Multimodal Interfaces in Support of Human-Human Interaction

Alex Waibel

Interactive Systems Laboratories,
Carnegie Mellon University, USA
and
University of Karlsruhe, Germany
waibel@cs.cmu.edu
http://www.cs.cmu.edu/%7Eahw/

1 Extended Abstract

After building computers that paid no intention to communicating with humans, the computer science community has devoted significant effort over the years to more sophisticated interfaces that put the "human in the loop" of computers. These interfaces have improved usability by providing more appealing output (graphics, animations), more easy to use input methods (mouse, pointing, clicking, dragging) and more natural interaction modes (speech, vision, gesture, etc.). Yet all these interaction modes have still mostly been restricted to human-machine interaction and made severely limiting assumptions on sensor setup and expected human behavior. (For example, a gesture might be presented clearly in front of the camera and have a clear start and end time). Such assumptions, however, are unrealistic and have, consequently, limited the potential productivity gains, as the machine still operates in a passive mode, requiring the user to pay considerable attention to the technological artifact.

As a departure from such classical user interfaces, we have turned our attention to developing user interface for use in computing services that place Computers in the midst of Humans, i.e. in the Human Interaction Loop (CHIL), rather than the other way round. CHIL services aim to provide assistance implicitly and proactively, while causing minimal interference. They operate in environments, where humans interact with humans and computers hover in the background providing assistance wherever needed. Providing such services in real life situations, however, presents formidable technical challenges. Computers must be made aware of the activities, locations, interactions, and cognitive states of the humans that they are to serve and they must become socially responsive. Services must be delivered and provided in a private, secure, and socially acceptable manner.

CHIL services require perceptual technology that provides a complete description of human activities and interactions to derive and infer user needs, i.e., they must describe the WHO, WHERE, HOW, TO WHOM, WHY, WHEN of human interaction and engagement. Describing human-human interaction in open, natural and unconstrained environments is further complicated by robustness issues, when noise, illumination, occlusion, interference, suboptimal sensor positioning, perspective, localization and segmentation all introduce uncertainty. Relevant perceptual cues

S. Kopp and I. Wachsmuth (Eds.): GW 2009, LNAI 5934, pp. 243–244, 2010.

therefore must be gathered, accumulated and fused across modalities and along time opportunistically, i.e., whenever and wherever such cues can be determined and merged reliably. And finally, gathering of such multimodal cues, should involve a proactive participation of the interface to seek out such cues, as the interface may move (Humanoid Robots), coordinate (multiple sensors), and calibrate its own sensors and data gathering.

In this talk, ongoing work and results were presented from perceptual interfaces under development in realistic human-human interaction environments, using data from smart rooms and humanoid robot interaction.