Pattern Recognition Letters xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

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 xxxx

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

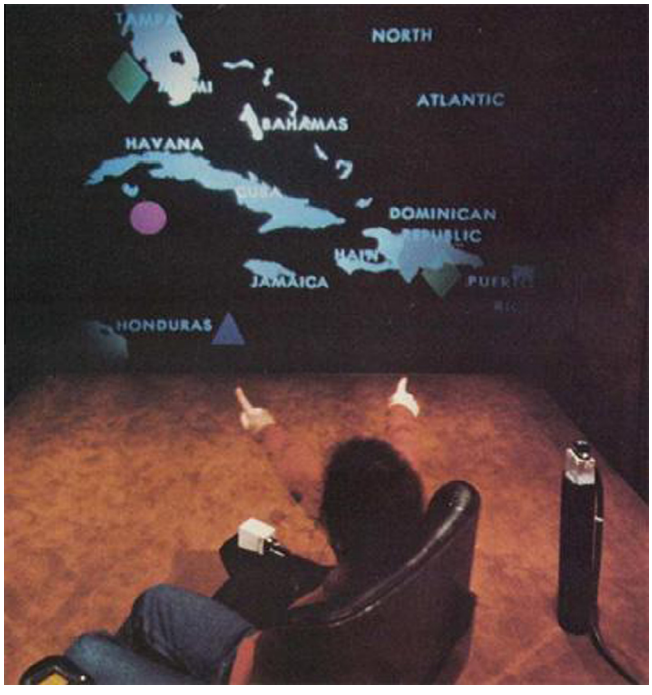

Fig. 1. Bolt's "Put That There" system (Bolt, 1980). (Photo by Christian Lischewski. Copyright 1980, Association for Computing Machinery, Inc. Used with permission.) [Intended for color reproduction].

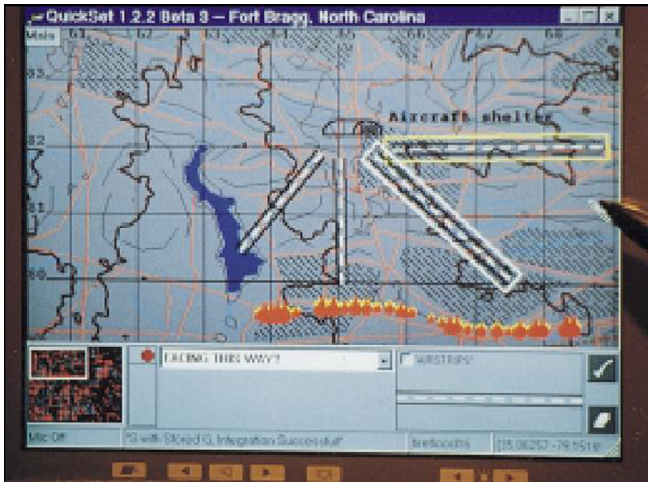

Fig. 2. The QuickSet tablet PC interface (Cohen et al., 1997). From Oviatt (1999) – reprinted with permission.

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

Please cite this article in press as: Turk, M. Multimodal interaction: A review. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.003

be preferred by users over unimodal alternatives, can offer better flexibility and reliability, can offer interaction alternatives to better meet the needs of diverse users with a range of usage patterns and preferences (Xiao et al., 2002; Xiao et al., 2003; Oviatt et al., 2005; Bohus and Horvitz, 2010). Multimodal interfaces can increase task efficiency, although perhaps not significantly, as pointed out by Dumas et al. (2009). Humans may process information faster and better when it is presented in multiple modalities (van Wassenhove et al., 2005). Other potential advantages of multimodal interfaces include the following (Oviatt et al., 2000):

- They permit the flexible use of input modes, including alternation and integrated use.
- They support improved efficiency, especially when manipulating graphical information.
- They can support shorter and simpler speech utterances than a speech-only interface, which results in fewer disfluencies and more robust speech recognition.
- They can support greater precision of spatial information than a speech-only interface, since pen input can be quite precise.
- They give users alternatives in their interaction techniques.
- They lead to enhanced error avoidance and ease of error resolution.
- They accommodate a wider range of users, tasks, and environmental situations.
- They are adaptable during continuously changing environmental conditions.
- They accommodate individual differences, such as permanent or temporary handicaps.
- They can help prevent overuse of any individual mode during extended computer usage.

While every combination of interface, task, user, and environment is different, and it is thus difficult to draw general conclusions for a whole category, the trend of existing studies points to a wide range of reasons that the pursuit of multimodal interfaces will be advantageous to users.

4. Input and output modalities

Some of the terms relevant to multimodal interaction – such as modes/modalities, channels, devices, multisensory, multimedia, and multimodal – have subtly or significantly different meanings in different communities. Blattner and Glinert (1996) addressed the terminology years ago; Table 1 updates their list of modalities and examples. In addition to input modalities listed in the table, emerging technologies such as indirect sensing of neural activity (e.g., brain–computer interfaces) may become practical components of multimodal interaction systems in the near future (see Leeb et al., 2013 for an example in the domain of virtual reality gaming).

Table 1
Human sensory modalities relevant to multimodal human–computer interaction, after Blattner and Glinert (1996).

Modality	Example
Visual	Face location
	Gaze
	Facial expression
	Lipreading
	Face-based identity (and other user characteristics such as age, sex, race, etc.)
	Gesture (head/face, hands, body)
	Sign language
Auditory	Speech input
	Non-speech audio
Touch	Pressure
	Location and selection
	Gesture
Other sensors	Sensor-based motion capture

Humans primarily interact with the world through their five major senses of sight, hearing, touch, smell, and taste. In perception, a mode or modality refers to receiving stimuli from a particular sense. A communication channel is a particular pathway through which information is transmitted. In typical HCI usage, a channel describes an interaction technique that utilizes a particular combination of user ability and device capability (such as the keyboard for inputting text, a mouse for pointing or selecting, or a 3D sensor used for gesture recognition). In this view, the following are all channels: text (which may use multiple modalities when typing in text or reading text on a monitor), sound, speech recognition, images/video, and mouse pointing and clicking. Multimodal interaction, then, may refer to systems that use either multiple modalities or multiple channels. Multimodal systems and architectures vary along several key dimensions or characteristics, including the number and type of input modalities; the number and type of communication channels; the ability to use modes in parallel, serially, or both; the size and type of recognition vocabularies; the methods of sensor and channel integration; and the kinds of applications supported.

In a system supporting multimodal input, the mapping from input modes and specific user actions to user intent – i.e., defining the vocabulary of ways to communicate a particular function, command, or parameter – is not straightforward. System designers tend to make such decisions based on intuition or some preliminary testing, but the appropriate assignment of multimodal input vocabulary to user intent is an open research question. See Ruiz et al. (2010) for an overview of multimodal input.

While the multimodal interaction community has focused more on input technologies such as speech and gesture recognition and haptic input, and multimodal output has been a key element of multimedia and visualization research communities, the overall goal of multimodal interaction is to fully support both directions of communication between human and machine – as well as to empower computer-supported human–human multimodal interaction. One small example of the latter is my lab's recent work on telecollaboration (Gauglitz et al., 2012), using computer vision and augmented reality techniques to improve remote collaboration among users.

Coincident with the growth of the multimodal interaction research community has been the commercial explosion of smartphones – powerful mobile devices that are well suited for multiple modes of interaction – as well as the maturing of speech recognition and speech understanding systems and the introduction of 3D vision sensors (such as the Microsoft Kinect and the Leap Motion Controller). Together, these advances have created a plethora of opportunity for multimodal interaction techniques and applications and have motivated researchers to leverage these opportunities. The ubiquity of smartphones has accelerated the jump to a post-WIMP world.

5. Biological sensory integration

While much is known about how biological systems sense and process sensory data, still more is unknown, especially about how sensory channels integrate with higher-level, cognitive processes and about how multiple sensory channels integrate at any level. In the past decade or so, there has been an influx of new discoveries and research in the area of crossmodal integration of sensory inputs – i.e., how some neurons respond to stimulation in more than one modality (Andersen, 1997; Calvert et al., 1998; Calvert, 2001; Lewkowicz and Kraebel, 2004; Cappe et al., 2012). Evidence is increasing that there are neurons in what are considered primary

Please cite this article in press as: Turk, M. Multimodal interaction: A review. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.003

sensory areas (e.g., area V1 in the visual cortex) where connections are made and outputs influenced by crossmodal sources – combining auditory and visual stimuli (Falchier et al., 2002; Arnal et al., 2009; Campi et al., 2010; Werner and Noppeney, 2011; Leitão et al., 2012), visual and tactile stimuli (Vasconcelos et al., 2011; Arabzadeh et al., 2008), visual and olfactory (Zhou et al., 2012) and others. Temporal issues of sensory integration are also becoming better understood (Powers et al., 2012; Swallow et al., 2012).

On a perceptual level, the well-known McGurk effect (McGurk and MacDonald, 1976) has provided a compelling example for decades of how auditory and visual stimuli can interact and affect the perception of an event, in some cases quite significantly. This knowledge has motivated the audio-visual speech processing (Stork and Hennecke, 1996; Chen, 2001) community and also motivated the search for equally compelling examples involving other modality pairs.

6. Designing and building multimodal interaction systems

Creating multimodal systems is challenging, as the typical design choices and intuitions from standard computing environments do not necessarily translate well to multimodal environments. Furthermore, each multimodal computing environment (combination of available modalities, application tasks, and user constraints) may warrant different design decisions. A set of multimodal myths and a set of multimodal design guidelines have proven to be useful both in designing systems and in considering research methods in the area.

Oviatt's "Ten Myths of Multimodal Interaction" (Oviatt, 1999) offers useful insights for those researching and building multimodal systems, with a few especially apropos:

- *Myth: If you build a multimodal system, users will interact multimodally.*Rather, users tend to intermix unimodal and multimodal interactions. Fortunately, multimodal interactions are often predictable based on the type of action being performed.
- *Myth: Multimodal input involves simultaneous signals.* Multimodal signals often do not co-occur temporally, and much of multimodal interaction involved the sequential (rather than simultaneous) use of modalities.
- *Myth: Multimodal integration involves redundancy of content between modes.* Complementarity of content may be more significant in multimodal systems than redundancy.
- *Myth: Enhanced efficiency is the main advantage of multimodal systems.* Multimodal systems may increase efficiency, but not always. Their main advantages may be found in other aspects, such as decreased errors, increased flexibility, or increased user satisfaction.
- *Myth: Individual error-prone recognition technologies combine multimodally to produce even greater unreliability.* In an appropriately flexible multimodal interface, people determine how to use the available input modes most effectively; mutual disambiguation of signals may contribute to a higher level of robustness.

Reeves et al. (2004) defined the following guidelines for multimodal user interface design:

- Multimodal systems should be designed for the broadest range of users and contexts of use. Designers should support the best modality or combination of modalities anticipated in changing environments (for example, private office vs. driving a car).
- Designers should take care to address privacy and security issues in multimodal systems. For example, non-speech alternatives should be available in a public context to prevent others from overhearing provide information or conversations.
- Maximize human cognitive and physical abilities, based on an understanding of users' human information processing abilities and limitations.
- Modalities should be integrated in a manner compatible with user preferences, context, and system functionality. For example, match the output to acceptable user input style, such as constrained grammar or unconstrained natural language.
- Multimodal interfaces should adapt to the needs and abilities of different users, as well as different contexts of use. Individual differences (for example, age, preferences, skill, sensory or motor impairment) can be captured in a user profile and used to determine interface settings.
- Be consistent – in system output, presentation and prompts, enabling shortcuts, state switching, etc.
- Provide good error prevention and error handling; make functionality clear and easily discoverable.

7. Multimodal integration

Multimodal integration – also referred to as the *fusion engine* – is the key technical challenge for multimodal interaction systems. In general, the meanings of input streams can vary according to context, task, user, and time. Modalities with very different characteristics – e.g., speech and eye gaze, facial expression and haptics input, touch-based gesture and prosody-based affect – may not have obvious points of similarity and straightforward ways to connect. Perhaps the most challenging aspect is the temporal dimension. Different modalities may have different temporal constraints and different signal and semantic endurance. Some modalities provide information at sparse, discrete points in time (e.g., some gestures) while others generate continuous but less time-specific output (e.g., affect). Some modal combinations are intended to be interpreted in parallel, which others may typically be offered sequentially.

Nigay and Coutaz (1993) classified multimodal interfaces in a 2×2 table depending on the fusion method (combined or independent) and the use of modalities (sequential or parallel) – see Table 2. In an *exclusive* multimodal system, the modalities are used sequentially and are available separately but not integrated by the system. In an *alternative* multimodal system, modalities are used sequentially but they are integrated to some degree (across time). In a *concurrent* multimodal system, modal information is available in parallel, but separately (not integrated). Finally, in a *synergistic* multimodal system, the modes are available in parallel and fully integrated. While synergistic multimodal systems are the assumed goal here, there are still possible benefits of the other styles of multimodal interfaces over unimodal systems.

Lalanne et al. (2009) provide a nice survey of multimodal integration methods as of 2009. They use the seven-layered protocol model of human–computer interaction of Nielsen (Nielsen, 1986), comprising the following levels:

- Goal – the current operational goal of an interaction
- Task/pragmatic – systems concepts to achieve the goal
- Semantic – specific operations the implement the desired tasks
- Syntactic – time and space sequencing of input and output information units on the underlying lexical level
- Lexical – the smallest information-carrying symbols (tokens) of the interaction
- Alphabetical – primitive symbols (letters, numbers, columns, lines, etc.)
- Physical – physically coded information (light, sound, movement, etc.)

The first three levels are conceptual, the next two perceptual, and the last two physical. Integration or fusion may be performed

Please cite this article in press as: Turk, M. Multimodal interaction: A review. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.003

Table 2
A classification of multimodal interfaces types, after Nigay and Coutaz (1993).

		Use of modalities	
		Sequential	Parallel
Fusion of modalities	Integrated	Alternative	Synergistic
	Not integrated	Exclusive	Concurrent

at any or all levels. This kind of analysis allows for a wide range of human–computer interaction scenarios to be formalized and modeled along with the possible contributions of various modalities, so that a system can predict what level(s) of integration are required in order to support a given interaction. Something like this will be needed in order for multimodal interaction to move beyond a large collection of specific approaches for certain sets of {users|modalities|environments|tasks} to more general solutions that can adapt to new situations.

Prior to ten years ago, there were only a few publications on multimodal integration as a specific topic (Lalanne et al., 2009). The interest in integration/fusion approaches has increased in the past decade, but much new research is needed in this critical area of multimodal interaction.

8. Multimodal integration – early or late?

As previously mentioned, the key issue in multimodal integration is how and when modalities should be integrated (see Johnston et al., 1997; Johnston, 1998; Wu et al., 1999; Nakamura, 2002; Chai et al., 2004; Johnston and Bangalore, 2005; Wasinger, 2006; Portillo et al., 2006; Mendonca et al., 2009; Song et al., 2012). That is, given streams of data from multiple modalities, such as voice, 3D gesture, and touch, should the data be processed separately and interpreted unimodally before being integrated with information from other modalities? This is a *late integration model*, or decision-level integration – merge the multimodal information only after unimodal processing and classification decisions have been made. Or should the data be integrated across modalities immediately (perhaps after initial low-level processing)? This is an *early integration model*, or feature-level integration of sensory data. A compromise, referred to as *mid-level integration*, allows for some degree of processing and perhaps classification before merging across modalities.

There are advantages to using late integration of multiple modalities in multimodal systems. For example, the input types can be recognized independently, and therefore do not have to occur simultaneously. The training requirements are smaller: generally O(2N) for two separately trained modes as opposed to $O(N^2)$ for two modes trained together. The software development process is also simpler in the late integration case, as exemplified by the QuickSet multimodal architecture (Cohen et al., 1997). QuickSet used temporal and semantic filtering, unification as the fundamental integration technique, and a statistical ranking to decide among multiple consistent interpretations. This approach required preprocessed and classified outputs from each modality.

On the other hand, late integration can miss key cross-modal interactions and force unimodal decisions apart from the full multimodal context. In considering the question of how sensory information should be integrated, Coen (2001) makes a compelling argument for the importance of early integration in biological systems (see also Wahlster, 2003). While our senses seem to be distinct – e.g., seeing is a qualitatively different experience from hearing or tasting – he points out the pervasiveness of cross-modal influence in perception. An extreme example is in synesthesia (or ideasthesia), which is a neurological condition in which stimulation of one sensory or cognitive pathway leads to automatic and involuntary experiences in a second sensory or cognitive pathway (Cytowic, 2002). For example, a person may hear sounds in response to viewing visual motion or flicker, or feel like an object is being held when experiencing a particular taste. How the brain coordinates and combines information from the different sensory modalities is known as the binding problem, and the traditional assumption has been that only at the highest levels of brain functioning in the cortex are sensory streams integrated, and they interrelate only through experience. Integration is thus a post-perceptual process, integrating sensory input after the fact – i.e., after the input has been classified. It is a "late integration" model, combining temporally proximal, abstracted unimodal inputs into an integrated event model.

This post-perceptual approach to integration, according to Coen, denies the possibility of cross-modal influence, which is pervasive in biological perception (see Section 1), and he argues that the default approach to building multimodal interfaces – using late integration – suffers from the same problem. Coen argues that late integration is an artifact of how people like to build computational systems, but that it is not well-suited for dealing with the cross-modal interdependencies of perceptual understanding:

> *Perception does not seem to be amenable to the clear-cut abstraction barriers that computer scientists find so valuable for solving other problems, and we claim this approach has [led] to the fragility of so many multimodal systems* (Coen, 2001).

He goes on to question the appropriateness of a discrete, symbolic event model, claiming that it may be more useful to view perception as a fluid, dynamic process.

This discussion by Coen, more than a decade later, is still quite apropos to the multimodal interaction community. Not only for the "early vs. late integration" discussion, but also the bigger picture question of what should be the output of multimodal integration. Is the task of multimodal integration to produce a *multimodal event*, which fits nicely in the computing model of discrete events, event loops, and event handlers – or is it to produce a more complex representation of perceptual activity that may better match the human interaction which the system is intended to support? Methods that look for deep, complex relationships between modalities, possibly at various levels, appear to be promising approaches for the field.

9. Challenges in multimodal HCI

Despite the significant progress on multimodal interaction systems in recent years, much work remains to be done before sophisticated multimodal interaction becomes a commonplace, indispensable part of computing. Many challenges remain, and the research agenda moving forward must include both the continued development of individual modalities and methods for multimodal integration. Each unimodal technology (vision-based tracking and recognition, speech and sound recognition, language understanding, dialogue management, haptics, touch-based gesture, user modeling, context modeling, etc.) is an active research area in itself. Fundamental improvements based on machine learning techniques are necessary for improved performance, personalization, and adaptability. Multimodal integration methods and architectures need to explore a wider range of methods and modality combinations; most current systems integrate only two modalities, such as speech along with touch or visual gesture. Large, ambitious research projects and prototype systems must be developed in order to tackle some of the deep problems that may not be apparent with simpler systems.

It is important to understand issues relating to cognitive load in multimodal systems, both in terms of what multimodal systems

Please cite this article in press as: Turk, M. Multimodal interaction: A review. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.003

can indicate about a user's cognitive load (Chen et al. 2012), when people naturally interact multimodally (Oviatt et al. 2004), and how alternative modalities may reduce or increase the cognitive load. Despite some insight into these questions to date, a more thorough understanding of the issues is required. From an interface designer's viewpoint, developing and evaluating multimodal interaction systems is a significant challenge in practice. The importance of this issue has been understood since the early days of multimodal interaction (Coutaz et al., 1993). Chang and Bourguet (2008) provide a more recent framework for design and evaluation, but better guidance and accumulated best practices are still needed.

For computer vision researchers interested in applying real-time vision algorithms to multimodal human–computer interaction, the main areas of application are well documented, including face detection and recognition, facial expression analysis, hand tracking and modeling, head and body tracking and pose extraction, gesture recognition, activity analysis, and object recognition. Building systems to perform these tasks robustly, with limited computing resources, in real-world scenarios – in the presence of occlusion by objects and other people, changes in illumination and camera pose, variations in the appearance of users, and multiple users – is a huge challenge for the field. A high level of robustness is paramount for practical deployment of these recognition technologies, and in the end robustness can only be determined by thorough testing under a wide range of conditions. To accomplish required tasks at acceptable levels of overall system performance, researchers must determine what the accuracy and robustness requirements are for each component. Testing a face recognition system may be straightforward, but what are the implications for testing when there are several recognition technologies and underlying user and context models all in one system? The whole is clearly not just the sum of the parts, and full-system testing is critical.

In addition to the many issues of sensing, recognition, usability, and interaction, there are potentially quite significant privacy issues associated with multimodal systems that must be considered early on in order to provide potential users with the assurance and confidence that such systems will not violate expectations of security and privacy. Waiting until the technologies are on the way to market is not the way to handle these serious issues.

There is much work to be done before multimodal interfaces revolutionize the human–computer interface. The grand challenge of creating powerful, efficient, natural, and compelling multimodal interfaces is an exciting pursuit, one that will keep us busy for some time.

Acknowledgement

This work was partially supported by the National Science Foundation under Grant No. 1219261.

References

Andersen, R.A., 1997. Multimodal integration for the representation of space in the posterior parietal cortex. Philos. Trans. R. Soc. Lond. B Biol. Sci. 352 (1360), 1421–1428.

Arabzadeh, E., Clifford, C.W., Harris, J.A., 2008. Vision merges with touch in a purely tactile discrimination. Psychol. Sci. 19 (7), 635–641.

Arnal, L.H., Morillon, B., Kell, C.A., Giraud, A.L., 2009. Dual neural routing of visual facilitation in speech processing. J. Neurosci. 29 (43), 13445–13453.

Blattner, M.M., Glinert, E.P., 1996. Multimodal integration. IEEE Multimedia 3 (4), 14–24.

Bohus, D., Horvitz, E., 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In: ACM International Conference on Multimodal Interfaces (ICMI-MLMI), Beijing, China.

Bolt, R.A., 1980. "Put-that-there": voice and gesture at the graphics interface. ACM Comput. Graphic. 14 (3), 262–270.

Bunt, H., Beun, R.-J., Borghuis, T., 1998. Multimodal human–computer communication systems, techniques, and experiments. Lect. Notes Comput. Sci. 1374.

Calvert, G., 2001. Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex 11 (12), 1110–1123.

Calvert, G.A., Brammer, M.J., Iversen, S.D., 1998. Crossmodal identification. Trends Cogn. Sci. 2 (7), 247–253.

Campi, K.L., Bales, K.L., Grunewald, R., Krubitzer, L., 2010. Connections of auditory and visual cortex in the prairie vole (Microtus ochrogaster): evidence for multisensory processing in primary sensory areas. Cereb. Cortex 20 (1), 89–108.

Cappe, C., Thelen, A., Romei, V., Thut, G., Murray, M.M., 2012. Looming signals reveal synergistic principles of multisensory integration. J. Neurosci. 32 (4), 1171–1182.

Chai, J.Y., Hong, P., Zhou, M.X., 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In: ACM International Conference on Intelligent User, Interfaces, pp. 70–77.

Chang, J., Bourguet, M.-L., 2008. Usability framework for the design an devaluation of multimodal interaction. In: The 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, vol. 2, pp. 123–126.

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, A., Taib, R., Yin, B., Wang, Y., 2012. Multimodal behaviour and interaction as indicators of cognitive load. ACM Trans. Interact. Intell. Syst. 2 (4).

Chen, T., 2001. Audiovisual speech processing. IEEE Signal Process. Mag., 9–21.

Coen, M.H., 2001. Multimodal integration – a biological view. Int. Joint Conf. Artif. Intell. 17 (1), 1417–1424.

Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J., 1997. QuickSet: multimodal interaction for distributed applications. In: ACM International Conference on Multimedia, Seattle, WA, pp. 31–40.

Coutaz, J., Salber, D., Balbo, S., 1993. Towards automatic evaluation of multimodal user interfaces. Knowl.-Based Syst. 6 (4), 267–274.

Cutugno, F., Leano, V. A., Rinaldi, R., Mignini, G., 2012. Multimodal framework for mobile interaction. In: International Working Conference on Advanced Visual Interfaces, Naples, Italy, pp. 197–203.

Cytowic, R.E., 2002. Synesthesia: A Union of the Senses, 2nd ed. MIT Press, Cambridge, Massachusetts.

Dumas, B., Lalanne, D., Oviatt, S., 2009. Multimodal interfaces: a survey of principles, models and frameworks. Human Machine Interaction. Lect. Notes Comput. Sci. 5440, 3–26, Springer.

Falchier, A., Clavagnier, S., Barone, P., Kennedy, H., 2002. Anatomical evidence of multimodal integration in primate striate cortex. J. Neurosci. 22 (13), 5749–5759.

Gauglitz, S., Lee, C., Turk, M., Höllerer, T., 2012. Integrating the physical environment into mobile remote collaboration. In: ACM International Conference on Human–Computer Interaction with Mobile Devices and Services (MobileHCI), San Francisco, CA.

Jaimes, A., Sebe, N., 2007. Multimodal human–computer interaction: a survey. Comput. Vis. Image Underst. 108 (1-2), 116–134, Elsevier, Amsterdam.

Johnston, M., 1998. Unification-based multimodal parsing. In: The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 624–630.

Johnston, M., Bangalore, S., 2005. Finite-state multimodal integration and understanding. Nat. Lang. Eng. 11 (2), 159–188.

Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A., Smith, I., 1997. Unification-based multimodal integration. In: 35th Annual Meeting of the Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 281–288.

Koons, D., Sparrell, C., Thorisson, K., 1993. Integrating simultaneous input from speech, gaze, and hand gestures. In: Maybury, M. (Ed.), Intelligent Multimedia Interfaces. MIT Press, Cambridge, pp. 257–276.

Lalanne, D., Nigay, L., Robinson, P., Vanderdonckt, J., Ladry, J.F., 2009. Fusion engines for multimodal input: a survey. In: ACM International Conference on Multimodal Interfaces (ICMI-MLMI), Cambridge, MA, pp. 153–160.

Leeb, R., Lancelle, M., Kaiser, V., Fellner, D.W., Pfurtscheller, G., 2013. Thinking penguin: multi-modal brain–computer interface control of a VR game. IEEE Trans. Comput. Intell. AI Games.

Leitão, J., Thielscher, A., Werner, S., Pohmann, R., Noppeney, U., 2012. Effects of parietal TMS on visual and auditory processing at the primary cortical level - a concurrent TMS-fMRI study. Cereb. Cortex. Oxford University Press. Advance Access available April 5, 2012.

Lewkowicz, D.J., Kraebel, K., 2004. The value of multimodal redundancy in the development of intersensory perception. In: Calvert, G., Spence, C., Stein, B. (Eds.), Handbook of Multisensory Processing. MIT Press.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature, 746–748.

Mendonça, H., Lawson, J.Y.L., Vybornova, O., Macq, B., Vanderdonckt, J., 2009. A fusion framework for multimodal interactive applications. In: ACM International Conference on Multimodal Interfaces (ICMI-MLMI), Cambridge, MA, pp. 161–168.

Nakamura, S., 2002. Statistical multimodal integration for audio-visual speech processing. IEEE Trans. Neural Networks 13 (4), 854–866.

Neal, J.G., Thielman, C.Y., Dobes, Z., Haller, S.M., Shapiro, S.C., 1989. Natural language with integrated deictic and graphic gestures. In: ACL Workshop on Speech and Natural Language. Association for Computational Linguistics, Stroudsburg, PA.

Nielsen, J., 1986. A virtual protocol model for computer–human interaction. Int. J. Man Mach. Stud. 24 (3), 301–312.

Nigay, L., Coutaz, J., 1993. A design space for multimodal systems: concurrent processing and data fusion. In: INTERCHI: Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, pp. 172–178.

Please cite this article in press as: Turk, M. Multimodal interaction: A review. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.003

Oviatt, S., 1999. Ten myths of multimodal interaction. Commun. ACM 42 (11), 74–81.
Oviatt, S., 2003. Advances in robust multimodal interface design. IEEE Comput. Graphics Appl. 23 (5), 62–68.
Oviatt, S., Cohen, P., 2000. Perceptual user interfaces: multimodal interfaces that process what comes naturally. Commun. ACM 43 (3), 45–53.
Oviatt, S.L., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., Ferro, D., 2000. Designing the user interface for multimodal speech and gesture applications: state-of-the-art systems and research directions. Human Comput. Interact. 15 (4), 263–322.
Oviatt, S., Coulston, R., and Lundsford, R., 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In: ACM International Conference on Multimodal Interfaces, State College, PA, pp. 129–136.
Oviatt, S., Lunsford, R., Coulston, R., 2005. Individual differences in multimodal integration patterns: what are they and why do they exist? In: ACM SIGCHI Conference on Human Factors in Computing Systems, vol. 2, No. 7, pp. 241–249.
Portillo, P.M., García, G.P., Carredano, G.A., 2006. Multimodal fusion: a new hybrid strategy for dialogue systems. In: ACM International Conference on Multimodal Interfaces, Banff, Canada, pp. 357–363.
Powers III, A.R., Hevey, M.A., Wallace, M.T., 2012. Neural correlates of multisensory perceptual learning. J. Neurosci. 32 (18), 6263–6274.
Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.F., Kirbas, C., McCullough, K.E., Ansari, R., 2002. Multimodal human discourse: gesture and speech. ACM Trans. Comput.–Human Interact. 9 (3), 171–193.
Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.-C., McTear, M., Raman, T.V., Stanney, K.M., Su, H., Wang, Q.Y., 2004. Guidelines for multimodal user interface design. Commun. ACM 47 (1), 57–59.
Ruiz, N., Chen, F., Oviatt, S., 2010. Multimodal input. In: Thiran, J.P., Marques, F., Bourlard, H. (Eds.), Multimodal Signal Processing: Theory and Applications for Human–Computer Interaction. Academic Press, pp. 231–255, Chapter 12.
Song, Y., Morency, L.P., Davis, R., 2012. Multimodal human behavior analysis: learning correlation and interaction across modalities. In: ACM International Conference on Multimodal Interfaces, Santa Monica, CA.
Stork, D.G., Hennecke, M.E., 1996. Speechreading by Humans and Machines. Springer.
Swallow, K.M., Makovski, T., Jiang, Y.V., 2012. Selection of events in time enhances activity throughout early visual cortex. J. Neurophysiol. 108, 3239–3252.
Turk, M., 1998. Moving from GUIs to PUIs. In: Symposium on Intelligent Information Media, Tokyo, Japan.
Turk, M., Robertson, G., 2000. Perceptual user interfaces. Commun. ACM 43 (3), 32–34.
Turk, M., Kölsch, M., 2004. Perceptual interfaces. In: Medioni, G., Kang, S.B. (Eds.), Emerging Topics in Computer Vision. Prentice Hall.
van Dam, A., 1997. Post-WIMP user interfaces. Commun. ACM 40 (2), 63–67.
van Wassenhove, V., Grant, K.W., Poeppel, D., 2005. Visual speech speeds up the neural processing of auditory speech. Proc. Nat. Acad. Sci. 102, 1181–1186.
Vasconcelos, N., Pantoja, J., Belchior, H., Caixeta, F.V., Faber, J., Freire, M.A.M., Cota, V.R., de Macedo, E.A., Laplagne, D.A., Gomes, H.M., Ribeiro, S., 2011. Cross-modal responses in the primary visual cortex encode complex objects and correlate with tactile discrimination. Proc. Nat. Acad. Sci. 108 (37), 15408–15413.
Wahlster, W., 2003. Towards symmetric multimodality: fusion and fission of speech, gesture, and facial expression. In: The 26th German Conference on Artificial Intelligence, Hamburg, Germany, pp. 1–18, September.
Waibel, A., Vo, M.T., Duchnowski, P., Manke, S., 1996. Multimodal interfaces. Artif. Intell. Rev. 10 (3), 299–319.
Wasinger, R., 2006. Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations. IOS Press.
Werner, S., Noppeney, U., 2011. The contributions of transient and sustained response codes to audiovisual integration. Cereb. Cortex 21 (4), 920–931.
Wu, L., Oviatt, S.L., Cohen, P.R., 1999. Multimodal integration – a statistical view. IEEE Trans. Multimedia 1 (4), 334–341.
Xiao, B., Girand, C., Oviatt, S.L., 2002. Multimodal integration patterns in children. In: International Conference on Spoken Language Processing, pp. 629–632.
Xiao, B., Lunsford, R., Coulston, R., Wesson, M., Oviatt, S., 2003. Modeling multimodal integration patterns and performance in seniors: toward adaptive processing of individual differences. In: ACM International Conference on Multimodal Interfaces, Vancouver, Canada, pp. 265–272.
Zhou, W., Zhang, X., Chen, J., Wang, L., Chen, D., 2012. Nostril-specific olfactory modulation of visual perception in binocular rivalry. J. Neurosci. 32 (48), 17225–17229.

Please cite this article in press as: Turk, M. Multimodal interaction: A review. Pattern Recognition Lett. (2013), http://dx.doi.org/10.1016/j.patrec.2013.07.003