Cogn Comput (2012) 4:383–385
DOI 10.1007/s12559-012-9180-1

Cognitive and Emotional Information Processing for Human–Machine Interaction

Stefano Squartini · Björn Schuller · Amir Hussain

Received: 10 July 2012 / Accepted: 1 August 2012 / Published online: 18 August 2012

Human–machine interaction (HMI) has been widely addressed in the literature and also encountered strong commercial interest in the last two decades: a number of advanced solutions have been proposed in a range of diverse application fields such as user/web interfaces, mobile computing and computer graphics and, from a wider perspective, robotics, ambient intelligence, entertainment and computer support to collaborative work/learning.

The basic HMI goal aims at improving the interactions between users and machines by making machines themselves more usable and receptive to the user's needs. This essentially includes studies on models and theories of interaction, on methodologies and processes for designing interfaces, on developing suitable techniques for evaluating and comparing interfaces and, of course, on exploring new hardware devices and software frameworks.

In particular, HMI researchers have been working hard to maximize the naturalness of the interaction itself by reducing the gap between the cognitive and emotional model behind human behavior and the machine awareness of "what is going on" during the task accomplishment. This asks for development of expert systems, able to manage large amount of information coming from sensory activity, to intelligently process it, and to promptly and knowledgeably respond to human actions according to natural interaction standards and by means of suitable actuary devices. Information processing therefore plays a central role from this perspective, operating at different levels, from multimodal digital data manipulation to semantic meta data processing, and necessarily encompassing the most challenging computational intelligence paradigms for contextual adaptation, social-emotional competence, and cognitive reasoning abilities.

This Special Issue is fully devoted to propose the most recent and stimulating advances within the multidisciplinary area of cognitive and emotional information processing for HMI. It collects eight original contributions, which cover some of the aforementioned topics providing to the reader an insightful panoramic view of the research achievements and open issues in the field of interest. The present papers are the result of a rigorous review procedure applied to the fifteen articles initially submitted. At least three independent experts have been involved for each paper (more than 50 in total) and up to four review rounds have been performed before final acceptance for publication.

The first two contributions deal with the processing of one of the widely used medium signals in HMI applications, namely speech, in order to suitably detect spoken activity in interactive environments.

In the work of Principi et al., an innovative person activity detection algorithm operating in reverberated environments and in the presence of multiple sources is proposed. It is composed of two main stages, both operating in real-time: the speech enhancement front-end and the activity detector. The former is able to reduce the distortions introduced by room reverberation, thus improving the speech quality for each speaker and consists

S. Squartini (✉)
Department of Information Engineering, Università Politecnica
delle Marche, Via Brecce Bianche 1, 60131 Ancona, Italy
e-mail: s.squartini@univpm.it

B. Schuller
Institute of Human–Machine Communication, Technische
Universität München, Arcisstrasse 21, 80333 Munich, Germany

A. Hussain
Department of Computing Science and Mathematics,
University of Stirling, Stirling FK9 4LA, Scotland, UK

in three cooperating blocks, which fulfill three fundamental tasks: speaker diarization, room impulse responses identification, and speech dereverberation. The activity estimation algorithm is based on bidirectional long short-term memory networks, performing context-sensitive activity classification from audio features computed via the real-time speech feature extraction toolkit open SMILE. The authors report in their paper the results obtained through an extensive experimental session where the AMI database has been employed, showing the positive behavior of the proposed algorithmic framework.

The second work by Kim et al. advances a novel speech emotion recognition framework for affective HMI, where the machine is a personal device. The authors base their work on the observation that most of the conventional techniques in this field adopt a speaker-independent model framework due to the sparseness of individual speech data. However, when dealing with a personal device, a large amount of individual data can be accumulated over time, and this allows for building suitable emotion models in accordance with a speaker adaptation procedure. The authors thus advance a modified version of the maximum likelihood linear regression approach, on the basis of selective label refinement, and are also able to iteratively take the accumulated individual data into account, in order to maximize the emotion modeling performance. Computer simulations carried out on a suitable emotional corpus allowed the authors to positively conclude about the appropriateness of their approach, especially in light of the performance attainable with conventional techniques.

Moving to the work of Jiu et al., an interesting Computational Intelligence approach for supervised codebook learning and optimization for bag of words models is investigated. Computer vision is the natural field of application of their algorithm, and in particular, it can deal with visual recognition tasks such as object class recognition or human action recognition. In these applicative scenarios, one is asked to recognize a certain entity, which is usually represented as a histogram of codewords that are traditionally clustered with unsupervised methods and then classified in a supervised way. The authors' idea consists in jointly accomplishing the codebook creation and the class learning by means of a supervised approach, which learns the cluster centers of the codebook in a goal-directed way using the class labels of the training set. As a result, the codebook is highly correlated to the recognition task under test and thus reveals much more discriminative potentialities with respect to conventional unsupervised approaches. The new method allows reducing the computational complexity if the same discriminative power of standard algorithms needs to be maintained. Some experiments have been carried out to assess the effectiveness of the approach: In particular, the algorithm was tested in a human action recognition task from video sequences, and encouraging results have been registered.

Then, this issue presents three relevant contributions, which deal with semantic-based processing for cognitive and emotional information extraction and manipulation in various scenarios. First, Das et al. propose a new method for tagging of sentence-level emotion and valence by using the SemEval 2007 affect sensing news corpus. In this work, three individual systems are developed on purpose. First, the baseline system operates such that each emotion class decides on the class label assigned to each word, without including any knowledge regarding word features. Second, a system based on the emotion lexicon WordNet Affect is considered, where updates from the *synsets* retrieved from the SentiWordNet and the morphology information for stemming purposes are also included. Then, a conditional random field (CRF)-based machine learning framework is employed for identifying word-level emotional expressions. Some computer simulations have been performed to the three approaches into relation: The proposed CRF-based one outperforms the others both at word and sentential levels. Similar conclusions are drawn in terms of valence, which is identified on the basis of the total sense scores of the word-level emotion tags along with their polarity.

Secondly, the work of Zhang addresses a particularly challenging and relevant cognitive topic in the field of affective computing: the suitable detection of significant context to inform affect detection. The author proposes a context-based affect detection algorithm and evaluates its performance by embedding it in an improvisational virtual platform, which allows up to five human characters and one intelligent agent to be engaged in one session to conduct creative improvisation within loose scenarios. In order to detect affect from such contexts, first a naïve Bayes classifier based on linguistic cues is used to categorize the simulated conversations. Then, two machine learning techniques have been tailored to provide affect detection, respectively, in the social and personal emotion contexts. The personal context affect interpretation uses the emotional history of each individual character to inform affect analysis, while the social context affect detection takes interpersonal relationships, sentence types, emotions implied by the potential target audiences in their most recent interactions, and discussion topics into account. The several experimental tests conducted by the author confirm the effectiveness of the approach.

Thirdly, D'Errico et al. deal with the analysis of cognitive and emotional aspects in a particular type of human interaction: the political debate. As the authors point out, in this kind of scenario, the persuader, beside bearing logical arguments and triggering emotions, presents herself/himself as a credible and reliable person, by enhancing assumed capabilities as competence, benevolence and dominance, and at the same time, s/he tries to discredit the opponent by criticizing, accusing or insulting. In this work, a description and a typology of multimodal discrediting moves is thus discussed. Based on an Italian corpus of political debates, the analysis carried out by the authors highlights which facial expressions, gaze behavior, gestures, postures, and prosodic features are used to convey discredit concerning the three target features of competence, benevolence, and dominance. Moreover, the effects of the different types of discrediting moves on potential electors are also experimentally assessed. Obtained results allow the authors to conclude that, in order to make the politician's arguments more shareable and convincing, s/he is supposed to cast discredit on the other's competence while performing gestures and on the other's dominance without gesturing.

Finally, there are the contributions by Cambria et al. and Grassi et al., which explore the development of advanced semantic tools, including innovative cognitive and emotional cues, to enhance the exploitation of multimedia content. The former contribution deals with the online personal photo management topic, which to the present day asks for further development in annotating, organizing and retrieving pictures available on the web so that they can be easily queried and visualized. As the authors observe, "existing content-based image retrieval systems apply statistics, pattern recognition, signal processing and computer vision techniques but these are still too weak to 'bridge the semantic gap' between the low-level data representation and the high-level concepts the user associates with images". Current Image meta search engines mainly depend on keyword-based rather than concept-based algorithms: That is why, the authors propose a new solution called Sentic Album that lets the user manage her/his personal photographs by suitably exploiting both data and metadata of online personal pictures, to intelligently annotate, organize, and retrieve them. To be more specific, Sentic Album uses not only colors and texture of online images (*content*), but also the cognitive and affective information associated with their metadata (*concept*), and their relative timestamp, geolocation, and user interaction metadata (*context*).

The last contribution of the issue describes a semantic web-based annotation system for multimedia resources, with special focus on video. The recent advent of Web 2.0 and the spreading of video sharing services over the Web have led to an explosion of online video content, which needs to be accessed and explored. In this context, the value of semantically structured data and metadata is recognized as a key factor both to improve search efficiency and to guarantee data interoperability. On the other hand, the annotation of video resources has been increasingly understood as a medium factor to enable in depth analysis of contents and collaborative study of online digital objects. However, as existing annotation tools provide poor support for semantically structured content, these objectives are hardly reached. The tool proposed by the authors (developed within the SemLib EU project) enables user annotations to form semantically structured knowledge at different levels of granularity and complexity. Annotation can be reused by external applications and mixed with web of data sources to enable "serendipity", the reuse of data produced for a specific task by different people and in different contexts from the one data originated from. The authors provide an interesting implementation of an online video annotation tool, with the aim to demonstrate how such technologies can enable a scenario where users' annotations are created while browsing the Web, naturally shared among users, stored in machine-readable format and then possibly recombined with external data and ontologies to enhance end-user experience.

To conclude, as guest editors, we would like to thank all authors for their contributions to this special issue. We also would like to express our sincere appreciation to all reviewers for their time and efforts. Finally, our gratitude goes to the Cognitive Computation Editorial board for their substantial support in the whole organizing procedure.