Pattern Recognition Letters 36 (2014) 189–195

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Review Article

Multimodal interaction: A review

Matthew Turk *

Department of Computer Science, University of California, Santa Barbara, CA 93106-5110, United States

ARTICLE INFO

Article history:
Available online 17 July 2013

Communicated by Luis Gomez Deniz

Keywords:
Multimodal interaction
Perceptual interface
Multimodal integration
Review

ABSTRACT

People naturally interact with the world multimodally, through both parallel and sequential use of multiple perceptual modalities. Multimodal human–computer interaction has sought for decades to endow computers with similar capabilities, in order to provide more natural, powerful, and compelling interactive experiences. With the rapid advance in non-desktop computing generated by powerful mobile devices and affordable sensors in recent years, multimodal research that leverages speech, touch, vision, and gesture is on the rise. This paper provides a brief and personal review of some of the key aspects and issues in multimodal interaction, touching on the history, opportunities, and challenges of the area, especially in the area of multimodal integration. We review the question of early vs. late integration and find inspiration in recent evidence in biological sensory integration. Finally, we list challenges that lie ahead for research in multimodal human–computer interaction.

1. Introduction

Human interaction with the world is inherently multimodal (Bunt et al., 1998; Quek et al., 2002). We employ multiple senses, both sequentially and in parallel, to passively and actively explore our environment, to confirm expectations about the world and to perceive new information. We experience external stimuli through sight, hearing, touch, and smell, and we sense our internal kinesthetic state through proprioception. A given sensing modality may be used to simultaneously estimate several useful properties of one's environment – for example, audio cues may be used to determine a speaker's identity and location, to recognize the speaker's words and interpret the prosody of the utterance, to estimate the size and other characteristics of the surrounding physical space, and to identify other characteristics of the environment and simultaneous peripheral activities. Multiple sensing modalities give us a wealth of information to support interaction with the world and with one another.

In stark contrast to human experience with the natural world, human–computer interaction has historically been focused on unimodal communication – i.e., information or data communicated between human and computer primarily through a single mode or channel, such as text on a screen with a keyboard for input. While, technically, almost all interaction with computers has been multimodal to some degree – combining typed text with switches, buttons, mouse movement and clicks, and providing various visual and auditory output signals (including unintentional but useful audio cues such as the sound of a hard drive being accessed) – for much of interactive computing's history, the model of a single primary channel for data input, and perhaps a different primary channel for data output, has been the norm.

Multimodal interfaces describes interactive systems that seek to leverage natural human capabilities to communicate via speech, gesture, touch, facial expression, and other modalities, bringing more sophisticated pattern recognition and classification methods to human–computer interaction. While these are unlikely to fully displace traditional desktop and GUI-based interfaces, multimodal interfaces are growing in importance due to advances in hardware and software, the benefits that they can provide to users, and the natural fit with the increasingly ubiquitous mobile computing environment (Cutugno et al., 2012). The goal of research in multimodal interaction is to develop technologies, interaction methods, and interfaces that remove existing constraints on what is possible in human–computer interaction, towards the full use of human communication and interaction capabilities in our interactions. This is an interdisciplinary endeavor that requires collaboration among computer scientists, engineers, social scientists, linguists, and many others who bring expertise to bear on understanding the user, the system, and the interaction.

There are good surveys available on various aspects of multimodal interaction – e.g., Jaimes and Sebe (2007) survey multimodal HCI research, with a particular emphasis on computer vision; Dumas et al. (2009) surveys multimodal principles, models, and frameworks; Lalanne et al. (2009) survey fusion engines for multimodal input.

2. A history of multimodal interaction

Richard Bolt's "Put That There" system (Bolt, 1980) is widely regarded as a groundbreaking demonstration that first communi-

* Tel.: +1 (805) 893 4236.
E-mail address: mturk@cs.ucsb.edu

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.07.003

cated the value and opportunity for multimodal interfaces. Bolt's group at the MIT Architecture Machine Group (later to become the Media Lab), built the Media Room, which integrated voice and gesture inputs to enable a user sitting in a chair to have a rather natural and efficient interaction with a wall display in the context of a spatial data management system (see Fig. 1). Commands such as "create a blue square there," "move that to the right of the green square," "make that smaller," and the canonical "put that there" illustrate the power of integrating modalities to resolve pronoun reference and eliminate ambiguity. None of these phrases can be interpreted properly from either the utterance or the gesture alone – both are required, but that multimodal combination (if interpreted correctly) creates a simple, expressive command that is natural for the user.

"Put That There" was followed by numerous systems that sought to integrate various aspects of speech and gesture in a range of application areas; speech-based systems drove the majority of multimodal interface research. These early multimodal systems were primarily focused on spatial tasks and map-based applications. Put That There was a spatial data management system. CUBRICON (Neal et al., 1989), which enabled a user to interact using spoken or typed natural language and gesture and displayed results using combinations of language, maps, and graphics, was in the context of map-based tactical mission planning. The Koons et al. (1993) system that integrated speech, gesture, and eye gaze used a map-based application. QuickSet (Cohen et al., 1997) was a pen/voice system running on an early tablet PC, used in the context of a US Marine Corps training simulator (see Fig. 2).

Alternative formulations also followed, bringing new modalities such as haptics and eventually mobile computing environments as a rich testbed for multimodality. While multimodal interaction can be viewed as expanding the traditional desktop experience, much of the focus in multimodal interaction has been on alternative, or "post-WIMP" computing environments. Van Dam (1997) described post-WIMP user interfaces as those moving beyond the desktop graphical user interfaces (GUI) paradigm, relying more on things

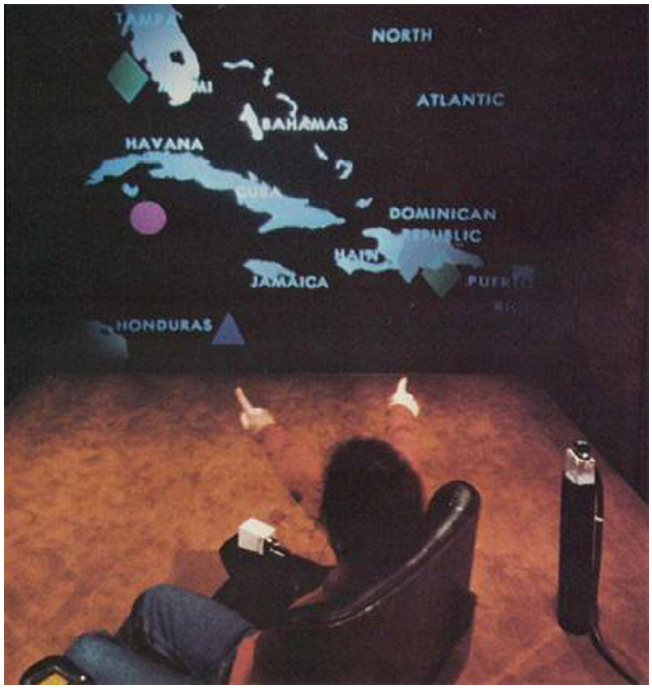

Fig. 1. Bolt's "Put That There" system (Bolt, 1980). (Photo by Christian Lischewski. Copyright 1980, Association for Computing Machinery, Inc. Used with permission.) [Intended for color reproduction].

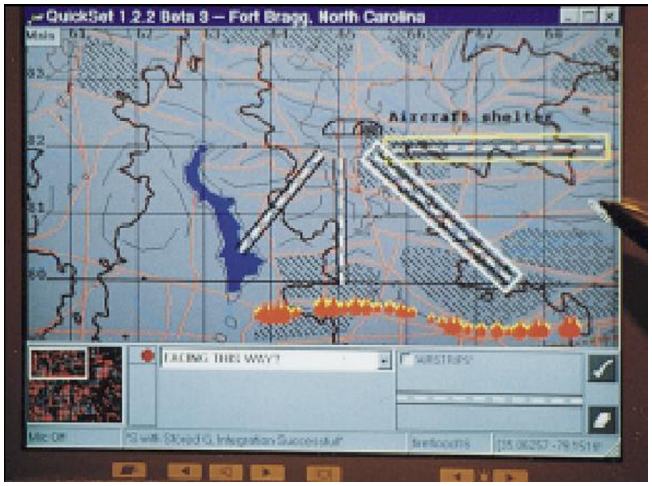

Fig. 2. The QuickSet tablet PC interface (Cohen et al., 1997). From Oviatt (1999) – reprinted with permission.

like speech, gesture, sketching, and 3D, though falling short of the longer-term vision of butler-like interfaces that understand the user's context, tastes, and idiosyncrasies and act accordingly, sometimes without needing explicit direction, just as a proper butler anticipates his employer's needs. Interaction with the "butler interface" will be more like interacting with a person, communicating via speaking, gesturing, facial expression, and other forms of human communication.

This view of post-WIMP interfaces with an eye towards more powerful "butler-like" interaction took on life in the push for "perceptual interfaces" (Turk, 1998; Turk and Robertson, 2000; Oviatt and Cohen, 2000; Turk and Kölsch, 2004), which seek to make the user interface more natural and compelling by taking advantage of the ways in which people naturally interact with each other and with the world, employing both verbal and non-verbal communications, along with interaction techniques that leverage an understanding of natural human capabilities (particularly communication, motor, cognitive, and perceptual skills) and employ machine perception and reasoning. Perceptual user interfaces (PUIs) are intended to be proactive multimodal interfaces, integrating perceptual capabilities into the human–computer interface. A series of PUI workshops began in 1997 and eventually merged with the International Conference on Multimodal Interfaces, which first met in 1996, to form a new ACM conference (keeping the ICMI name) that has become the premier venue for research in multimodal interaction. In recent years ICMI also merged with a European-focused workshop on machine learning and multimodal interaction (MLMI), expanding its focus and enlarging its community. As of 2013, the International Conference on Multimodal Interaction is an annual ACM meeting that showcases the state of the art in the field. In addition, a new ACM journal was founded in 2011, the Transactions on Interactive Intelligent Systems, that includes multimodal interaction as one of its core areas of focus.

3. Advantages of multimodal interaction

Multimodal interaction systems aim to support the recognition of naturally occurring forms of human language and behavior through the use of recognition-based technologies (Oviatt, 2003; Waibel et al., 1996). Multimodal interfaces are generally intended to deliver natural and efficient interaction, but it turns out that there are several specific advantages of multimodality. Although the literature on formal assessment of multimodal systems is still sparse, various studies have shown that multimodal interfaces may