Multimodal Human Computer Interaction: A Survey

Alejandro Jaimes[*] and Nicu Sebe[&]

[*]FXPAL, Japan, Fuji Xerox Co., Ltd.
alex.jaimes@fujixerox.co.jp
[&]University of Amsterdam, The Netherlands
nicu@science.uva.nl

Abstract. In this paper we review the major approaches to multimodal human computer interaction from a computer vision perspective. In particular, we focus on body, gesture, gaze, and affective interaction (facial expression recognition, and emotion in audio). We discuss user and task modeling, and multimodal fusion, highlighting challenges, open issues, and emerging applications for Multimodal Human Computer Interaction (MMHCI) research.

1 Introduction

Multimodal Human computer interaction (MMHCI) lies at the crossroads of several research areas including computer vision, psychology, artificial intelligence, and many others. As computers become integrated into everyday objects (ubiquitous and pervasive computing), effective natural human-computer interaction becomes critical: in many applications, users need to be able to interact naturally with computers the way face-to-face human-human interaction takes place. We communicate through speech and use body language (posture, gaze [48], hand motions) to express emotion, mood, attitude, and attention [41].

In human-human communication, interpreting the mix of audio-visual signals is essential in understanding communication. Researchers in many fields recognize this, and thanks to advances in the development of unimodal techniques (in speech and audio processing, computer vision, etc.), and in hardware technologies (inexpensive cameras and sensors), there has been a significant growth in MMHCI research. Unlike in traditional HCI applications (a single user facing a computer and interacting with it via a mouse or a keyboard), in new applications (e.g., intelligent homes [43], remote collaboration, arts, etc.), interactions are not always explicit commands, and often involve multiple users.

Although much progress has been achieved in MMHCI, most researchers still treat each modality (e.g., vision, speech) separately, and integrate the results at the application stage. One reason for this is that the roles of multiple modalities and their interplay remain to be quantified and scientifically understood. Additionally, many open issues remain in processing each modality individually.

In this paper we highlight the main vision problems that in our view should be solved for successful MMHCI applications, and give an overview of the research areas we consider essential for MMHCI. We group vision techniques according to the human body (Figure 1). Large-scale body movement, gesture (e.g., hands), and gaze analysis are used for tasks such as emotion recognition in affective interaction, and for a variety of applications. We discuss affective computer interaction, issues in multi-modal fusion, modeling, and data collection, and a variety of emerging MMHCI applications. Since MMHCI is a very dynamic and broad research area we do not intend to present a complete survey. The main contribution of this paper, therefore, is to consolidate some of the main issues and approaches, and to highlight some of the techniques and applications developed recently within the context of MMHCI.

1.1. Related Surveys

Extensive surveys have been previously published in several areas such as face detection [88][26], face recognition [91], facial expression analysis [17][54], vocal emotion [46][95], gesture recognition [38][78][57], human motion analysis [27][83][84][22][1][44], and eye tracking [12]. A review of vision-based HCI is presented in [62] with a focus on head tracking, face and facial expression recognition, eye tracking, and gesture recognition. Adaptive and intelligent HCI is discussed in [14] with a review of computer vision for human motion analysis, and a discussion of techniques for lower arm movement detection, face processing, and gaze analysis. Multimodal interfaces are discussed in [49][50][51][52][69]. Real-time vision for HCI (gestures, object tracking, hand posture, gaze) is discussed in [33]. Here, we discuss work not included in previous surveys, expand the discussion to areas not covered previously (e.g., in [33][14][62][50]), and discuss new applications in emerging areas while highlighting the main research issues.

2. Overview of Multimodal Interaction

The term multimodal has been used in many contexts and across several disciplines. For our interests, *a multimodal HCI system is simply one that responds to inputs in more than one modality or communication channel* (e.g., speech, gesture, writing, and others). We use a human-centered approach in our definition: by modality we mean mode of communication according to human senses *or* type of computer input devices. In terms of human senses the categories are *sight, touch, hearing, smell,* and *taste*. In terms of computer input devices we have modalities that are equivalent to human senses: cameras (*sight*), haptic sensors (*touch*), microphones (*hearing*), olfactory (*smell*), and even taste [36]. In addition, however, there are input devices that do not map directly to human senses: keyboard, mouse, writing tablet, motion input (e.g., the device itself is moved for interaction), and many others.

In our definition, a system that uses any combination of modalities in the categories above is multimodal. For our purposes, however, interest is exclusively on systems that include vision (cameras) as a modality[1]. A system that responds only to facial expressions and hand gestures, for example, is not multimodal, even if integration of both inputs (simultaneous or not) is used (using the same argument, a system with multiple keys is not multimodal, but a system with mouse a keyboard input is). The issue of where integration of modalities takes place, if at all, is of great importance and is discussed throughout the paper.

[1] Others have studied multimodal interaction using multiple devices such as mouse and keyboard, keyboard and pen, and so on.

As depicted in Figure 1, we place input modalities in two major groups: based on human senses (*vision*, *audio*, *haptic*, *olfactory* and *touch*), and others (mouse, keyboard, etc.). The visual modality includes any form of interaction that can be interpreted visually, and the audio modality any form that is audible (including multi-language input). We only discuss vision in detail, but as many new applications show (see Section 6), other modalities have gained importance for interaction (e.g., haptic [4]).

As depicted in Figure 1, multimodal techniques can be used to construct a variety of interfaces. Of particular interest for our goals are perceptual and attentive interfaces. Perceptual interfaces [80] as defined in [81], are highly interactive, multimodal interfaces that enable rich, natural, and efficient interaction with computers. Perceptual interfaces seek to leverage sensing (input) and rendering (output) technologies in order to provide interactions not feasible with standard interfaces and common I/O devices such as the keyboard, the mouse and the monitor [81]. Attentive interfaces, on the other hand, are context-aware interfaces that rely on a person's attention as the primary input [71] — the goal of these interfaces [47] is to use gathered information to estimate the best time and approach for communicating with the user.

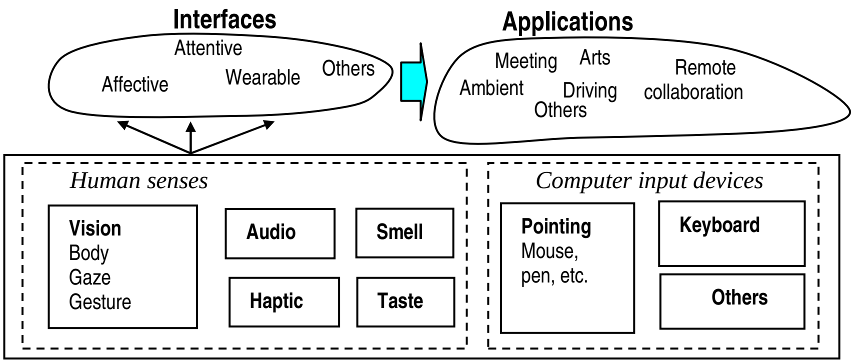

Figure 1. Overview of multimodal interaction using a human-centered approach.

Vision plays a fundamental role in several types of interfaces. As argued in [71], attention is epitomized by eye contact (even though other measures, such as cursor movement can also be indicative). Perceptual interfaces aim at natural interaction, making vision an essential component. The key point is that vision plays a major role in human-computer interfaces that aim at natural interaction. As we will see in Section 6, vision in multimodal interaction is applied in a variety of applications and interface types.

Although there have been many advances in MMHCI, as our discussions will show, the majority of research approaches focus on one mode independently and fuse the results at the highest level possible (in the application). Accordingly, in the next section we survey Computer Vision techniques for MMHCI and in the following sections we discuss fusion, interaction, and applications.

3. Core Vision Techniques

We classify vision techniques for MMHCI using a human-centered approach and divide them according to how humans may interact with the system: (1) large-scale body movements, (2) gestures, and (3) gaze. We make a distinction between *command* (actions can be used to explicitly execute commands: select menus, etc.) and *non-command* interfaces (actions or events used to indirectly tune the system to the user's needs) [45][7].

In general, vision-based human motion analysis systems used for MMHCI can be thought of as having mainly 4 stages: (1) motion segmentation, (2) object classification, (3) tracking, and (4) interpretation. While some approaches use geometric primitives to model different components (e.g., cylinders for limbs, head, and torso for body movements, or for hand and fingers in gesture recognition), others use feature representations based on appearance. In the first approach, external markers are often used to estimate body posture and relevant parameters. While markers can be accurate, they place restrictions on clothing and require calibration, so they are not desirable in many applications. Appearance based methods, on the other hand, do not require markers, but require training (e.g., with machine learning, probabilistic approaches, etc.). Methods that do not require markers place fewer constraints on the user and are more desirable, as are those that do not use geometric primitives (which are computationally expensive and often not suitable for real-time processing).

Next, we discuss some specific techniques for body, gesture, and gaze. The motion analysis steps are similar, so there is some inevitable overlap in the discussions. Some of the issues for gesture recognition, for instance, apply to body movements and gaze detection.

3.1. Large-Scale Body Movements

Tracking of large-scale body movements (head, arms, torso, and legs) is necessary to interpret pose and motion in many MMHCI applications Since extensive surveys have been published [83][84][22][1][44], we discuss the topic briefly.

The authors of [87] identify three important issues in articulated motion analysis: representation (joint angles or motion of all the sub-parts), computational paradigms (deterministic or probabilistic), and computation reduction. They propose a dynamic Markov network that uses Mean Field Monte Carlo algorithms so that a set of low dimensional particle filters interact with each other to solve a high dimensional problem collaboratively.

Body posture analysis is important in many MMHCI applications. In [77], the authors use a stereo and thermal infrared video system to estimate driver posture for deployment of smart air bags. The authors of [64] propose a method for recovering articulated body pose without initialization and tracking (using learning). The authors of [3] use pose and velocity vectors to recognize body parts and detect different activities, while the authors of [5] use temporal templates.

In some emerging MMHCI applications, group and non-command actions play an important role. The authors of [40] present an approach to segment a meeting according to actions such as monologue, presentation, white-board, discussion, and note taking. HMMs are used with a combination of audiovisual features. Visual features are extracted from head and hand/forearm blobs: the head blob is represented by the vertical position of its centroid, and hand blobs are represented by eccentricity and angle with respect to the horizontal. Audio features include energy, pitch, and speaking rate, among others. The authors of [24] use only computer vision, but make a distinction between body movements, events, and behaviors, within a rule-based system framework.

Important issues for large-scale body tracking include whether the approach uses 2D or 3D, desired accuracy, speed, occlusion and other constraints. Some of the issues pertaining to gesture recognition, discussed next, can also apply to body tracking.

3.2. Gesture Recognition

Psycholinguistic studies for human-to-human communication [41] describe gestures as the critical link between our conceptualizing capacities and our linguistic abilities. Humans use a very wide variety of gestures ranging from simple actions of using the hand to point at objects to the more complex actions that express feelings and allow communication with others. Gestures should therefore play an essential role in MMHCI [32][86][19]. A major motivation for these research efforts is the potential of using hand gestures in various applications aiming at natural interaction between the human and the computer-controlled interface. These applications range from virtual environments [31], to smart surveillance [78] and remote collaboration [19].

There are several important issues that should be considered when designing a gesture recognition system [57]. The first phase of a recognition task is choosing a mathematical model that may consider both the spatial and the temporal characteristics of the hand and hand gestures. The approach used for modeling plays a crucial role in the nature and performance of gesture interpretation. Once the model is detected, an analysis stage is required for computing the model parameters from the features that are extracted from single or multiple input streams. These parameters represent some description of the hand pose or trajectory and depend on the modeling approach used. Among the important problems involved in the analysis are that of hand localization [94], hand tracking [89], and the selection of suitable features [32]. After the parameters are computed, the gestures represented by them need to be classified and interpreted based on the accepted model and based on some grammar rules that reflect the internal syntax of gestural commands. The grammar may also encode the interaction of gestures with other communication modes such as speech, gaze, or facial expressions. As an alternative, some authors have explored using combinations of simple 2D motion based detectors for gesture recognition [29].

In any case, to fully exploit the potential of gestures for an MMHCI application, the class of possible recognized gestures should be as broad as possible and ideally any gesture preformed by the user should be unambiguously interpretable by the interface. However, most of the gesture-based HCI systems allow only symbolic commands based on hand posture or 3D pointing. This is due to the complexity associated with gesture analysis and the desire to build real-time interfaces. Also, most of the systems accommodate only single-hand gestures. Yet, human gestures, especially communicative, naturally employ actions of both hands. However, if the two-hand gestures are to be allowed, several ambiguous situations may appear (e.g., occlusion of hands, intentional vs. unintentional, etc.) and the processing time will likely increase. Another important aspect that is increasingly considered is the use of other modalities (e.g., speech) to augment the MMHCI system [51][72]. The use of such multimodal approaches can reduce the complexity and increase the naturalness of the interface for MMHCI [50].

3.3. Gaze Detection

Gaze, defined as the direction to which the eyes are pointing in space, is a strong indicator of attention, and it has been studied extensively since as early as 1879 in psychology, and more recently in neuroscience and in computing applications [12]. While early eye tracking research focused only on systems for in-lab experiments, many commercial and experimental systems are available today for a wide range of applications.

Eye tracking systems can be grouped into wearable or non-wearable, and infrared-based or appearance-based. In infrared-based systems, a light shining on the subject whose gaze is to be tracked creates a "red-eye effect:" the difference in reflection between the cornea and the pupil is used to determine the direction of sight. In appearance-based systems, computer vision techniques are used to find the eyes in the image and then determine their orientation. While wearable systems are the most accurate (approximate error rates under 1.4° vs. errors under 1.7° for non-wearable infrared), they are also the most intrusive. Infrared systems are more accurate than appearance-based, but there are concerns over the safety of prolonged exposure to infrared lights. In addition, most non-wearable systems require (often cumbersome) calibration for each individual.

Appearance-based systems use both eyes to predict gaze direction, so the resolution of the image of each eye is often small, which makes them less accurate. In [82], the authors propose using a single high-resolution image of one eye to improve accuracy. Infrared-based systems usually use only one camera. The authors of [66] have proposed using multiple cameras to improve accuracy.

One trend has been to improve non-wearable systems for use in MMHCI and other applications where the user is stationary (e.g., [74][66]). For example, the authors of [74] monitor driver visual attention using a single, non-wearable camera placed on a car's dashboard to track face features and for gaze detection.

There have also been advances in wearable eye trackers for novel applications. In [90], eye tracking data is combined with video from the user's perspective, head directions, and hand motions to learn words from natural interactions with users; the authors of [58] use a wearable eye tracker to understand hand-eye coordination in natural tasks, and the authors of [13] use a wearable eye tracker to detect eye contact and record video for blogging.

The main issues in developing gaze tracking systems are intrusiveness, speed, robustness, and accuracy. The type of hardware and algorithms necessary, however, depend highly on the level of analysis desired. Gaze analysis can be performed at three different levels [7]: (a) highly detailed low-level micro-events, (b) low-level intentional events, and (c) coarse-level goal-based events. Micro-events include micro-saccades, jitter, nystagmus, and brief fixations, which are studied for their physiological and psychological relevance by vision scientists and psychologists. Low-level intentional events are the smallest coherent units of movement that the user is aware of during visual activity, which include sustained fixations and revisits. Although most of the work on HCI has focused on coarse-level goal-based events (e.g., using gaze as a pointer [73]), it is easy to foresee the importance of analysis at lower levels, particularly to infer the user's cognitive state in affective interfaces (e.g., [25]). Within this context, an important issue often overlooked is how to interpret eye-tracking data (see [67] for discussion on eye tracking data clustering).

4. Affective Human-computer Interaction

There is a vast body of literature on affective computing and emotion recognition [2][55][61]. Affective states are intricately linked to other functions such as attention, perception, memory, decision-making, and learning [15]. This suggests that it may be beneficial for computers to recognize the user's emotions and other related cognitive states and expressions.

Researchers use mainly two different methods to analyze emotions. One approach is to classify emotions into discrete categories such as *joy, fear, love, surprise, sadness,* etc., using different modalities as inputs to emotion recognition models. The problem is that the stimuli may contain blended emotions and the choice of these categories may be too restrictive, or culturally dependent. Another way is to have multiple dimensions or scales to describe emotions. Two common scales are valence and arousal. Valence describes the pleasantness of the stimuli, with positive or pleasant (e.g, *happiness*) on one end, and negative or unpleasant (e.g., *disgust)* on the other. The other dimension is arousal or activation. For example, *sadness* has low arousal, whereas *surprise* has a high arousal level. The different emotional labels could be plotted at various positions on a two-dimensional plane spanned by these two axes to construct a 2D emotion model [35][23].

Facial expressions and vocal emotions are particularly important in this context, so we discuss them in more detail below.

4.1 Facial Expression Recognition

Most facial expression recognition research (see [54] and [17] for two comprehensive reviews) has been inspired by the work of Ekman [15] on coding facial expressions based on the basic movements of facial features called action units (AUs). In this scheme, expressions are classified into a predetermined set of categories. Some methods follow a "feature-based" approach, where one tries to detect and track specific features such as the corners of the mouth, eyebrows, etc. Other methods use a "region-based" approach in which facial motions are measured in certain regions on the face such as the eye/eyebrow and the mouth. In addition, we can distinguish two types of classification schemes: dynamic and static. Static classifiers (e.g., Bayesian Networks) classify each frame in a video to one of the facial expression categories based on the results of a particular video frame. Dynamic classifiers (e.g., HMM) use several video frames and perform classification by analyzing the temporal patterns of the regions analyzed or features extracted. They are very sensitive to appearance changes in the facial expressions of different individuals so they are more suited for person-dependent experiments [10]. Static classifiers, on the other hand, are easier to train and in general need less training data but when used on a continuous video sequence they can be unreliable especially for frames that are not at the peak of an expression.

4.2 Emotion in Audio

The vocal aspect of a communicative message carries various kinds of information. If we disregard the manner in which a message is spoken and consider only the textual content, we are likely to miss the important aspects of the utterance and we might even completely misunderstand the meaning of the message. Nevertheless, in contrast to spoken language processing, which has recently witnessed significant advances, the processing of emotional speech has not been widely explored.

Starting in the 1930s, quantitative studies of vocal emotions have had a longer history than quantitative studies of facial expressions. Traditional as well as most recent studies on emotional contents in speech (see [46], [95], and [68]) use "prosodic" information which includes the pitch, duration, and intensity of the utterance. Recent studies seem to use the "Ekman six" basic emotions, although others in the past have used many more categories. The reasons for using these basic categories are often not justified since it is not clear whether there exist "universal" emotional characteristics in the voice for these six categories [11].

The most surprising issue regarding the multimodal affect recognition problem is that although recent advances in video and audio processing could make the multimodal analysis of human affective state tractable, there are only a few research efforts [30][70][92] that have tried to implement a multimodal affective analyzer.

5. Modeling, Fusion, and Data Collection

5.1 User, context, and task modeling

Multimodal interface design [63] is important because the principles and techniques used in traditional GUI-based interaction do not necessarily apply in MMHCI systems. Issues to consider, as identified in [63] include design of inputs and outputs, adaptability, consistency, and error handling, among others. In addition, one must consider dependency of a person's behavior on his/her personality, cultural, and social vicinity, current mood, and the context in which the observed behavioral cues are encountered.

Many design decisions dictate the underlying techniques used in the interface. For example, adaptability can be addressed using machine learning: rather than using a priori rules to interpret human behavior, we can potentially learn application-, user-, and context-dependent rules by watching the user's behavior in the sensed context [59]. Probabilistic graphical models have an important advantage here: well known algorithms exist to adapt the models, and it is possible to use prior knowledge when learning new models. For example, a prior model of emotional expression recognition trained based on a certain user can be used as a starting point for learning a model for another user, or for the same user in a different context. Although context sensing and the time needed to learn appropriate rules are significant problems in their own right, many benefits could come from such adaptive MMHCI systems.

5.2 Fusion

A typical issue of multimodal data processing is that multisensory data is typically processed separately and only combined at the end. Yet, people convey multimodal (e.g., audio and visual) communicative signals in a complementary and redundant manner (as shown experimentally by Chen [11]). Therefore, in order to accomplish a human-like multimodal analysis of multiple input signals acquired by different sensors, the signals cannot be considered mutually independently and cannot be combined in a context-free manner at the end of the intended analysis but, on the contrary, the input data should be processed in a joint feature space and according to a context-dependent model. In practice, however, besides the problems of context sensing and developing context-dependent models for combining multisensory information, one should cope with the size of the required joint feature space. Problems include large dimensionality, differing feature formats, and time-alignment. A potential way to achieve multisensory data fusion is to develop context-dependent versions of a suitable method such as the Bayesian inference method proposed by Pan et al. [53].

In spite of its importance, the problem of fusing multiple modalities is often largely ignored. For example, the studies in facial expression recognition and vocal affect recognition have been done largely independent of each other. Most works in facial expression recognition use still photographs or video sequences without speech. Similarly, works on vocal emotion detection often use only audio information. A legitimate question that should be considered in MMHCI, is how much information does the face, as compared to speech, and body movement, contribute to natural interaction. Most experimenters suggest that the face is more accurately judged, produces higher agreement, or correlates better with judgments based on full audiovisual input than on voice input [42].

A multimodal system should be able to deal with imperfect data and generate its conclusion so that the certainty associated with it varies in accordance to the input data. A way of achieving this is to consider the time-instance versus time-scale dimension of human nonverbal communicative signals [55]. By considering previously observed data (time scale) with respect to the current data carried by functioning observation channels (time instance), a statistical prediction and its probability might be derived about both the information that has been lost due to malfunctioning/inaccuracy of a particular sensor and the currently displayed action/reaction. Probabilistic graphical models, such as Hidden Markov Models (including their hierarchical variants), Bayesian networks, and Dynamic Bayesian networks are very well suited for fusing such different sources of information. These models can handle noisy features, temporal information, and missing values of features all by probabilistic inference. Hierarchical HMM-based systems [10] have been shown to work well for facial expression recognition. Dynamic Bayesian Networks and HMM variants [21] have been shown to fuse various sources of information in recognizing user intent, office activity recognition, and event detection in video using both audio and visual information [20]. This suggests that probabilistic graphical models are a promising approach to fusing realistic (noisy) audio and video for context-dependent detection of behavioral events such as affective states.

Despite important advances, further research is still required to investigate fusion models able to efficiently use the complementary cues provided by multiple modalities.

5.3. Data Collection and Testing

Collecting MMHCI data and obtaining the ground truth for it is a challenging task. Labeling is time-consuming, error prone, and expensive. In developing multimodal techniques for emotion recognition, for example, one approach consists of asking actors to read material aloud while simultaneously portraying particular emotions chosen by the investigators. Another approach is to use emotional speech from real conversations or to induce emotions from speakers using various methods (e.g., showing photos or videos to induce reactions). Using actor portrayals ensures control of the verbal material and the encoder's intention, but raises the question about the similarity between posed and naturally occurring expressions. Using real emotional speech, on the other hand, ensures high validity, but renders the control of verbal material and encoder intention more difficult. Induction methods are effective in inducing moods, but it is harder to induce intense emotional states in controlled laboratory settings.

In general, collection of data for an MMHCI application is challenging because there is wide variability in the set of possible inputs (consider the number of possible gestures), often only a small set of training examples is available, and the data is often noisy. Therefore, it is very beneficial to construct methods that use scarcely available labeled data and abundant unlabeled data.

Probabilistic graphical models are ideal candidates for tasks in which labeled data is scarce, but abundant unlabeled data is available. Efficient and convergent probabilistic graphical model algorithms exist for handling missing and unlabeled data. Cohen et al. [9] showed that unlabeled data can be used together with labeled data for MMHCI applications using Bayesian networks. However, they have shown that care must be taken when attempting such schemes. In the purely supervised case (only labeled data), adding more labeled data always improves the performance of the classifier. Adding unlabeled data, however, can be detrimental to the performance. Such detrimental effects occur when the assumed classifier's model does not match the data's distribution.

To conclude, further research is necessary to achieve maximum utilization of unlabeled data for MMHCI problems since it is clear that such methods could provide great benefit.

6. Applications

Throughout the paper we have discussed techniques applied in a wide variety of application scenarios, including video conferencing and remote collaboration, intelligent homes, and driver monitoring.

As many of these applications show, the model of user interface in which one person sits in front of a computer is quickly changing. In some cases, the actions or events to be recognized are not explicit commands. In smart conference room applications, multimodal analysis has been applied mostly for video indexing [40] (see [60] for a social analysis application). Although such approaches are not meant to be used in real-time, they are useful in investigating how multiple modalities can be fused in interpreting communication. It is easy to foresee applications in which "smart meeting rooms" actually react to multimodal actions in the same way that intelligent homes should [43].

Perhaps one of the most exciting application areas of MMHCI is art. Vision techniques can be used to allow audience participation [39] and influence a performance. In [85], the authors use multiple modalities (video, audio, pressure sensors) to output different "emotional states" for Ada, an intelligent space that responds to multimodal input from its visitors. In [37], a wearable camera pointing at the wearer's mouth interprets mouth gestures to generate MIDI sounds (so a musician can play other instruments while generating sounds by moving his mouth). In [56], limb movements are tracked to generate music. MMHCI can also be used in museums to augment exhibitions [76].

Robotics is yet another interesting area for MMHCI. The authors of [18] give a comprehensive review of socially active robots and discuss the role of "human-oriented perception" (speech, gesture, and gaze).

People with disabilities can benefit greatly from MMHCI technologies [34]. The authors of [75] propose a component-based smart wheel chair system and discuss other approaches that integrate various types of sensors (not only vision). In [12], computer vision is used to interpret facial gestures for wheel chair navigation. In [6], the authors present two techniques (head tilt and gesture with audio feedback) to control a mobile device. The approach could be beneficial for people with disabilities, but it points to another interesting area: use of MMHCI for mobile devices that have limited input/output resources. Finally, [65] introduces a system for presenting digital pictures non-visually (multimodal output). Other important application areas include gaming [92], and education, "safety-critical applications" (e.g., medicine, military, etc. [8]) among others.

7. Conclusion

We have highlighted major vision approaches for multimodal human-computer interaction. We discussed techniques for large-scale body movement, gesture recognition, and gaze detection. We discussed facial expression recognition, emotion analysis from audio, user and task modeling, multimodal fusion, and a variety of emerging applications.

One of the major conclusions of this survey is that most researchers process each channel (visual, audio) independently, and multimodal fusion is still in its infancy. On one hand, the whole question of how much information is conveyed by "separate" channels may inevitably be misleading. There is no evidence that individuals in actual social interaction selectively attend to another person's face, body, gesture, or speech, or that the information conveyed by these channels is simply additive. The central mechanisms directing behavior cut across channels, so that, for example, certain aspects of face, body, and speech are more spontaneous and others are more closely monitored and controlled. It might well be that observers selectively attend not to a particular channel but to a particular type of information (e.g., cues to emotion, deception, or cognitive activity), which may be available within several channels. No investigator has yet explored this possibility or the possibility that different individuals may typically attend to different types of information.

Another important issue is the affective aspect of communication that should be considered when designing an MMHCI system. Emotion modulates almost all modes of human communication—facial expression, gestures, posture, tone of voice, choice of words, respiration, skin temperature and clamminess, etc. Emotions can significantly change the message: often it is not what was said that is most important, but how it was said. As noted by Picard [61] affect recognition is most likely to be accurate when it combines multiple modalities, information about the user's context, situation, goal, and preferences. A combination of low-level features, high-level reasoning, and natural language processing is likely to provide the best emotion inference in the context of MMHCI. Considering all these aspects, Pentland [59] believes that multimodal context-sensitive human-computer interaction is likely to become the single most widespread research topic of the artificial intelligence research community. Advances in this area could change not only how professionals practice computing, but also how mass consumers interact with technology.

References

[1] J.K. Aggarwal and Q. Cai, "Human motion analysis: A review," *CVIU*, 73(3):428-440, 1999.

[2] Application of Affective Computing in Human-computer Interaction, *Int. J. of Human-Computer Studies*, 59(1-2), 2003.

[3] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. On PAMI*, 24(8):1091-1104, 2002.

[4] M. Benali-Khoudja, M. Hafez, J.-M. Alexandre, and A. Kheddar, "Tactile interfaces: a state-of-the-art survey*," Int. Symposium on Robotics*, 2004.

[5] A.F. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on PAMI*, 23(3):257–267, 2001.

[6] S.A. Brewster, J. Lumsden, M. Bell, M. Hall, S. and Tasker," Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices," in proc. *ACM CHI 2003*.

[7] C. S. Campbell and P.P. Maglio, "A Robust Algorithm for Reading Detection," *ACM Workshop on Perceptive User Interfaces*, 2001.

[8] P.R. Cohen, D.R. McGee, "Tangible Multimodal Interfaces for Safety-critical Applications," *Communications of the ACM*, Vol. 47, Issue 1, pp. 41-46, January 2004.

[9] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T.S. Huang, "Semi-supervised learning of classifiers: Theory, algorithms, and their applications to human-computer interaction," *IEEE Trans. on PAMI*, 22(12):1553-1567, 2004.

[10] I. Cohen, N. Sebe, A. Garg, L. Chen, and T.S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *CVIU*, 91(1-2):160–187, 2003.

[11] L.S. Chen, *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction*, PhD thesis, UIUC, 2000.

[12] A.T. Duchowski, "A Breadth-First Survey of Eye Tracking Applications," *Behavior Research Methods, Instruments, and Computing*, 34(4):455-70, 2002.

[13] C. Dickie, R. Vertegaal, D. Fono, C. Sohn, D. Chen, D. Cheng, J.S. Shell and O. Aoudeh, "Augmenting and Sharing Memory with eyeBlog," in *CARPE* 2004.

[14] Z. Duric, W. Gray, R. Heishman, F. Li, A. Rosenfeld, M. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proc. of the IEEE*, 90(7):1272-1289, 2002.

[15] P. Ekman, ed., *Emotion in the Human Face*, Cambridge University Press, 1982.

[16] C. Fagiani, M. Betke, and J. Gips, "Evaluation of tracking methods for human-computer interaction," *IEEE Workshop on Applications in Computer Vision*, 2002.

[17] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey" *Patt. Recogn.*, 36:259–275, 2003.

[18] T. Fong, I. Nourbakhsh, K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, 42(3-4):143-166, 2003.

[19] S. Fussell, L. Setlock, J. Yang, J. Ou, E. Mauer, A. Kramer, "Gestures over video streams to support remote collaboration on physical tasks," *Human-Computer Int.*, 19(3):273-309, 2004.

[20] A. Garg, M. Naphade, and T.S. Huang, "Modeling video using input/output Markov models with application to multi-modal event detection," *Handbook of Video Databases: Design and Applications*, 2003.

[21] A. Garg, V. Pavlovic, and J. Rehg. "Boosted learning in dynamic Bayesian networks for multimodal speaker detection," *Proceedings of the IEEE*, 91(9):1355–1369, 2003.
[22] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *CVIU*, 73(1):82-98, 1999.
[23] A. Hanjalic and L-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. on Multimedia*, 7(1):143– 154, 2005.
[24] A. Hakeem and M. Shah, "Ontology and taxonomy collaborated framework for meeting classification," *ICPR*, 2004.
[25] R. Heishman, Z. Duric, and H. Wechsler, "Using eye region biometrics to reveal affective and cognitive states," *CVPR Workshop on Face Processing in Video*, 2004.
[26] E. Hjelmas and B. K. Low, "Face detection: A survey," *CVIU*, 83:236–274, 2001.
[27] W. Hu, T. Tan, L. Wang, and S. Maybank, "A Survey on Visual Surveillance of Object Motion and Behaviors", *in IEEE Trans. On Systems, Man, and Cybernetics*, Vol. 34, No. 3, Aug. 2004.
[28] S. Intille, K. Larson, J. Beaudin, J. Nawyn, E. Tapia, P. Kaushik, "A living laboratory for the design and evaluation of ubiquitous computing technologies," *Conf. on Human Factors in Computing Systems*, 2004.
[29] A. Jaimes and J. Liu, "Hotspot Components for Gesture-Based Interaction," in proc. *IFIP Interact 2005*, Rome, Italy, Sept. 2005.
[30] R. El Kaliouby and P. Robinson: Real time inference of complex mental states from facial expressions and head gestures, *CVPR Workshop on Real-time Vision for HCI*, 2004.
[31] S. Kettebekov and R. Sharma, "Understanding gestures in multimodal human computer interaction," *Int. J. on Artificial Intelligence Tools*, 9(2):205-223, 2000.
[32] T. Kirishima, K. Sato, and K. Chihara, "Real-time gesture recognition by learning and selective control of visual interest points," *IEEE Trans. on PAMI*, 27(3):351–364, 2005.
[33] B. Kisacanin, V. Pavlovic, and T.S.Huang (eds.). *Real-Time Vision for Human-Computer Interaction*, Springer-Verlag, New York, 2005.
[34] Y. Kuno, N. Shimada, and Y. Shirai, "Look where you're going: A robotic wheelchair based on the integration of human and environmental observations," *IEEE Robotics and Automation*, 10(1):26-34, 2003.
[35] P. Lang, "The emotion probe: Studies of motivation and attention," *American Psychologist*, 50(5):372–385, 1995.
[36] A. Legin, A. Rudnitskaya, B. Seleznev, Yu. Vlasov, "Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie," *Analytica Chimica Acta*, Vol. 534, pp. 129-135, April 2005.
[37] M.J. Lyons, M. Haehnel, and N. Tetsutani, "Designing, playing, and performing, with a vision-based mouth Interface," *Conf. on New Interfaces for Musical Expression*, 2003.
[38] S. Marcel, "Gestures for multi-modal interfaces: A Review," *Technical Report IDIAP-RR 02-34*, 2002.
[39] D. Maynes-Aminzade, R. Pausch, and S. Seitz, "Techniques for interactive audience participation," *ICMI* 2002.
[40] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. on PAMI*, 27(3):305-317, 2005.
[41] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*, Univ. of Chicago Press, Chicago, IL, 1992.
[42] A. Mehrabian, "Communication without words," *Psychology Today*, 2(4):53–56, 1968.
[43] S. Meyer and A. Rakotonirainy, "A Survey of research on context-aware homes," *Australasian Information Security Workshop Conference on ACSW Frontiers*, 2003
[44] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *CVIU*, 81(3):231-258, 2001.
[45] J. Nielsen, "Non-command user interfaces," *Comm. of the ACM*, 36(4):83-99, 1993.
[46] P.Y. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms," *Int. J. of Human-Computer Studies*, 59(1-2):157–183, 2003.
[47] A. Oulasvirta and A. Salovaara, "A cognitive meta-analysis of design approaches to interruptions in intelligent environments," in *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI'04): Extended Abstracts (2004)*.
[48] P. Qvarfordt, and S. Zhai, "Conversing with the user based on eye-gaze patterns," *Conf. Human-Factors in Computing Syst.*, 2005.
[49] S. Oviatt, T. Darrell, and M. Flickner, "Multimodal Interfaces that Flex, Adapt, and Persist," (eds.) special issue, *Communications of the ACM*, Vol. 47, Issue 1, January 2004.
[50] S.L. Oviatt and P. Cohen, "Multimodal interfaces that process what comes naturally," *Comm. of the ACM*, 43(3):45-48, 2000.

[51] S.L. Oviatt, “Multimodal interfaces,” *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, chap.14, 286-304, 2003.
[52] S.L. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, “Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions,” *Human-Computer Int.*, 15:263-322, 2000.
[53] H. Pan, Z.P. Liang, T.J. Anastasio, and T.S. Huang. “Exploiting the dependencies in information fusion,” *CVPR*, vol. 2:407–412, 1999.
[54] M. Pantic and L.J.M. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” *IEEE Trans. on PAMI*, 22(12):1424–1445, 2000.
[55] M. Pantic and L.J.M. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction,” *Proceedings of the IEEE*, 91(9):1370–1390, 2003.
[56] J. Paradiso and F. Sparacino, “Optical Tracking for Music and Dance Performance,” *Optical 3-D Measurement Techniques IV*, A. Gruen, H. Kahmen, eds., pp. 11-18, 1997.
[57] V.I. Pavlovic, R. Sharma and T.S. Huang, “Visual interpretation of hand gestures for human-computer interaction: a review”, *IEEE Trans. on PAMI*, 19(7):677-695, 1997.
[58] J.B. Pelz, “Portable eye-tracking in natural behavior,” *J. of Vision*, 4(11), 2004.
[59] A. Pentland, “Looking at People,” *Comm. of the ACM,”* 43(3):35-44, 2000.
[60] A. Pentland, “Socially Aware Computation and Communication,” IEEE Computer Vol 38,No. 3, March 2005.
[61] R. W. Picard, *Affective Computing*, MIT Press, 1997.
[62] M. Porta, “Vision-based user interfaces: methods and applications,” *Int. J. Human-Computer Studies*, 57(1):27-73, 2002.
[63] L. M. Reeves et. al., “Guidelines for multimodal user interface design,” *Communications of the ACM*, Vol. 47, Issue 1, pp. 57-69, January 2004.
[64] R. Rosales and S. Sclaroff, “Learning body pose via specialized maps,” *NIPS*, Vol. 14, pp 1263-1270, 2001.
[65] P. Roth, and T. Pun, “Design and evaluation of a multimodal system for the non-visual exploration of digital pictures,” *INTERACT* 2003.
[66] R. Ruddaraju, A. Haro, K. Nagel, Q. Tran, I. Essa, G. Abowd, E. Mynatt, “Perceptual user interfaces using vision-based eye tracking,” *ICMI*, 2003.
[67] A. Santella and D. DeCarlo, “Robust clustering of eye movement recordings for quantification of visual interest,” *Eye Tracking Research and Applications (ETRA)*, pp. 27-34, 2004.
[68] N. Sebe, I. Cohen, and T.S. Huang, Multimodal emotion recognition, *Handbook of Pattern Recognition and Computer Vision*, World Scientific, 2005.
[69] E. Schapira and R. Sharma, “Experimental evaluation of vision and speech based multimodal interfaces,” W*orkshop on Perceptive User Interfaces*, pp. 1-9, 2001.
[70] B. Schuller, M. Lang, G. Rigoll: “Multimodal emotion recognition in audiovisual communication,” *ICME*, 2002.
[71] T. Selker, “Visual Attentive Interfaces,” *BT Technology Journal*, Vol. 22, No. 4, pp. 146-150, Oct. 2004.
[72] R. Sharma, M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, I. Brewer, A. MacEachren, and K. Sengupta, “Speech–gesture driven multimodal interfaces for crisis management,” *Proceedings of the IEEE*, 91(9):1327–1354, 2003.
[73] L.E. Sibert and R.J.K. Jacob, “Evaluation of eye gaze interaction,” *Conf. Human-Factors in Computing Syst.*, pp. 281-288, 2000.
[74] P. Smith, M. Shah, and N.d.V. Lobo, “Determining driver visual zttention with one camera,” *IEEE Trans. on Intelligent Transportation Systems*, 4(4), 2003.
[75] R. Simpson, E. LoPresti, S. Hayashi, I. Nourbakhsh, D. Miller, “The smart wheelchair component system,” *J. of Rehabilitation Research and Development*, May/June 2004.
[76] F. Sparacino, “The museum wearable: Real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences,” *Museums and the Web*, 2002.
[77] M. M. Trivedi, S. Y. Cheng, E. M. C. Childers, S. J. Krotosky, “Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation,” *IEEE Trans. on Vehicular Technology*, 53(6):1698-1712, 2004.
[78] M. Turk, "Gesture recognition," *Handbook of Virtual Environment Technology*, K. Stanney (ed.), 2001.
[79] M. Turk, "Computer vision in the interface," *Communications of the ACM*, Vol. 47, Issue 1, pp. 60-67, January 2004.

[80] M. Turk and G. Robertson, “Perceptual Interfaces,” *Communications of the ACM*, Vol. 43, Issue 3, pp. 32-34, 2000.
[81] M. Turk and M. Kölsch, "Perceptual Interfaces," G. Medioni and S.B. Kang (eds.), Emerging Topics in Computer Vision, Prentice Hall, 2004.
[82] J.-G. Wang, E. Sung, and R. Venkateswarlu, “Eye gaze estimation from a single image of one eye,” *ICCV*, pp. 136-143, 2003.
[83] L. Wang, W. Hu and T. Tan “Recent developments in human motion analysis,” *Patt. Recogn.*, 36 (2003) 585-601
[84] J.J.L. Wang and S. Singh, “Video analysis of human dynamics – A survey,” *Real-Time Imaging*, 9(5):321-346, 2003.
[85] K.C. Wassermann, K. Eng, P.F.M.J. Verschure, and J. Manzolli, “Live soundscape composition based on synthetic emotions,” *IEEE Multimedia Magazine*, 10(4), 2003.
[86] Y. Wu and T. Huang. Vision-based gesture recognition: A review, *3rd Gesture Workshop*, 1999.
[87] Y. Wu, G. Hua and T. Yu, “Tracking articulated body by dynamic Markov network,” *ICCV*, pp.1094-1101, 2003.
[88] M.-H. Yang, D. Kriegman, and N. Ahuja, “Detecting faces in images: A survey,” *IEEE Trans. on PAMI*, 24(1):34–58, 2002.
[89] Q. Yuan, S. Sclaroff and V. Athitsos, “Automatic 2D hand tracking in video sequences,” *IEEE Workshop on Applications of Computer Vision*, 2005.
[90] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensorimotor experience", *ACM Trans. on Applied Perception*, 2004.
[91] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips, “Face recognition: A literature survey,” *ACM Computing Surveys*, 12:399–458, 2003.
[92] K. Salen and E. Zimmerman, *Rules of Play: Game Design Fundamentals*, MIT Press, 2003.
[93] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, and S. Levinson, “Bimodal HCI-related affect recognition,” *ICMI*, 2004.
[94] Y. Wu and T.S. Huang, “Human hand modeling, analysis and animation in the context of human computer interaction,” *IEEE Signal Processing*, 18(3):51-60, 2001.
[95] I.R. Murray and J.L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion,” *J. of the Acoustic Society of America*, 93(2):1097–1108, 1993.