconscious human signals for empathic stimulation, such as the real-time analysis of psychological user states and the smooth adaptation of the human-machine interface based on this analysis. A classical application field that might benefit from the idea of unconscious human-computer interaction is the exploration of massive datasets.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Interaction styles*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms: Human Factors

Additional Key Words and Phrases: Emotion recognition, social signal processing, multimodal interfaces

ACM Reference Format:

Andre, E. 2013. Exploiting unconscious user signals in multimodal human-computer interaction. ACM Trans. Multimedia Comput. Commun. Appl. 9, 1s, Article 48 (October 2013), 5 pages.
DOI: http://dx.doi.org/10.1145/2502433

Recent years have led to a shift from pure task-based human-machine interaction to more human-like social dialogue. The driving force behind this shift is the hope that a user interface is more likely to be accepted by the user if the machine is aware of the user as a social actor. A system should not only analyze what the user said or gestured at, but also consider more subtle cues, such as head movements or body posture, to infer information on the users' emotional and attentive state. Equipping a machine with social and emotional intelligence is one of the greatest challenges in human-computer interaction and multimedia computing.

When looking back at 20 years of ACM Multimedia, we see a number of trends that reflect the move towards more social human-computer interaction (see Figure 1). The first papers that focused on the recognition of affect and attention in interactive applications appeared in 1998 within a special session on face and gesture recognition. These papers already considered a variety of modalities, such as facial expressions, body postures, and vocal emotions [Nakatsu 1998]. Also, an approach that

This work is funded by EU under research grant CEEDS (FP7-ICT-2009-5). The author is solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

Author's email: andre@informatik.uni-augsburg.de.

DOI: http://dx.doi.org/10.1145/2502433

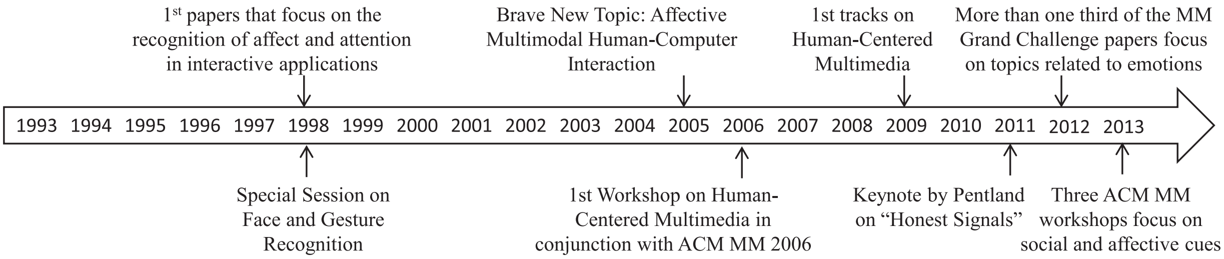
1st papers that focus on the recognition of affect and attention in interactive applications
Brave New Topic: Affective Multimodal Human-Computer Interaction
1st tracks on Human-Centered Multimedia
More than one third of the MM Grand Challenge papers focus on topics related to emotions
1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
Special Session on Face and Gesture Recognition
1st Workshop on Human-Centered Multimedia in conjunction with ACM MM 2006
Keynote by Pentland on "Honest Signals"
Three ACM MM workshops focus on social and affective cues

Fig. 1. Twenty years of ACM multimedia: work related to affective computing and social signal processing.

integrated facial expressions using surface electromyography and vocal emotions was presented and demonstrated the superiority of multimodal fusion over unimodal recognition techniques [Cohn and Katz 1998].

After 1998, the interest in methods for analyzing attention and affect was steadily growing, focusing on applications such as multimedia meeting browsers [Stiefelhagen et al. 1999] or interactions with embodied agents [Lisetti and Nasoz 2002]. A number of interactive affective applications were presented within the art programme that accompanied ACM Multimedia. Characteristic of these applications was their focus on aesthetic emotions as opposed to basic emotions [Ekman 1999] that were dominantly used in the more technically oriented work. Some of them even explored the idea of relying on emotions as the only channel of expression. A typical example was an artistic installation of an augmented reality tree which responded to the spectators' spontaneous emotional reactions by dynamically changing its appearance [Gilroy et al. 2008].

At ACM Multimedia 2005, a brave new topic session on affective multimodal human-computer interaction [Pantic et al. 2005] was organized, attracting papers from leading researchers in the field. Interactive applications were presented that made use of techniques for the analysis of multimodal human signals, such as tone of voice, facial expressions, gestures or body postures, to determine the engagement of children in a learning scenario [Kapoor and Picard 2005] or to characterize the dynamics of social groups by measurable features, such as speaking time or prosodic variation [Pentland 2005]. The paper by Pentland already indicated the move from affective human-computer interaction to the broader and even more challenging field of socially-aware computing.

At ACM Multimedia 2008, social signal processing was chosen as one out of three brave new topics and introduced in a survey paper by Vinciarelli et al. [2008] that investigated a broad range of human behavioral signals, such as mutual gaze, interpersonal distance, and mirroring. Around the same time, the need to give more emphasis to the human perspective in multimedia computing was recognized, leading to the 1st Workshop on Human-Centered Multimedia in conjunction with ACM Multimedia 2006 and special tracks on Human-Centered Multimedia in 2009, see Jaimes et al. [2006] for an introduction to this field. The growing interest of the multimedia community in the field of socially-aware computing was also reflected by Pentland's keynote "Honest Signals" at ACM Multimedia 2011.

Overall, the analysis of affective and social signals has not been mainstream at ACM Multimedia. However, hardly any other topic has attracted more attention as one of the big challenges that still need to be solved. Furthermore, advances in this area would not have been possible without methods from multimedia computing, such as video and audio processing. As technology for analyzing affective and social signals is getting more robust, time has come to explore an exciting new idea. Usually, affective and attentive signals are not deliberately communicated by users. Nevertheless, they may reveal information that may be used to adjust interfaces or services to them. For example, users normally do not consciously control their eye movements. Nevertheless, unconsciously provided gaze information could be employed as an indicator of user preferences or interest [Buscher et al. 2012]. We propose a new form of human-computer interaction that is based on the concept of empathic stimulation and exploits not only attentive but also affective signals. Unconsciously provided behavioral cues are employed to smoothly adapt a system without making users aware of the fact that they are implicitly controlling it.

A classical field of ACM Multimedia that might benefit from unconscious human-computer interaction is multimedia data retrieval and exploration. Size and complexity of databases nowadays make it increasingly difficult to make sense of the explored data, not only for novices, but also for experts. Within the CEEDS project, we are currently exploring the concept of empathic stimulation as a means to optimize the selection and presentation of data [Wagner et al. 2013]. To this end, users are immersed in a mixed-reality space, allowing them to explore complex data while freely moving around and interacting with the physical and virtual world. However, apart from recording explicit user input (e.g., gestures, motion, verbal commands), wearable sensors are used to capture implicit responses (e.g., heart rate, electrodermal activity, and gaze behaviour). The implicit responses are exploited to guide the users' discovery of patterns in the data space, for example, by directing their attention to relevant areas in the data space.

The realization of an empathic feedback loop comes with a lot of challenges for the fields of human computer interaction and multimedia computing. In a cross-disciplinary effort, the following issues need to be addressed.

—*Recognizing Unconscious User Signals.* Existing research in the area of social signal processing may serve as a starting point for the recognition of unconscious user signals. However, to meet the requirements of a smooth interaction between a human and a machine, a number of unrealistic assumptions have to be given up. Traditionally, research has concentrated on posteriori analyses of social cues under laboratory-like conditions. Such an approach leads, however, to over-optimistic assessments of recognition rates that cannot be reproduced in naturalistic settings. A common example includes voice data from actors for which developers of emotion recognition systems reported accuracy rates of over 80% for seven emotion classes [Vogt and André 2005]. In realistic applications, there is, however, no guarantee that emotions are expressed in a prototypical manner. To cope with nonprototypical emotional behaviors, recognition algorithms should no longer be based on hard labels, but rather on emotion profiles [Mower et al. 2011]. Furthermore, unconscious user signals have to be analyzed in (near) real-time while the user is interacting with a system. Consequently, we can no longer rely on global statistics for recognition tasks. Instead, a recognition result has to be provided at each increment in time.

—*Responding to Unconscious Cues.* Unconscious user signals offer powerful methods for personalizing a user interface. However, the decisive question is how and when to exploit unconscious user signals. Obviously, not every user signal should trigger a system response. Rather, a system needs to find the right level of sensitivity. As an example, let us consider a system that adapts presentations to the user's interest based on eye gaze. Basically, such a system would have to avoid two risks. First, it should not show exaggerated attentiveness by modifying the presentation at each fixation of the eyes that indicates potential user interest. Second, the adaptation should not interfere with the natural presentation flow. Apart from the fact that interruptions of the presentation flow may be perceived as disturbing, there is also the danger that badly integrated adjustments confuse users because they did not initiate them deliberately. To avoid such issues, it would be desirable to control human-computer interaction beyond user awareness. A promising idea to explore is a modified form of active learning that extracts new data points from unconsciously provided signals, such as physiological data.

—*Developing Evaluation Metrics.* Unconscious processes are typically not reflected by easy-to-interpret observable behaviors. Thus, it is not obvious how to acquire ground truth data against which to evaluate the performance of emotion recognition components. In the area of social signal processing, a lot of effort has been spent to systematically label multimodal corpora based on validated annotation schemes that typically follow an underlying emotional model [André 2011]. An evaluation is performed by comparing labels manually created by human raters with the labels provided by the automated recognition process. If unconscious processes are reflected by visible cues, a similar approach may be employed to obtain ground truth data. Alternatively, one might make use of physiological recordings with the caveat that they are hard to interpret. Besides measuring the robustness of recognition components, the effects of unconscious interaction as a whole have to be assessed. The evaluation criteria depend on the envisioned application. For example, the effect of unconscious human-computer interaction on data exploration could be measured by assessing the users' ability to discover certain patterns in the data space with and without considering their unconsciously conveyed signals.

REFERENCES

ANDRÉ, E. 2011. Experimental methodology in emotion-oriented computing. *IEEE Perv. Comput. 10*, 3, 54–57.

BUSCHER, G., DENGEL, A., BIEDERT, R., AND ELST, L. V. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Trans. Interact. Intell. Syst. 1*, 2, Article 9, 30 pages.

COHN, J. F. AND KATZ, G. S. 1998. Bimodal expression of emotion by face and voice. In *Proceedings of the 6th ACM International Conference on Multimedia: Face/Gesture Recognition and their Applications (MULTIMEDIA'98)*. ACM, New York, NY, 41–44.

EKMAN, P. 1999. Basic emotions. In *Handbook of Cognition and Emotion*, Wiley, 45–60.

GILROY, S. W., CAVAZZA, M., CHAIGNON, R., MÄKELÄ, S.-M., NIRANEN, M., ANDRÉ, E., VOGT, T., URBAIN, J., BILLINGHURST, M., SEICHTER, H., AND BENAYOUN, M. 2008. E-tree: Emotionally driven augmented reality art. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*. ACM, New York, NY, 945–948.

JAIMES, A., SEBE, N., AND GATICA-PEREZ, D. 2006. Human-centered computing: a multimedia perspective. In *Proceedings of the 14th Annual ACM International Conference on Multimedia (MULTIMEDIA'06)*. ACM, New York, NY, 855–864.

KAPOOR, A. AND PICARD, R. W. 2005. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM International Conference on Multimedia (MULTIMEDIA'05)*. ACM, New York, NY, 677–682.

LISETTI, C. L. AND NASOZ, F. 2002. MAUI: A multimodal affective user interface. In *Proceedings of the 10th ACM International Conference on Multimedia (MULTIMEDIA '02)*. ACM, New York, NY, 161–170.

MOWER, E., MATARIC, M. J., AND NARAYANAN, S. S. 2011. A framework for automatic human emotion classification using emotion profiles. *IEEE Trans. Audio Speech Lang. Process. 19*, 5, 1057–1070.

NAKATSU, R. 1998. Nonverbal information recognition and its application to communications. In *Proceedings of the 6th ACM International Conference on Multimedia: Face/Gesture Recognition and their Applications (MULTIMEDIA '98)*. ACM, New York, NY, 2–9.

PANTIC, M., SEBE, N., COHN, J. F., AND HUANG, T. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*. ACM, New York, NY, 669–676.

PENTLAND, A. 2005. Socially aware media. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA'05)*. ACM, New York, NY, 690–695.

STIEFELHAGEN, R., YANG, J., AND WAIBEL, A. 1999. Modeling focus of attention for meeting indexing. In *Proceedings of the 7th ACM International Conference on Multimedia (Part 1) (MULTIMEDIA'99)*. ACM, New York, NY, 3–10.

VINCIARELLI, A., PANTIC, M., BOURLARD, H., AND PENTLAND, A. 2008. Social signal processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*. ACM, New York, NY, 1061–1070.

Vogt, T. and André, E. 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'05)*. IEEE, Los Alamitos, CA, 474–477.

Wagner, J., Lingenfelser, F., André, E., Mazzei, D., Tognetti, A., Lanatà, A., De Rossi, D., Betella, A., Zucca, R., Omedas, P., and Verschure, P. F. M. J. 2013. A sensing architecture for empathetic data systems. In *Proceedings of the 4th Augmented Human International Conference*. ACM, New York, NY, 96–99.