Neurocomputing 161 (2015) 17–37

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Fusion paradigms in cognitive technical systems for human–computer interaction

Michael Glodek [a], Frank Honold [b], Thomas Geier [c], Gerald Krell [d], Florian Nothdurft [e], Stephan Reuter [f], Felix Schüssel [b], Thilo Hörnle [c], Klaus Dietmayer [f], Wolfgang Minker [e], Susanne Biundo [c], Michael Weber [b], Günther Palm [a], Friedhelm Schwenker [a]

[a] Institute of Neural Information Processing, University of Ulm, Germany
[b] Institute of Media Informatics, University of Ulm, Germany
[c] Institute of Artificial Intelligence, University of Ulm, Germany
[d] Institute of Information Technology and Communications, Otto-von-Guericke University Magdeburg, Germany
[e] Institute of Information Technology, University of Ulm, Germany
[f] Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form
7 January 2015
Accepted 9 January 2015
Available online 27 February 2015

Keywords:
Cognitive technical systems
Scene understanding
Human–computer interaction

ABSTRACT

Recent trends in human–computer interaction (HCI) show a development towards cognitive technical systems (CTS) to provide natural and efficient operating principles. To do so, a CTS has to rely on data from multiple sensors which must be processed and combined by fusion algorithms. Furthermore, additional sources of knowledge have to be integrated, to put the observations made into the correct context. Research in this field often focuses on optimizing the performance of the individual algorithms, rather than reflecting the requirements of CTS. This paper presents the information fusion principles in CTS architectures we developed for Companion Technologies. Combination of information generally goes along with the level of abstractness, time granularity and robustness, such that large CTS architectures must perform fusion gradually on different levels — starting from sensor-based recognitions to highly abstract logical inferences. In our CTS application we sectioned information fusion approaches into three categories: perception-level fusion, knowledge-based fusion and application-level fusion. For each category, we introduce examples of characteristic algorithms. In addition, we provide a detailed protocol on the implementation performed in order to study the interplay of the developed algorithms.

1. Introduction

Modern computer systems are designed to improve on efficiency and user experience by dynamically adapting to situations, incorporating additional knowledge and enhancing the interaction. These features are realized by enabling the computer to perceive its environment, to extract relevant information and to compare this to previously acquired data. In the literature cognitive technical systems (CTS) are known as Companion Systems. State-of-the-art systems available on the market claim to provide these kinds of features, however they are still far below their possible potential, mostly due to the demanding information processing required [1,2].

Perception in a CTS can be divided into three major categories which are virtually omnipresent in any given human–computer interaction (HCI) setting: (1) the implicit user input [3] (e.g. emotion or disposition [4]); (2) the explicit user input (e.g. multimodal instructions by gesture and speech); and (3) the recognition of the user's environment as well as the context of use [5] (e.g. activities, state and manipulation of objects nearby). It must be emphasized that necessary perceptions usually strongly depend on the application at hand. Therefore, it is always important to identify the relevant and application-specific perceptions in a first step. In case of emotions, useful classes are not necessarily the most obvious ones, e.g. happiness or anger. In fact, to improve an interaction it is more beneficial to focus on negative user

E-mail addresses: michael.glodek@uni-ulm.de (M. Glodek), frank.honold@uni-ulm.de (F. Honold), thomas.geier@uni-ulm.de (T. Geier), gerald.krell@ovgu.de (G. Krell), florian.nothdurft@uni-ulm.de (F. Nothdurft), stephan.reuter@uni-ulm.de (S. Reuter), felix.schuessel@uni-ulm.de (F. Schüssel), thilo.hoernle@uni-ulm.de (T. Hörnle), klaus.dietmayer@uni-ulm.de (K. Dietmayer), wolfgang.minker@uni-ulm.de (W. Minker), susanne.biundo@uni-ulm.de (S. Biundo), michael.weber@uni-ulm.de (M. Weber), guenther.palm@uni-ulm.de (G. Palm), friedhelm.schwenker@uni-ulm.de (F. Schwenker).

http://dx.doi.org/10.1016/j.neucom.2015.01.076

dispositions being directly related to the system, like boredom or stress [2].

1.1. Fusion categories in cognitive technical systems

In CTS, the problems often arise from the endeavor to develop architectures which implement multiple requirements simultaneously. The perception of classes, especially of the second and the third category, is encumbered by the open world scenario in which unusual events may occur and classes often have a wide range of variability. This perception problem can be addressed by enriching the recognition approach with additional domain knowledge. However, what appears to be a straightforward solution entails many open research questions. The most important one is How to realize a seamless integration of symbolic and sub-symbolic information, also with respect of how to exchange information in both directions, i.e. from sensory to high-level representations and back. Furthermore, it is not sufficient that these algorithms perform well on convenient pre-segmented datasets but have to provide good results in real-time in ubiquitous applications. In turn, more requirements arise: how to (1) compensate sensor failures; (2) draw information from the temporal dimension; (3) take uncertainty into account; and (4) deal with the open world setting. However, the problems mentioned so far only address the perceptive periphery. Another central issue represents the combination of the uncertain perceptions with symbolic domain knowledge. The combination is crucial to enhance the recognition results and to bring them into the correct context. The integration of domain knowledge is also inevitable, since the recording of datasets covering all possible observations is usually infeasible.

In addition, the inferred classes are typically more abstract, which is helpful to create a truly relevant user history and which in turn is necessary to carry out a reasonable interaction. Indeed, the concept of combining explicit user inputs with knowledge about the ongoing and past behavioral patterns of the user bears another challenge. In a first step, an abstract input representation has to be derived from inputs of multiple modalities which then has to be combined with available knowledge, the dialog management and the application. In other words, an algorithm has to mediate between the user's input, acquired knowledge (possibly afflicted with uncertainty), goal priorities, and the interface provided by the application.

In the last instance, the CTS's fission component has to reason about how to provide a situation and input dependent output based on an abstract representation of the core system. The presented work addresses these challenges and shows ways how it is possible to solve them by taking advantages from information fusion methods.

Shifting the view from outer requirements to the characteristics of information, it becomes evident that the processing can be grouped into different stages. Fig. 1 exemplifies the stages and how the information is condensed by fusion. Algorithms close to the sensory usually recognize patterns which are directly observable in the scene, e.g. the presence or the identification and attributes of a person [6,7]. By adding spatial and temporal context, more meaningful classes can be derived, e.g. recognizing activities or the mood of a person [8,4]. In the next layer, relationship of entities can be taken into account, e.g. persons with respect to each other or connection between a person and objects [9]. Again, a large history of observations can allow the discovery of more complex attitudes and salient events. Ultimately, this kind of high-level information is of relevance for the application and interaction. We regard the decisions based on the high-level information as the last step of fusion. The requirements and characteristics motivate the partitioning of the fusion algorithms and architectures into these categories: *perception-level fusion*, *knowledge-based fusion* and *application-level fusion*. The new taxonomy will be used throughout the paper as an aid to orientation and explained in greater detail in the corresponding sections.

1.2. Architecture of cognitive technical systems

An alternative view on CTS is to take a closer look at its architecture design. The systematical decomposition of a CTS architecture is depicted in Fig. 2. The schema shows the exchange of information between the user and the system, where the red arrows represent the input and the blue arrows the output of the system. Basically, the system itself is organized in two basic blocks: (1) peripheral block consisting of the user interface and perception component; and (2) an inner block consisting of a knowledge model and associated components such as planning, ontologies and the application and dialog management itself.

The red input arrow in the lower right, which is leading into the perception component of the first basic block, represents the recognizers perceiving the environment and the intrinsic user state. The multimodal inputs, e.g. video cameras or microphones, are mapped to classes by the perception component. In case one class is recognized by multiple modalities, the perception-level fusion combines them to a single output. The perception component is connected to the knowledge model, not only to derive more sophisticated information but also to enhance the perception by back-propagating beliefs. This is achieved by the knowledge-base fusion, represented by the lower bi-directional arrow. On the upper right red and blue arrows represent the input and the output of the system, respectively. The commands, which are combined and interpreted by the user interface, are, if necessary, forwarded to the inner basic block which then adapts the knowledge model and planning accordingly. The bi-directional arrows in the system show that in the ideal case information is exchanged in both directions in a seamless manner. Therefore, the bi-directional arrows represent not only the fusion of information but also fission

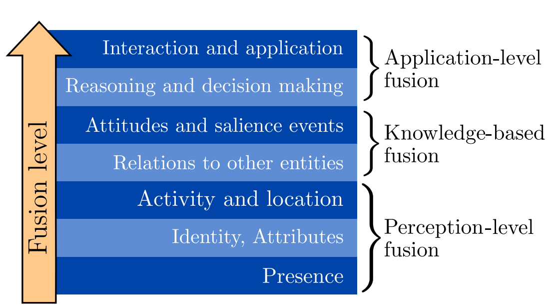

Fig. 1. *Information fusion in CTS grouped into three layers: perception-level fusion, knowledge-based fusion and application-level fusion.* The higher the layer of fusion, the more abstract the derived and processed knowledge. The procedure is usually accompanied with an increase of the temporal granularity and the variability of the occurrences covered.

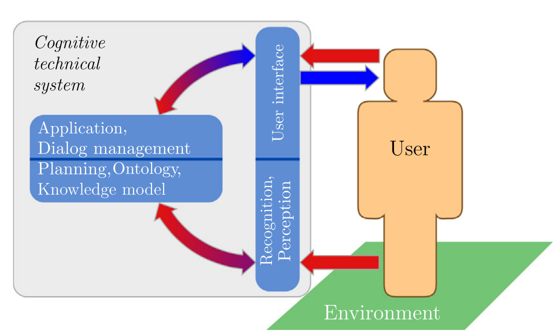

Fig. 2. *Architecture design of a CTS.* The CTS perceives two kinds of input, namely the implicit input (lower arrow) and the explicit input (upper arrow). Within the CTS, the information needs to be processed gradually with respect to the temporal granularity, the level of abstractness and uncertainty in order to allow a robust extraction. (For interpretation of the references to color in the text, the reader is referred to the web version of this paper.)