J Multimodal User Interfaces (2014) 8:1–3
DOI 10.1007/s12193-014-0152-5

EDITORIAL

From multimodal analysis to real-time interactions with virtual agents

Ronald Poppe · Ronald Böck · Francesca Bonin · Nick Campbell · Iwan de Kok · David Traum

Published online: 00 0000
© OpenInterface Association 2014

1 Introduction

One of the aims in building multimodal user interfaces is to make the interaction between user and systems as natural as possible. Possibly the most natural form of interaction we know is the way we communicate with other humans. By building virtual agents, we aim to recreate this natural form of interaction in human–machine communication. This is even more important for virtual agents that communicate with humans in a real-time face-to-face setting.

While the promises of such natural interfaces are long-standing [1,9], their development is not straightforward. Understanding of human–human interaction is needed to an extent that it can be detected, modeled and generated by a system. This is not only challenging due to the large variation in human communicative behavior, but also due to the requirement that the interactions with humans and a system should be in real-time. Although challenging, the goal of a system interacting as a companion seems attainable.

The development of virtual agent systems capable of recreating natural interactions with humans typically involves several fundamental steps: recording and analyzing natural interaction data, extracting and recognizing relevant multimodal features, crafting or learning models from these features, generating the appropriate behavior in real-time based on these models and evaluating the system in a methodologically sound experiment. The papers in this special issue advance the state of the art for these different stages:

- Analysis of natural interactions: [2,4,7,10]
- Dialog modeling: [3,6,11,13]
- Experimental design: [5]

2 Analysis of natural interactions

Collecting multimodal natural corpora for analyzing natural interactions presents many challenges, such as capturing naturalistic behavior, enhancing the corpus with reliable ground truth annotations and dealing with the inter-personal differences in behavior. The collection of multimodal datasets and their analysis is a fundamental step for understanding human behavior and for finding features that can be extracted in real-time from the recorded signals. Such features include, but are not limited to, speech and its content, prosodic and paralinguistic features, eye gaze, facial expressions, body movements, or more advanced interpretations of such features such

R. Poppe (✉) · I. de Kok
Human Media Interaction Group, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: r.w.poppe@utwente.nl

I. de Kok
e-mail: i.a.dekok@utwente.nl

R. Böck
Cognitive Systems Group, Otto von Guericke University,
Universitaetsplatz 2, 39106 Magdeburg, Germany
e-mail: ronald.boeck@ovgu.de

F. Bonin
Computational Linguistic Group and Speech Communication Lab,
Trinity College Dublin, Dublin 2, Ireland
e-mail: boninf@tcd.ie

N. Campbell
Speech Communication Lab, Trinity College Dublin,
Dublin 2, Ireland
e-mail: nick@tcd.ie

D. Traum
USC Institute for Creative Technologies, 12015 Waterfront Drive,
Playa Vista, CA 90094, USA
e-mail: traum@ict.usc.edu

as the affective state, personality, mood or intentions of the user (e.g. [8,14,15]).

In order to fully exploit a multimodal dataset, metadata and annotations are essential. Manually labeling audio-visual data of human behavior is a subjective process as different human coders might not always agree. Moreover, the process is time-consuming as it involves manual effort, with annotators often going through the material several times. In this special issue, Schels et al. [7] and Siegert et al. [10] address these issues. The former investigate and discuss the subjectivity problem, presenting a discussion to reason about inter-rater reliability in the context of the annotation of emotions from audio and/or video. Schels et al. [7] address the time consumption issue. There is typically a trade-off between the amount of coded material and the accuracy of automatic classifiers trained on these data. In an attempt to mitigate this effect, the authors use unlabeled material in addition to a small sample of labeled data to improve the classification of emotional states from physiological data.

Human–human interactions are often seen in a multimodal setting. This face-to-face situation of two dialog partners is often the same for interactions with virtual agents. Therefore, the understanding of behavior from multimodal input deserves special attention. In this context, Lefter et al. [4] propose the analysis of stressful situations from both speech and hand gestures from a newly recorded corpus of videos. Further, they present material which is rich of naturalistic, prototypical interactions that can be used to derive hypotheses how different modalities influence each other. A multimodal analysis is offered also by De Carolis and Novielli [2], who perform corpus analyses to arrive at a model that infers social attitude in a dialog from language, prosodic and gesture cues.

3 Dialog modeling

The automatic and real-time detection, modeling and generation of the communicative behavior from a virtual agent's point of view are three aspects that are most suitably studied jointly. In this special issue, particular attention is given to the analysis of the human–human and human–machine communicative behavior from a conversational point of view. Regarding the analysis of human–human interactions, Visser et al. [13] present a comprehensive model for the analysis of the conversational grounding, the process of establishing common ground between dialog partners. They consider modalities such as speech and gesture in human–human interaction. Their empirical model is based on the work by Traum [12]. On the other hand, regarding the analysis of human–machine interactions, Prylipko et al. [6] investigate specific events within the dialog with the aim of improving the detection of human reactions to the system. Differences are found between age and gender groups in naturalistic interactions.

In modeling spoken conversations, one needs to take in consideration that, in addition to *what* is said, it is essential to deliver also *how* something is said. Linguistic and paralinguistic aspects are both key in the unfolding construction of the conversation. Szekely et al. [11] present an approach for speech-to-speech translation that explicitly addresses maintaining the paralinguistic information. To this end, facial expression analysis is used to analyze the affective state of the speaker and to adjust the generation of the speech accordingly.

Content and paralinguistic features of the speech, appropriate accompanying gestures and facial expressions constitute virtual agents' responses in an interaction. To perform its response, the virtual agent needs to reason about, plan and realize the actions with the correct timing. Timing is a key factor in human–human and human–machine interactions. Continuous perception, interpretation, reasoning and generation are required to keep the interaction between the user and the virtual agent as natural and fluent as a human–human interaction. Kopp et al. [3] address this issue by presenting an architecture to fluidly adapt the timing of the generation based on (partially) processed input.

4 Experimental design

Finally, this special issue's papers contribute to the research methodologies for the development and evaluation of real-time continuous virtual agent systems. Many aspects can influence the evaluation of the system as a whole and individual aspects of the system may require continuous evaluation as well. To this end, Poppe et al. [5] introduce a methodology to evaluate the systematic variation of the behavior of a virtual agent in an online dialog setting with a human. This will also lead to data sets which provided naturalistic interactions.

5 Outlook

With the advances described in the papers in this special issue, we are confident that the promise of having real-time conversations between humans and virtual agents has come closer. While some challenges remain, we have a better understanding, tools and methodology to address them. We hope that this special issue will inspire others to work on the topic, bringing virtual agents from the lab into the real world.

References

1. Cassell J, Prevost S, Sullivan J, Churchill E (2000) Embodied Conversational Agents. MIT Press, Cambridge

2. De Carolis, B., Novielli, N.: Recognizing signals of social attitude in interacting with ambient conversational systems. J. Multimodal User Interf. doi:10.1007/s12193-013-0143-Y (this issue)
3. Kopp, S., van Welbergen, H., Yaghoubzadeh, R., Buschmeier, H.: An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. J. Multimodal User Interf. doi:10.1007/s12193-013-0130-3 (this issue)
4. Lefter, I., Burghouts, G.J., Rothkrantz, L.J.: An audio-visual dataset of human–human interactions in stressful situations. J. Multimodal User Interf. doi:10.1007/s12193-014-0150-7 (this issue)
5. Poppe, R., ter Maat, M., Heylen, D.: Switching wizard of Oz for the online evaluation of backchannel behavior. J. Multimodal User Interf. doi:10.1007/s12193-013-0131-2 (this issue)
6. Prylipko, D., Rösner, D., Siegert, I., Günther, S., Friesen, R., Haase, M., Vlasenko, B., Wendemuth, A.: Analysis of significant dialog events in realistic human–computer interaction. J. Multimodal User Interf. doi:10.1007/s12193-013-0144-X (this issue)
7. Schels, M., Kächele, M., Glodek, M., Hrabal, D., Walter, S., Schwenker, F.: Using unlabeled data to improve classification of emotional states in human computer interaction. J. Multimodal User Interf. doi:10.1007/s12193-013-0133-0 (this issue)
8. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. 1(2):119–131
9. Shneiderman B, Plaisant C (2010) Designing the User Interface: Strategies for Effective Human–computer Interaction. Addison-Wesley, Boston
10. Siegert, I., Böck, R., Wendemuth, A.: Inter-rater reliability for emotion annotation in human–computer interaction. J. Multimodal User Interf. doi:10.1007/s12193-013-0129-9 (this issue)
11. Szekely, E., Steiner, I., Ahmed, Z., Carson-Berndsen, J.: Facial expression-based affective speech translation. J. Multimodal User Interf. doi:10.1007/s12193-013-0128-X (this issue)
12. Traum, D.: A computational theory of grounding in natural language conversation. Ph.D. thesis, University of Rochester, Rochester (1994)
13. Visser, T., Traum, D., DeVault, D., op den Akker, R.: A model for incremental grounding in spoken dialogue systems. J. Multimodal User Interf. doi:10.1007/s12193-013-0147-7 (this issue)
14. Vogt, T., André, E.: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: IEEE International Conference on Multimedia and Expo, pp. 474–477. IEEE, New Jersey (2005)
15. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31(1):39–58