Multimodal Human Computer Interaction: A Survey

Alejandro Jaimes[*,1] and Nicu Sebe[&]

[*] IDIAP, Switzerland
ajaimes@ee.columbia.edu
[&] University of Amsterdam, The Netherlands
nicu@science.uva.nl

Abstract. In this paper we review the major approaches to Multimodal Human Computer Interaction, giving an overview of the field 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. We study MMHCI to determine how we can make computer technology more usable by people, which invariably requires the understanding of at least three things: *the user* who interacts with it, *the system* (the computer technology and its usability), and the *interaction* between the user and the system. By considering these aspects, it is obvious that MMHCI is a multi-disciplinary subject since the designer of an interactive system should have expertise in a range of topics: psychology and cognitive science to understand the user's perceptual, cognitive, and problem solving skills, sociology to understand the wider context of interaction, ergonomics to understand the user's physical capabilities, graphic design to produce effective interface presentation, computer science and engineering to be able to build the necessary technology, etc.

The multidisciplinary nature of MMHCI motivates our approach to this survey. Instead of focusing only on Computer Vision techniques for MMHCI, we give a general overview of the field, discussing the major approaches and issues in MMHCI from a computer vision perspective. Our contribution, therefore, is giving researchers in Computer Vision or any other area who are interested in MMHCI a broad view of the state of the art and outlining opportunities and challenges in this exciting area.

1.1. Motivation

In human-human communication, interpreting the mix of audio-visual signals is essential in communicating. Researchers in many fields recognize this, and thanks to

[1] This work was performed while Alejandro Jaimes was with FXPAL Japan, Fuji Xerox Co., Ltd.

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 the new applications (e.g., intelligent homes [105], remote collaboration, arts, etc.), interactions are not always explicit commands, and often involve multiple users. This is due in part to the remarkable progress in the last few years in computer processor speed, memory, and storage capabilities, matched by the availability of many new input and output devices that are making ubiquitous computing [185][67][66] a reality. Devices include phones, embedded systems, PDAs, laptops, wall size displays, and many others. The wide range of computing devices available, with differing computational power and input/output capabilities, means that the future of computing is likely to include novel ways of interaction. Some of the methods include gestures [136], speech [143], haptics [9], eye blinks [58], and many others. Glove mounted devices [19] and graspable user interfaces [48], for example, seem now ripe for exploration. Pointing devices with haptic feedback, eye tracking, and gaze detection [69] are also currently emerging. As in human-human communication, however, effective communication is likely to take place when different input devices are used in combination.

Multimodal interfaces have been shown to have many advantages [34]: they prevent errors, bring robustness to the interface, help the user to correct errors or recover from them more easily, bring more bandwidth to the communication, and add alternative communication methods to different situations and environments. Disambiguation of error-prone modalities using multimodal interfaces is one important motivation for the use of multiple modalities in many systems. As shown by Oviatt [123], error-prone technologies can compensate each other, rather than bring redundancy to the interface and reduce the need for error correction. It should be noted, however, that multiple modalities alone do not bring benefits to the interface: the use of multiple modalities may be ineffective or even disadvantageous. In this context, Oviatt [124] has presented the common misconceptions (myths) of multimodal interfaces, most of them related to the use of speech as an input modality.

In this paper, we review the research areas we consider essential for MMHCI, giving an overview of the state of the art, and based on the results of our survey, identify major trends and open issues in 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 provide an overview of the main computer vision techniques used in the context of MMHCI while giving an overview of the main research areas, techniques, applications, and open issues in MMHCI.

1.2. Related Surveys

Extensive surveys have been previously published in several areas such as face detection [190][63], face recognition [196], facial expression analysis [47][131], vocal emotion [119][109], gesture recognition [96][174][136], human motion analysis [65][182][182][56][3][46][107], audio-visual automatic speech recognition [143], and eye tracking [41][36]. Reviews of vision-based HCI are presented in [142] and [73] with a focus on head tracking, face and facial expression recognition, eye tracking, and gesture recognition. Adaptive and intelligent HCI is discussed in [40] 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 [125][126][127][128][144][158][135][171]. Real-time vision for HCI (gestures, object tracking, hand posture, gaze, face pose) is discussed in [84] and [77]. Here, we discuss work not included in previous surveys, expand the discussion to areas not covered previously (e.g., in [84][40][142][126][115]), and discuss new applications in emerging areas while highlighting the main research issues.

Related conferences and workshops include the following: ACM CHI, IFIP Interact, IEEE CVPR, IEEE ICCV, ACM Multimedia, International Workshop on Human-Centered Multimedia (HCM) in conjunction with ACM Multimedia, International Workshops on Human-computer Interaction in conjunction with ICCV and ECCV, Intelligent User Interfaces (IUI) conference, and International Conference on Multimodal Interfaces (ICMI), among others.

1.3. Outline

The rest of the paper is organized as follows. In section 2 we give an overview of MMHCI. Section 3 covers core computer vision techniques. Section 4 surveys affective HCI, and section 5 deals with modeling, fusion, and data collection, while section 6 discusses relevant application areas for MMHCI. We conclude with section 7.

2. Overview of Multimodal Interaction

The term multimodal has been used in many contexts and across several disciplines (see [10][11][12] for a taxonomy of modalities). 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 and by modality we mean mode of communication according to human senses and computer input devices activated by humans or measuring human qualities[2] (e.g., blood pressure, see Figure 1). The human senses are *sight, touch, hearing*, *smell,* and *taste*. The input modalities of many computer input devices can be considered to correspond to human senses: cameras (*sight*), haptic sensors (*touch*) [9], microphones (*hearing*), olfactory (*smell*), and even taste [92]. Many other computer

[2] Robots or other devices could communicate in a multimodal way with each other. For instance, a conveyor belt in a factory could carry boxes and a system could identify the boxes using RFID tags on the boxes. The orientation of the boxes could then be estimated using cameras. Our interest in this survey, however, is only on human-centered multimodal systems.

input devices activated by humans, however, can be considered to correspond to a combination of human senses, or to none at all: keyboard, mouse, writing tablet, motion input (e.g., the device itself is moved for interaction), galvanic skin response, and other biometric sensors.

In our definition, the word *input* is of great importance, as in practice most interactions with computers take place using multiple modalities. For example, as we type we touch keys on a keyboard to input data into the computer, but some of us also use sight to read what we type or to locate the proper keys to be pressed. Therefore, it is important to keep in mind the differences between what the human is doing and what the system is actually receiving as input during interaction. For instance, a computer with a microphone could potentially understand multiple languages or only different types of sounds (e.g., using a humming interface for music retrieval). Although the term multimodal has often been used to refer to such cases (e.g., multilingual input in [13] is considered multimodal), in this survey only a system that uses any combination of different modalities (i.e., communication channels) such as those depicted in Figure 1 is multimodal. For example, a system that responds only to facial expressions and hand gestures using only cameras as input is not multimodal, even if signals from various cameras are used. Using the same argument, a system with multiple keys is not multimodal, but a system with mouse and keyboard input is. Although others have studied multimodal interaction using multiple devices such as mouse and keyboard, keyboard and pen, and others, for the purposes of our survey, we are only interested in that the combination of visual (camera) input with other types of input for Human-Computer Interaction.

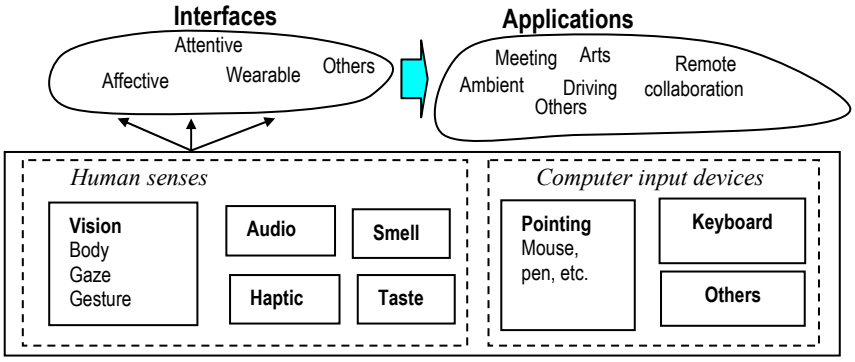

Figure 1. Overview of multimodal interaction using a human-centered approach.

In the context of HCI, multimodal techniques can be used to construct many different types of interfaces (Figure 1). Of particular interest for our goals are perceptual, attentive, and enactive interfaces. Perceptual interfaces [176], as defined in [177], 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 [177], making computer vision a central component in many cases. Attentive interfaces [180] are context-aware interfaces that rely on a person's attention as the pri-

mary input [160] — that is, attentive interfaces [120] use gathered information to estimate the best time and approach for communicating with the user. Since attention is epitomized by eye contact [160] and gestures (although other measures such as mouse movement can be indicative), computer vision plays a major role in attentive interfaces. Enactive interfaces are those that help users communicate a form of knowledge based on the active use of the hands or body for apprehension tasks. Enactive knowledge is not simply multisensory mediated knowledge, but knowledge stored in the form of motor responses and acquired by the act of "doing". Typical examples are the competence required by tasks such as typing, driving a car, dancing, playing a musical instrument, and modeling objects from clay. All of these tasks would be difficult to describe in an iconic or symbolic form.

In the next section, we survey Computer Vision techniques for MMHCI and in the following sections we discuss fusion, interaction, and applications in more detail.

3. Human-Centered Vision

We classify vision techniques for MMHCI using a human-centered approach and we divide them according to the human body: (1) large-scale body movements, (2) hand 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) [111][23].

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 used to model limbs, head, and torso for body movements, or for hand and fingers in gesture recognition), others use feature representations based on appearance (appearance-based methods). 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. Moreover, the attempt to fit geometric shapes to body parts can be computationally expensive and these methods are often not suitable for real-time processing. Appearance based methods, on the other hand, do not require markers, but require training (e.g., with machine learning, probabilistic approaches, etc.). Since they do not require markers, they place fewer constraints on the user and are therefore more desirable.

Next, we briefly 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. However, since extensive surveys have been published in this area [182][182][56][1][107], we discuss the topic briefly.

There are three important issues in articulated motion analysis [188]: representation (joint angles or motion of all the sub-parts), computational paradigms (deterministic or probabilistic), and computation reduction. Body posture analysis is important in many MMHCI applications. In [172], the authors use a stereo and thermal infrared video system to estimate the driver's posture for deployment of smart air bags. The authors of [148] propose a method for recovering articulated body pose without initialization and tracking (using learning). The authors of [8] use pose and velocity vectors to recognize body parts and detect different activities, while the authors of [17] use temporal templates.

In some emerging MMHCI applications, group and non-command actions play an important role. In [102], 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. These features together with audio features (e.g., energy, pitch, and speaking rate, among others) are used for segmenting meeting videos according to actions such as monologue, presentation, white-board, discussion, and note taking. The authors of [60] 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. Hand Gesture Recognition

Although in human-human communication gestures are often performed using a variety of body parts (e.g., arms, eyebrows, legs, entire body, etc.), most researchers in computer vision use the term gesture recognition to refer exclusively to hand gestures. We will use the term accordingly and focus on hand gesture recognition in this section.

Psycholinguistic studies of human-to-human communication [103] 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 [83][186][52], as they seem intrinsic to natural interaction between the human and the computer-controlled interface in many applications, ranging from virtual environments [82] and smart surveillance [174], to remote collaboration applications [52].

There are several important issues that should be considered when designing a gesture recognition system [136]. 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. Typically, features are extracted from the images or video, and once these features are extracted, model parameters are estimated based on subsets of them until a right match is found. For example, the system might detect **n** points and attempt to determine if these **n** points (or a subset of them) could match the characteristics of points extracted from a hand in a

particular pose or performing a particular action. The parameters of the model are then a description of the hand pose or trajectory and depend on the modeling approach used. Among the important problems involved in the analysis are hand localization [187], hand tracking [194], and the selection of suitable features [83]. 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 to modeling, some authors have explored the use of combinations of simple 2D motion based detectors for gesture recognition [71].

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 performed 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 gestures, 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 [127][162]. The use of such multimodal approaches can reduce the complexity and increase the naturalness of the interface for MMHCI [126].

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 [41]. 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 below 1.4° vs. errors below 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 [108].

Appearance-based systems usually capture both eyes using two cameras to predict gaze direction. Due to the computational cost of processing two streams simultaneously, the resolution of the image of each eye is often small. This makes such sys-

tems less accurate, although increasing computational power and lower costs mean that more computationally intensive algorithms can be run in real time. As an alternative, in [181], the authors propose using a single high-resolution image of one eye to improve accuracy. On the other hand, infrared-based systems usually use only one camera, but the use of two cameras has been proposed to further increase accuracy [152].

Although most research on non-wearable systems has focused on desktop users, the ubiquity of computing devices has allowed for application in other domains in which the user is stationary (e.g., [168][152]). For example, the authors of [168] monitor driver visual attention using a single, non-wearable camera placed on a car's dashboard to track face features and for gaze detection.

Wearable eye trackers have also been investigated mostly for desktop applications (or for users that do not walk wearing the device). Also, because of advances in hardware (e.g., reduction in size and weight) and lower costs, researchers have been able to investigate uses in novel applications (eye tracking while users walk). For example, in [193], eye tracking data are combined with video from the user's perspective, head directions, and hand motions to learn words from natural interactions with users; the authors of [137] use a wearable eye tracker to understand hand-eye coordination in natural tasks, and the authors of [38] 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 [23]: (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 [165]), 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., [62]). Within this context, an important issue often overlooked is how to interpret eye-tracking data. In other words, as the user moves his eyes during interaction, the system must decide what the movements mean in order to react accordingly. We move our eyes 2-3 times per second, so a system may have to process large amounts of data within a short time, a task that is not trivial even if processing does not occur in real-time. One way to interpret eye tracking data is to cluster fixation points and assume, for instance, that clusters correspond to areas of interest. Clustering of fixation points is only one option, however, and as the authors of [154] discuss, it can be difficult to determine the clustering algorithm parameters. Other options include obtaining statistics on measures such as number of eye movements, saccades, distances between fixations, order of fixations, and so on.

4. Affective Human-computer Interaction

Most current MMHCI systems do not account for the fact that human-human communication is always socially situated and that we use emotion to enhance our communication. However, since emotion is often expressed in a multimodal way, it is an important area for MMHCI and we will discuss it in some detail. HCI systems that can sense the affective states of the human (e.g., stress, inattention, anger, boredom, etc.) and are capable of adapting and responding to these affective states are likely to be perceived as more natural, efficacious, and trustworthy. In her book, Picard [140] suggested several applications where it is beneficial for computers to recognize human emotions. For example, knowing the user's emotions, the computer can become a more effective tutor. Synthetic speech with emotions in the voice would sound more pleasing than a monotonous voice. Computer agents could learn the user's preferences through the users' emotions. Another application is to help the human users monitor their stress level. In clinical settings, recognizing a person's inability to express certain facial expressions may help diagnose early psychological disorders.

The research area of machine analysis and employment of human emotion to build more natural and flexible HCI systems is known by the general name of affective computing [140]. There is a vast body of literature on affective computing and emotion recognition [67][132][140][133]. Emotion is intricately linked to other functions such as attention, perception, memory, decision-making, and learning [43]. This suggests that it may be beneficial for computers to recognize the user's emotions and other related cognitive states and expressions. Addressing the problem of affective communication, Bianchi-Berthouze and Lisetti [14] identified three key points to be considered when developing systems that capture affective information: *embodiment* (experiencing physical reality), *dynamics* (mapping the experience and the emotional state onto a temporal process and a particular label), and *adaptive interaction* (conveying emotive response, responding to a recognized emotional state).

Researchers use mainly two different methods to analyze emotions [133]. One approach is to classify emotions into discrete categories such as *joy, fear, love, surprise, sadness,* etc., using different modalities as inputs. 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 [88][60].

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 [131] and [47] for two comprehensive reviews) has been inspired by the work of Ekman [43] on coding facial expressions based on the basic movements of facial features called action units (AUs).

In order to offer a comprehensive description of the visible muscle movement in the face, Ekman proposed the Facial Action Coding System (FACS). In the system, a facial expression is a high level description of facial motions represented by regions or feature points called action units. Each AU has some related muscular basis and a given facial expression may be described by a combination of AUs. 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. Dynamic classifiers are very sensitive to appearance changes in the facial expressions of different individuals so they are more suited for person-dependent experiments [32]. 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.

Mase [99] was one of the first to use image processing techniques (optical flow) to recognize facial expressions. Lanitis et al. [90] used a flexible shape and appearance model for image coding, person identification, pose recovery, gender recognition, and facial expression recognition. Black and Yacoob [15] used local parameterized models of image motion to recover non-rigid motion. Once recovered, these parameters are fed to a rule-based classifier to recognize the six basic facial expressions. Yacoob and Davis [189] computed optical flow and used similar rules to classify the six facial expressions. Rosenblum et al. [149] also computed optical flow of regions on the face, then applied a radial basis function network to classify expressions. Essa and Pentland [45] also used an optical flow region-based method to recognize expressions. Otsuka and Ohya [117] first computed optical flow, then computed their 2D Fourier transform coefficients, which were then used as feature vectors for a hidden Markov model (HMM) to classify expressions. The trained system was able to recognize one of the six expressions near real-time (about 10 Hz). Furthermore, they used the tracked motions to control the facial expression of an animated Kabuki system [118]. A similar approach, using different features was used by Lien [93]. Nefian and Hayes [110] proposed an embedded HMM approach for face recognition that uses an efficient set of observation vectors based on the DCT coefficients. Martinez [98] introduced an indexing approach based on the identification of frontal face images under different illumination conditions, facial expressions, and occlusions. A Bayesian approach was used to find the best match between the local observations and the learned local features model and an HMM was employed to achieve good recognition even when the new conditions did not correspond to the conditions previously encountered during the learning phase. Oliver et al. [116] used lower face tracking to extract mouth shape features and used them as inputs to an HMM based facial expression recognition system (recognizing neutral, happy, sad, and an open mouth). Chen [28] used a suite of static classifiers to recognize facial expressions, reporting on both person-dependent and person-independent results.

In spite of the variety of approaches to facial affect analysis, the majority suffer from the following limitations [132]:

- handle a small set of posed prototypic facial expressions of six basic emotions from portraits or nearly-frontal views of faces with no facial hair or glasses recorded under constant illumination;
- do not perform a context-dependent interpretation of shown facial behavior;
- do not analyze extracted facial information on different time scales (short videos are handled only); consequently, inferences about the expressed mood and attitude (larger time scales) cannot be made by current facial affect analyzers.

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 [119], [109], [72], and [155]) use "prosodic" information, that is information on intonation, rhythm, lexical stress, and other features in speech. This is extracted using measures such as pitch, duration, and intensity of the utterance. Recent studies use "Ekman's 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 [27].

The limitations of existing vocal-affect analyzers are [132]:

- perform singular classification of input audio signals into a few emotion categories such as anger, irony, happiness, sadness/grief, fear, disgust, surprise and affection;
- do not perform a context-sensitive analysis (environment-, user- and task-dependent analysis) of the input audio signal;
- do not analyze extracted vocal expression information on different time scales (proposed inter-audio-frame analyses are used either for the detection of supra-segmental features, such as the pitch and intensity over the duration of a syllable or word, or for the detection of phonetic features) – inferences about moods and attitudes (longer time scales) are difficult to be made based on the current vocal-affect analyzers;
- adopt strong assumptions (e.g., the recordings are noise free, the recorded sentences are short, delimited by pauses, carefully pronounced by non-smoking actors) and use the test data sets that are small (one or more words or one or more short sentences spoken by few subjects) containing exaggerated vocal expressions of affective states.

4.3 Multimodal Approaches to Emotion Recognition

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 [80][159][153][195][157] that have tried to implement a multimodal affective analyzer.

Although studies in psychology on the accuracy of predictions from observations of expressive behavior suggest that the combined face and body approaches are the most informative [4][59], with the exception of a tentative attempt of Balomenos et al. [7], there is virtually no other effort reported on automatic human affect analysis from combined face and body gestures. In the same way, 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 [104][195].

Examples of existing works combining different modalities into a single system for human affective state analysis are those of Chen [27], Yoshitomi et al. [192], De Silva and Ng [166], Go et al. [57], and Song et al. [169], who investigated the effects of a combined detection of facial and vocal expressions of affective states. In brief, these works achieve an accuracy of 72% to 85% when detecting one or more basic emotions from clean audiovisual input (e.g., noise-free recordings, closely-placed microphone, non-occluded portraits) from an actor speaking a single word and showing exaggerated facial displays of a basic emotion. Although audio and image processing techniques in these systems are relevant to the discussion on the state of the art in affective computing, the systems themselves have most of the drawbacks of unimodal affect analyzers. Many improvements are needed if those systems are to be used for multimodal HCI where clean input from a known actor/announcer cannot be expected and a context independent separate processing and interpretation of audio and visual data does not suffice.

5. Modeling, Fusion, and Data Collection

Multimodal interface design [146] 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 Section 2, include the 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 [164][70][75].

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

[138]. 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.

First we discuss architectures, followed by modeling, fusion, data collection, and testing.

5.1 System Integration Architectures

The most common infrastructure that has been adopted by the multimodal research community involves multi-agent architectures such as the *Open Agent Architecture* [97] and *Adaptive Agent Architecture* [86][31]. Multi-agent architectures provide essential infrastructure for coordinating the many complex modules needed to implement multimodal system processing and permit this to be done in a distributed manner. In a multi-agent architecture, the components needed to support the multimodal system (e.g., speech recognition, gesture recognition, natural language processing, multimodal integration) may be written in different programming languages, on different machines, and with different operating systems. Agent communication languages are being developed that handle asynchronous delivery, triggered responses, multi-casting, and other concepts from distributed systems.

When using a multi-agent architecture, for example, speech and gestures can arrive in parallel or asynchronously via individual modality agents, with the results passed to a *facilitator*. These results, typically an *n*-best list of conjectured lexical items and related time-stamp information, are then routed to appropriate agents for further language processing. Next, sets of meaning fragments derived from the speech, or other modality, arrive at the multimodal *integrator* which decides whether and how long to wait for recognition results from other modalities, based on the system's temporal thresholds. It fuses the meaning fragments into a semantically and temporally compatible whole interpretation before passing the results back to the facilitator. At this point, the system's final multimodal interpretation is confirmed by the interface, delivered as multimedia feedback to the user, and executed by the relevant application.

Despite the availability of high-accuracy speech recognizers and the maturing of devices such as gaze trackers, touch screens, and gesture trackers, very few applications take advantage of these technologies. One reason for this may be that the cost in time of implementing a multimodal interface is very high. If someone wants to equip an application with such an interface, he must usually start from scratch, implementing access to external sensors, developing ambiguity resolution algorithms, etc. However, when properly implemented, a large part of the code in a multimodal system can be reused. This aspect has been identified and many multimodal application frameworks (using multi-agent architectures) have recently appeared such as VTT's Japis framework [179], Rutgers CAIP Center framework [49], and the Embassi system [44].

5.2 Modeling

There have been several attempts for modeling humans in human-computer interaction literature [191]. Here we present some proposed models and we discuss their particularities and weaknesses.

One of the most commonly used models in HCI is the *Model Human Processor*. The model, proposed in [24] is a simplified view of the human processing involved in interacting with computer systems. This model comprises three subsystems namely, *the perceptual system* handling sensory stimulus from the outside world, *the motor system* that controls actions, and *the cognitive system* that provides the necessary processing to connect the two. Retaining the analogy of the user as an information processing system, the components of an MMHCI model include an input-output component (sensory system), a memory component (cognitive system), and a processing component (motor system). Based on this model, the study of input-output channels (vision, hearing, touch, movement), human memory (sensory, short-term, and working or long-term memory), and processing capabilities (reasoning, problem solving, or acquisition skills) should all be considered when designing MMHCI systems and applications. Many studies in the literature analyze each subsystem in detail and we point the interested reader to [39] for a comprehensive analysis.

Another model proposed by Card et al. [24] is the GOMS (**Goals, Operators, Methods, and Selection rules)** model. GOMS is essentially a reduction of a user's interaction with a computer to its elementary actions and all existing GOMS variations [24] allow for different aspects of an interface to be accurately studied and predicted. For all of the variants, the definitions of the major concepts are the same. **Goals** are what the user intends to accomplish. An **operator** is an action performed in service of a goal. A **method** is a sequence of operators that accomplish a goal and if more than one method exists, then one of them is chosen by some **selection rule**. Selection rules are often ignored in typical GOMS analyses. There is some flexibility for the designers/analysts definition of all of these entities. For instance, one person's operator may be another's goal. The level of granularity is adjusted to capture what the particular evaluator is examining.

All of the GOMS techniques provide valuable information, but they all also have certain drawbacks. None of the techniques address user fatigue. Over time a user's performance degrades simply because the user has been performing the same task repetitively. The techniques are very explicit about basic movement operations, but are generally less rigid with basic cognitive actions. Further, all of the techniques are only applicable to expert users and the functionality of the system is ignored while only the usability is considered.

The **human action cycle** [114] is a psychological model which describes the steps humans take when they interact with computer systems. The model can be used to help evaluate the efficiency of a user interface (UI). Understanding the cycle requires an understanding of the user interface design principles of affordance, feedback, visibility, and tolerance. This model describes how humans may form goals and then develop a series of steps required to achieve that goal, using the computer system. The user then executes the steps, thus the model includes both cognitive and physical activities.

5.3 Adaptability

The number of computer users (and computer-like devices we interact with) has grown at an incredible pace in the last few years. An immediate consequence of this is that there is much larger diversity in the "types" of computer users. Increasing differences in skill level, culture, language, and goals have resulted in a significant trend towards adaptive and customizable interfaces, which use modeling and reasoning about the domain, the task, and the user, in order to extract and represent the user's knowledge, skills, and goals, to better serve the users with their tasks. The goal of such systems is to adapt their interface to a specific user, give feedback about the user's knowledge, and predict the user's future behavior such as answers, goals, preferences, and actions [76]. Several studies [173] provide empirical support for the concept that user performance can be increased when the interface characteristics match the user skill level, emphasizing the importance of adaptive user interfaces.

Adaptive human-computer interaction promises to support more sophisticated and natural input and output, to enable users to perform potentially complex tasks more quickly, with greater accuracy, and to improve user satisfaction. This new class of interfaces promises knowledge or agent-based dialog, in which the interface gracefully handles errors and interruptions, and dynamically adapts to the current context and situation, the needs of the task performed, and the user model. This interactive process is believed to have great potential for improving the effectiveness of human-computer interaction [100], and therefore, is likely to play a major role in MMHCI. The overarching aim of intelligent interfaces is to both increase the interaction bandwidth between human and machine and, at the same time, increase interaction effectiveness and naturalness by improving the quality of interaction. Effective human machine interfaces and information services will also increase access and productivity for all users [89]. A grand challenge of adaptive interfaces is therefore to represent, reason, and exploit various models to more effectively process input, generate output, and manage the dialog and interaction between human and machine so that to maximize the efficiency, effectiveness, and naturalness, if not joy, of interaction [133].

One central feature of adaptive interfaces is the manner in which the system uses the learned knowledge. Some works in applied machine learning are designed to produce expert systems that are intended to replace the human. However, works in adaptive interfaces intend to construct advisory-recommendation systems, which only make recommendations to the user. These systems suggest information or generate actions that the user can always override. Ideally, these actions should reflect the preferences of the individual users, thus providing personalized services to each one.

Every time the system suggests a choice to the user he/she accepts or rejects it, thus giving feedback to the system to update its knowledgebase either implicit or explicit [6]. The system should carry out online learning, in which the knowledgebase is updated each time an interaction with the user occurs. Since adaptive user interfaces collect data during their interaction with the user, one naturally expects them to improve during the interaction process, making them "learning" systems rather than "learned" systems. Because adaptive user interfaces must learn from observing the behavior of their users, another distinguishing characteristic of these systems is their need for rapid learning. The issue here is the number of training cases needed by the system to generate good advice. Thus, it is recommended the use of learning methods

and algorithms that achieve high accuracy from small training sets. On the other hand, the speed of interface adaptation to user's needs is desirable but not essential.

Adaptive user interfaces should not be considered a panacea for all problems. The designer should seriously take under consideration if the user really needs an adaptive system. The most common concern regarding the use of adaptive interfaces is the violation of standard usability principles. In fact, there exists evidence that suggests that static interface designs sometimes promote superior performance than adaptive ones [64][163]. Nevertheless, the benefits that adaptive systems can bring are undeniable and therefore more and more research efforts are being paid towards this direction.

An important issue is how the interaction techniques should change to take this varying input and output hardware devices into account. The system might choose the appropriate interaction techniques taking into account the input and output capabilities of the devices and the user preferences. So, nowadays, many researchers are focusing on such fields as context aware interfaces, recognition-based interfaces, intelligent and adaptive interfaces, and multimodal perceptual interfaces [76][100][89][176][177].

Although there have been many advances in MMHCI, the level of adaptability in current systems is rather limited and there are many challenges left to be investigated.

5.4 Fusion

Fusion techniques are needed to integrate input from different modalities and many fusion approaches have been developed. Early multimodal interfaces were based on a specific control structure for multimodal fusion. For example, Bolt's "Put-That-There" system [18] combined pointing and speech inputs and searched for a synchronized gestural act that designates the spoken referent. To support more broadly functional multimodal systems, general processing architectures have been developed which handle a variety of multimodal integration patterns and support joint processing of modalities [16][86][97].

A typical issue of multimodal data processing is that multisensory data are 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 [27]). Therefore, in order to accomplish a human-like multimodal analysis of multiple input signals acquired by different sensors, the signals cannot be always considered mutually independently and might not be combined in a context-free manner at the end of the intended analysis but, on the contrary, the input data might preferably 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. [130].

Multimodal systems usually integrate signals at the *feature level* (early fusion), at a *higher semantic level* (late fusion), or *something in between* (intermediate fusion) [178][37]. In the following we present in detail these fusion techniques and we analyze their advantages and disadvantages.

Early Fusion Techniques

In an early fusion architecture, the signal-level recognition process in one mode influences the course of recognition in the other and so, this type of fusion is considered more appropriate for closely temporally synchronized input modalities, such as speech and lip movements. This class of techniques utilizes, therefore, a single classifier avoiding the explicit modeling of the different modalities.

To give an example of audio-visual integration using an early fusion approach, one simply concatenates the audio and visual feature vectors to obtain a single combined audio-visual vector [1]. In order to reduce the length of the resulting feature vector, dimensionality reduction techniques like LDA are usually applied before the feature vector finally feeds the recognition engine. The classifier utilized by most early integration systems is a conventional HMM which is trained with the mixed audio-visual feature vector.

Intermediate Fusion Techniques

Since the early fusion techniques avoid explicit modeling of the different modalities, they fail to model both the fluctuations in the relative reliability and the asynchrony problems between the distinct (e.g., audio and visual) streams. Moreover, 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 non-verbal communicative signals [132]. By considering previously observed data with respect to the current data carried by functioning observation channels, 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 [156]. These models can handle noisy features, temporal information, and missing values of features all by probabilistic inference. Hierarchical HMM-based systems [32] have been shown to work well for facial expression recognition. Dynamic Bayesian Networks and HMM variants [54] 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 [53]. 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.

Late Integration Techniques

Multimodal systems based on late (semantic) fusion integrate common meaning representations derived from different modalities into a combined final interpretation. This requires a common meaning representation framework for all modalities used

and a well-defined operation for integrating partial meanings. Meaning representation uses data structures such as *frames* [106], *feature structures* [81], or *typed feature structures* [25]. *Frames* represent objects and relations as consisting of nested sets of attribute/value pairs while *feature structures* go further to use shared variables to indicate common substructures. *Typed feature structures* are pervasive in natural language processing and their primary operation is unification, which determines the consistency of two representational structures and, if they are consistent, their combination.

Late integration models often utilize independent classifiers (e.g., HMMs), one for each stream, which can be trained independently. The final classification decision is reached by combining the partial outputs of the unimodal classifiers. The correlations between the channels are taken into account only later during the integration step. There are advantages though when using late integration [178]. Since the input types can be recognized independently, they do not have to occur simultaneously. Moreover, the training requirements are smaller $O(2N)$ for two separately trained modalities as compared to $O(N^2)$ for two modalities trained together. The software development process is also simpler in the late integration case [178].

Systems using the late fusion approach have been applied to processing multimodal speech and pen input or manual gesturing, for which the input modes are less coupled temporally and provide different but complementary information. Late semantic integration systems use individual recognizers that can be trained using unimodal data and can be scaled up more easily in a number of input modes or vocabulary size. To give an example, in an audio-visual speech recognition application [143], for small-vocabulary independent word speech recognition, late integration can be easily implemented by combining the audio- and visual-only log-likelihood scores for each word model in the vocabulary, given the acoustics and visual observations [1]. However, this approach is intractable in the case of connected word recognition where the number of alternative paths explodes. A good heuristic alternative in that case is through lattice rescoring. The *n*-most promising hypotheses are extracted from the audio-only recognizer and they are rescored after taking the visual evidence into account. The hypothesis with the highest combined score is then selected. More details about this approach can be found in [143].

Very important for MMHCI applications is the fusion between the visual and audio channels. Successful audio and visual feature integration requires utilization of advanced techniques and models for cross-model information fusion. This research area is currently very active and many different paradigms have been proposed for addressing the general problem. Therefore, we will confine ourselves to discussing the problem of feature fusion in the context of audiovisual feature integration. We note, however, that similar issues exist in the integration of *any* other modalities.

When considering audio-visual integration, the main task that needs to be addressed is the fusion of the heterogeneous pool of audio and visual features in a way that ensures that the combined audiovisual system outperforms its audio- or video-only counterpart in all practical scenarios. This task is complicated due to a couple of issues, the most important being:

(1) Audio and visual speech asynchrony. Although the audio and visual observation sequences are certainly correlated over time, they exhibit state asynchrony

(e.g., in speech, lip movement is preceding auditory activity by as much as 120 ms [21], close to the average duration of a phoneme). This asynchrony is critical in applications such as audio-visual speech recognition because it renders modeling audiovisual speech with conventional HMMs problematic [145];

(2) The relative discriminating power of the audio and visual streams can vary dramatically in unconstrained environments, making their optimal fusion a challenging task.

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

5.5 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. For example, in developing multimodal techniques for emotion recognition, 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. Asking someone to smile often does not create the same picture as an authentic smile. The fundamental reason of course is that the subject often does not feel happy so his smile is artificial and in many subtle ways quite different than a genuine smile [133]. 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. Cohen et al. [31] showed that unlabeled data could 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 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. As a consequence, 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.

Picard et al. [141] outlined five factors that influence the affective data collection. We list them below since they also apply to the more general problem of MMHCI data collection:

- *Spontaneous* versus *posed*: Is the emotion elicited by a situation or stimulus that is outside the subject's control or the subject is asked to elicit the emotion?
- *Lab setting* versus *real-world*: Is the data recording taking place in a lab or the emotion is recorded in the usual environment of the subject?
- *Expression* versus *feeling*: Is the emphasis on external expression or on internal feeling?
- *Open recording* versus *hidden recording*: Is the subject aware that he is being recorded?
- *Emotion-purpose* versus *other-purpose*: Does the subject know that he is a part of an experiment and the experiment is about emotion?

Note that these factors are not necessarily independent. The most natural setup would imply that the subject feels the emotion internally (*feeling*), and the emotion occurs *spontaneously*, while the subject is in his usual environment (*real-world*). Also, the subject should not know that he is being recorded (*hidden recording*) and that he is part of an experiment (*other-purpose*). Such data are usually impossible to obtain because of privacy and ethics concerns. As a consequence, most researchers who tackled the problem of establishing a comprehensive human-affect expression database used a setup that is rather far from the natural setup [133]: a *posed*, *lab-based*, *expression-oriented*, *open-recording*, and *emotion-purpose* methodology.

5.6 Evaluation

Evaluation is a very important issue in the design of multimodal systems. Here, we outline the most important features that could be used as measures in the evaluation of various types of adaptive MMHCI systems namely, *efficiency*, *quality*, *user satisfaction*, and *predictive accuracy*.

People typically invoke computational decision aids because they expect the system will help them accomplish their tasks more rapidly and with less effort than they do on their own. This makes *efficiency* an important measure to use in evaluating adaptive systems. One natural measure of efficiency is the time the user takes to accomplish his task. Another facet is the effort the user must exert to make a decision or solve a problem. In this case, the measure would be the number of user actions or commands that take place during the solving of a problem.

Another main reason the users turn to MMHCI systems is to improve the *quality* of solutions of their task. As with efficiency, there are several ways in which one can define the notion of quality or *accuracy* of the system. For example, if there is a certain object the user wants to find then the success of finding it constitutes an objective measure of quality. However, it is clear that in some cases it is necessary to rely on a separate measure of *user satisfaction* to determine the quality of the system's behavior. One way to achieve this is to present each user with a questionnaire that asks about his subjective experience. Another measure of user's satisfaction involves giving the user some control over certain features of the system. If the user turns the sys-

tem's advisory capability off or disables its personalization module, one can then conclude that the user has not been satisfied by his experience with these features.

Since many adaptive system user models make predictions about the user's responses, it is natural to measure the *predictive accuracy* to determine the success of a system. Although this measure can be a useful analytical tool for understanding the details of the system's behavior, it does not necessarily reflect the overall efficiency or quality of solutions, which should be the main concern.

6. Applications

Throughout the paper we have discussed techniques in a wide variety of application scenarios, including video conferencing and remote collaboration, intelligent homes, and driver monitoring. The types of modalities used, as well as the integration models vary widely from application to application. The literature on applications that use MMHCI is vast and could well deserve a survey of its own [74]. Therefore, we do not attempt a complete survey of MMHCI applications. Instead we give a general overview of some of the major areas (also see [84]) by focusing on specific application areas in which interesting progress has been made. In particular, we focus on the areas below.

Ambient Spaces

Computing is expanding beyond the desktop, integrating with everyday objects in a variety of scenarios. As our discussions show, this implies that the model of user interface in which a person sits in front of a computer is no longer the only model [5]. One of the implications of this is that the actions or events to be recognized by the "interface" are not necessarily explicit commands. In smart conference room applications, for instance, multimodal analysis has been applied mostly for video indexing [102] (see [139] and [55] for social analysis applications). 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 [105]. Projects in the video domain include MVIEWS [30], a system for annotating, indexing, extracting, and disseminating information from video streams for surveillance and intelligence applications. An analyst watching one or more live video feeds is able to use pen and voice to annotate the events taking place. The annotation streams are indexed by speech and gesture recognition technologies for later retrieval, and can be quickly scanned using a timeline interface, then played back during review of the film. Pen and speech can also be used to command various aspects of the system, including image processing functions, with multimodal utterances such as "Track this" or "If any object enters this area, notify me immediately." In [20], the authors present a multimodal attentive cookbook, which combines eye tracking and speech. In [29], human interaction events are detected in a nursing home environment. The authors of [91] use multimodal input in a smart home environment and the authors of [2] propose a design studio that uses multimodal input. Interestingly, techniques used for video analysis can also be used in the context of MMHCI. Examples include human activity recognition [8][17][138], work in the context of meeting video analysis [28][102], event detection [53], surveillance

[65], and others. One of the main differences between some of the approaches developed for video analysis and techniques for MMHCI are the requirements in processing speed. Nonetheless, much of what can be learned from video analysis applications can be applied in the context of MMHCI.

Mobile/wearable

The recent drop in costs of hardware has led to an explosion in the availability of mobile computing devices. One of the major challenges is that while devices such as PDAs and mobile phones have become smaller and more powerful, there has been little progress in developing effective interfaces to access the increased computational and media resources available in such devices. Mobile devices, as well as wearable devices, constitute a very important area of opportunity for research in MMHCI because natural interaction with such devices can be crucial in overcoming the limitations of current interfaces. Several researchers have recognized this, and many projects exist on mobile and wearable MMHCI applications. The authors of [42] integrate pen and speech input for PDA interaction. The use of computer vision, however, is also being explored in projects such as [51], in which a tourist can take photographs of a site to obtain additional information about the site. In [22], the authors present two techniques (head tilt and gesture with audio feedback) to control a mobile device. The authors of [85] use MMHCI to augment human memory: RFID tags are used in combination with a Head Mounted Display and a camera to capture video and information of all the objects the user touches. The authors of [193] combine eye tracking with video, head tracking, and hand motion information. The authors of [137] use eye tracking to understand eye-hand coordination in natural tasks, and in [38] eye tracking is used in a video blogging application, a very interesting area.

Virtual Environments

Virtual reality has been a very active research area at the crossroads of computer graphics, computer vision, and human-computer interaction. One of the major difficulties of VR systems is the HCI component, and many researchers are currently exploring the use of MMHCI to enhance the user experience. One reason MMHCI is very attractive in VR environments is that it helps disambiguate communication between users and the machine (in some cases virtual characters, the virtual environment, or even other users represented by virtual characters [113]). The authors of [95] integrate speech and gesture recognition for interaction in an immersive environment. Speech and gesture inputs are also used in [129], where the user communicates with an autonomous farmer.

Art

Perhaps one of the most exciting application areas of MMHCI is art. Vision techniques can be used to allow audience participation [101] and influence a performance. In [184], 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 [94], 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 [134], limb

movements are tracked to generate music. MMHCI can also be used in museums to augment exhibitions [170].

Users with Disabilities

People with disabilities can benefit greatly from MMHCI technologies [87]. The authors of [167] propose a component-based smart wheel chair system and discuss other approaches that integrate various types of sensors (not only vision). The authors of [87] also present a wheel chair navigation system. In [41], computer vision is used to interpret facial gestures for wheel chair navigation. The authors of [151] introduce a system for presenting digital pictures non-visually (multimodal output), and the techniques in [58] can be used for interaction using only eye blinks and eye brow movements. Some of the approaches in other application areas (e.g., [22]) could also be beneficial for people with disabilities—MMHCI has great potential in making computers and other resources accessible to people with disabilities.

Public and Private Spaces

In this category we place applications in which interfaces are implemented to access devices used in public or private spaces. One attractive example of implementation in public spaces is the use of MMHCI in information kiosks [79][147]. In some ways these are ideal and challenging applications for natural multimodal interaction: the kiosks are often intended to be used by a wide audience, thus there may be few assumptions about the types of users of the system. The range of tasks may also be wide, providing rich opportunities for MMHCI. On the other hand, we have MMHCI applications in private spaces. One interesting area is that of implementation in vehicles. The authors of [172], [168], and [77] present various approaches to monitor vehicle occupants, including the driver. This is an interesting application area due to the constraints: since the driver must focus on the driving task, traditional computer interfaces (e.g., GUIs) are not suitable. Thus, it is an important area of opportunity for MMHCI research, particularly because depending on the particular deployment, vehicle interfaces can be considered safety-critical.

Other

Other applications include biometrics [150][135], surveillance, remote collaboration [52], gaming and entertainment [153], education, and robotics ([50] gives a comprehensive review of socially active robots). MMHCI can also play an important role in safety-critical applications (e.g., medicine, military [31][34], etc.) and in situations in which a lot of information from multiple sources has to be viewed in short periods of time. A good example of this is crisis management [162].

Although not all of the specific applications mentioned above have a strong vision component (or a strong multimodal component), they do highlight some of the major application areas for MMHCI (see [84] for a different classification). Indeed, the range of application areas for MMHCI touches on every aspect of computing, and as computing becomes more ubiquitous, practically every aspect of human interaction with objects, the environment, and human-human interaction (e.g., remote collaboration, etc.) will make use of MMHCI techniques.

Trends in computing suggest that MMHCI will alleviate many of the problems with existing HCI paradigms. Nonetheless, many challenges lay ahead in making widespread use of MMHCI a reality, even in the most constrained applications. Robustness and accuracy are important, and even though computational power continues to increase at lower costs, the limitations on efficient processing of multiple modalities remains strong. Therefore, more research is needed not only on improving separate processing of each modality, but also in efficiently integrating the outputs of multiple processors so that responses can be generated fast enough to make them suitable in their application domains. A factor which cannot be ignored, in considering MMHCI applications, is the adaptability of the user to MMHCI paradigms. Although MMCHI promises natural interaction, at least for the foreseeable future, users will have to make efforts to use MMHCI systems properly. Although MMHCI may not replace WIMP[3]-based graphical user interfaces in the near future [178], the promise of using MMHCI in many applications makes this a very exciting research area full of challenges and opportunities for researchers in various fields. Given the importance of Vision in human communication, we foresee Computer Vision as one of the driving technologies for MMHCI applications [175].

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 (see [122] for a recent study on this topic).

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 [35], 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 [140] affect recognition is most likely to be accurate when it combines multiple modalities, information about the user's context,

[3] WIMP: Windows, Icons, Menus and Pointing device

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, multimodal context-sensitive human-computer interaction is likely to become the single most widespread research topic of the artificial intelligence research community [138]. Advances in this area could change not only how professionals practice computing, but also how mass consumers interact with technology.

Acknowledgements. The work of Nicu Sebe was partially supported by the Muscle NoE and MIAUCE FP6 EU projects.

References

[1] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," *Speech Reading by Humans and Machines,* D. Stork and M. Hennecke, eds., Springer, 1996.
[2] A. Adler, J. Eisenstein, M. Oltmans, L. Guttentag, and R. Davis, "Building the design studio of the future," *AAAI Fall Symposium on Making Pen-Based Interaction Intelligent and Natural*, 2004.
[3] J.K. Aggarwal and Q. Cai, "Human motion analysis: A review," *CVIU*, 73(3):428-440, 1999.
[4] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin,* 111(2):256-274, 1992.
[5] E. Arts, "Ambient intelligence: A multimedia perspective," *IEEE Multimedia*, 11(1):12-19, 2004.
[6] M. Balabanovic, "Exploring versus exploiting when learning user models for text recommendations," *User Modeling and User-adapted Interaction,* 8:71-102, 1998.
[7] T. Balomenos, A. Raouzaiou, S. Ioannou, S. Drosopoulos, A. Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," *Workshop on Machine Learning for Multimodal Interaction,* 2005.
[8] 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.
[9] M. Benali-Khoudja, M. Hafez, J.-M. Alexandre, and A. Kheddar, "Tactile interfaces: A state-of-the-art survey*," Int. Symposium on Robotics*, 2004.
[10] N.O. Bernsen, "A reference model for output information in intelligent multimedia presentation systems," *European Conf. on Artificial Intelligence,* 1996.
[11] N.O. Bernsen, "Foundations of multimodal representations: A taxonomy of representational modalities," *Interacting with Computers,* 6(4):347-71, 1994.
[12] N.O. Bernsen, "Defining a taxonomy of output modalities from an HCI perspective," *Computer Standards and Interfaces, Special Double Issue*, 18(6-7):537-553, 1997.
[13] N.O. Bernsen, "Multimodality in language and speech systems - From theory to design support tool," *Multimodality in Language and Speech Systems*, Granström, B., ed., Kluwer Academic Publishers 2001.
[14] N. Bianchi-Berthouze and C. Lisetti, "Modeling multimodal expression of user's affective subjective experience," *User Modeling and User-adapted Interaction,* 12:49-84, 2002.
[15] M. Black and Y. Yacoob, Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion, *Int'l Conference on Computer Vision,* pp. 374-381, 1995.
[16] M. Blattner and E. Glinert, "Multimodal integration," *IEEE Multimedia,* 3(4):14-24, 1996.
[17] A.F. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on PAMI*, 23(3):257–267, 2001.
[18] R. Bolt, "Put-That-There: Voice and gesture at the graphics interface," *Computer Graphics,* 14(3):262-270, 1980.
[19] C. Borst and R. Volz, "Evaluation of a haptic mixed reality system for interactions with a virtual control panel," *Presence: Teleoperators and Virtual Environments*, 14(6), 2005.
[20] J.S. Bradbury, J.S. Shell, and C.B. Knowles, "Hands on cooking: Towards an attentive kitchen", *ACM Conf. Human Factors in Computing Systems (CHI)*, 2003.
[21] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," *Int. Conf. on Acoustics, Speech, and Signal Processing,* 1994.
[22] S.A. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker, "Multimodal 'Eyes-Free' interaction techniques for wearable devices," *ACM Conf. Human Factors in Computer Systems (CHI,)* 2003.

[23] C.S. Campbell and P.P. Maglio, "A robust algorithm for reading detection," *ACM Workshop on Perceptive User Interfaces*, 2001.
[24] S.K. Card, T. Moran, and A. Newell, *The Psychology of Human-computer Interaction,* Lawrence Erlbaum Associates, 1983.
[25] R. Carpenter, *The Logic of Typed Feature Structures,* Cambridge University Press, 1992.
[26] D. Chen, R. Malkin, and J. Yang, "Multimodal detection of human interaction events in a nursing home environment," *Conf. on Multimodal Interfaces (ICMI)*, 2004.
[27] L.S. Chen, *Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction*, PhD thesis, Univ. of Illinois at Urbana-Champaign, 2000.
[28] L.S. Chen, R. Travis Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, and T.S. Huang, "VACE multimodal meeting corpus," *MLMI* 2005.
[29] D. Chen, R. Malkin, and J. Yang, "Multimodal detection of human interaction events in a nursing home environment," *Conf. on Multimodal Interfaces (ICMI)*, 2004.
[30] A. Cheyer and L. Julia, "MVIEWS: Multimodal tools for the video analyst," *Conf. on Intelligent User Interfaces (IUI),* 1998.
[31] 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.
[32] 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.
[33] P.R. Cohen, M. Johnston, D.R. McGee, S.L. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow, "QuickSet: Multimodal interaction for distributed applications," *ACM Multimedia*, pp. 31-40, 1997.
[34] P.R. Cohen and D.R. McGee, "Tangible multimodal interfaces for safety-critical applications," *Communications of the ACM*, 47(1):41-46, 2004.
[35] P.R. Cohen and S.L. Oviatt, "The role of voice in human-machine communication," *Voice Communication between Humans and Machines,* D. Roe and J. Wilpon, eds., National Academy Press, 1994.
[36] *Computer Vision and Image Understanding,* Special Issue on Eye Detection and Tracking, 98(1), 2005.
[37] A. Corradini, M. Mehta, N. Bernsen, and J.-C. Martin, "Multimodal input fusion in human-computer interaction," *NATO-ASI Conf. on Data Fusion for Situation Monitoring, Incident Detection, Alert, and Response Management*, 2003
[38] 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.
[39] A. Dix, J. Finlay, G. Abowd, and R. Beale, *Human-computer Interaction,* Prentice Hall, 2003.
[40] 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.
[41] A.T. Duchowski, "A breadth-first survey of eye tracking applications," *Behavior Research Methods, Instruments, and Computing*, 34(4):455-70, 2002.
[42] S. Dusan, G.J. Gadbois, and J. Flanagan, "Multimodal interaction on PDA's integrating speech and pen inputs," *Eurospeech* 2003.
[43] P. Ekman, ed., *Emotion in the Human Face*, Cambridge University Press, 1982.
[44] C. Elting, S. Rapp, G. Mohler, and M. Strube, "Architecture and implementation of multimodal plug and play," *ICMI,* 2003.
[45] I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. on PAMI,* 19(7):757-763, 1997.
[46] C. Fagiani, M. Betke, and J. Gips, "Evaluation of tracking methods for human-computer interaction," *IEEE Workshop on Applications in Computer Vision*, 2002.
[47] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, 36:259–275, 2003.
[48] G. Fitzmaurice, H. Ishii, and W.Buxton, "Bricks: Laying the foundations for graspable user interfaces," *ACM Conf. Human Factors in Computer Systems (CHI)*, 1(442-449), 1995.
[49] F. Flippo, A. Krebs, and I. Marsic, "A framework for rapid development of multimodal interfaces," *ICMI,* 2003.
[50] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems,* 42(3-4):143-166, 2003.
[51] G. Fritz, C. Seifert, P. Luley, L. Paletta, and A. Almer, "Mobile vision for ambient learning in urban environments in urban environments," *Int. Conf. on Mobile Learning (MLEARN)*, 2004.

[52] S. Fussell, L. Setlock, J. Yang, J. Ou, E. Mauer, and A. Kramer, "Gestures over video streams to support remote collaboration on physical tasks," *Human-computer Interaction*, 19(3):273-309, 2004.
[53] 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.
[54] 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.
[55] D. Gatica-Perez, "Analyzing group interactions in conversations: A survey", *IEEE Int. Conf. Multisensor Fusion and Integration for Intelligent Systems,* pp. 41-46, 2006.
[56] D.M. Gavrila, "The visual analysis of human movement: A survey," *CVIU*, 73(1):82-98, 1999.
[57] H.J. Go, K.C. Kwak, D.J. Lee, and M. Chun, "Emotion recognition from facial image and speech signal," *Conf. on the Society of Instrument and Control Engineers,* 2003.
[58] K. Grauman, M. Betke, J. Lombardi, J. Gips, and G. Bradski, "Communication via eye blinks and eyebrow raises: Video-based human-computer interfaces," *Universal Access in the Information Society*, 2(4):359-373, 2003.
[59] H. Gunes, M. Piccardi, and T. Jan, "Face and body gesture recognition for a vision-based multimodal analyzer," *Pan-Sydney Area Workshop on Visual Information Processing*, Vol. 36, 2004.
[60] A. Hakeem and M. Shah, "Ontology and taxonomy collaborated framework for meeting classification," *ICPR*, 2004.
[61] A. Hanjalic and L-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. on Multimedia,* 7(1):143– 154, 2005.
[62] 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.
[63] E. Hjelmas and B. K. Low, "Face detection: A survey," *CVIU*, 83:236–274, 2001.
[64] K. Hook, "Designing and evaluating intelligent user interfaces," *Int. Conf. on Intelligent User Interfaces,* 1999.
[65] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors", *IEEE Trans. on Systems, Man, and Cybernetics*, 34(3), 2004.
[66] *IEEE Computer, Special Issue on Human-centered Computing,* A. Jaimes, N. Sebe, D. Gatica-Peres, and T.S. Huang, eds., May 2007.
[67] *Int. J. of Human-Computer Studies*, *Special Issue on Applications of Affective Computing in Human-computer Interaction*, 59(1-2), 2003.
[68] S. Intille, K. Larson, J. Beaudin, J. Nawyn, E. Tapia, and P. Kaushik, "A living laboratory for the design and evaluation of ubiquitous computing technologies," *ACM Conf. Human Factors in Computing Systems (CHI)*, 2004.
[69] R. Jacob, "The use of eye movements in human-computer interactions techniques: What you look at is what you get," *ACM Trans. Information Systems,* 9(3):152-169, 1991.
[70] A. Jaimes, "Human-centered multimedia: Culture, deployment, and access", *IEEE Multimedia*, 13(1):12 –19, 2006.
[71] A. Jaimes and J. Liu, "Hotspot components for gesture-based interaction," *IFIP Interact*, 2005.
[72] A. Jaimes, T. Nagamine, J. Liu, K. Omura, and N. Sebe, "Affective meeting video analysis," *IEEE Int. Conf. on Multimedia and Expo,* 2005.
[73] A. Jaimes and N. Sebe, "Multimodal human computer interaction: A survey," *IEEE Int. Workshop on Human-computer Interaction*, 2005.
[74] A. Jaimes, N. Sebe, and Daniel Gatica-Perez, "Human-centered computing: A multimedia perspective," *ACM Multimedia,* pp 855-864, 2006.
[75] R. Jain, "Folk computing," *Comm. of the ACM*, 46(4):27-29, 2003.
[76] A. Jameson, R. Schafer, T. Weis, A, Berthold, and T. Weyrath, "Making systems sensitive to the user's time and working memory constraints," *Int. Conf. on Intelligent User Interfaces,* 1997.
[77] Q. Ji and X. Yang, "Real-time eye, gaze, and face pose tracking for monitoring driver vigilance," *Real-Time Imaging,* 8:357-377, 2002.
[78] M. Johnston, P. Cohen, D. McGee, S.L. Oviatt, J. Pittman, I. Smith, "Unification-based multimodal integration," *Annual Meeting of the Association for Computational Linguistics,* 1998.
[79] M. Johnston and S. Bangalore, "Multimodal Applications from Mobile to Kiosk," *W3C Workshop on Multimodal Interaction,* 2002.
[80] 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.
[81] M. Kay, "Functional unification grammar: A formalism for machine translation," *Int. Conf. on Computational Linguistics,* 1984.

[82] S. Kettebekov and R. Sharma, "Understanding gestures in multimodal human computer interaction," *Int. J. on Artificial Intelligence Tools*, 9(2):205-223, 2000.
[83] 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.
[84] B. Kisacanin, V. Pavlovic, and T.S.Huang, eds., *Real-Time Vision for Human-Computer Interaction*, Springer-Verlag, 2005.
[85] Y. Kono, T. Kawamura, T. Ueoka, S. Murata, and M. Kidode, "Real world objects as media for augmenting human memory," *Workshop on Multi-User and Ubiquitous User Interfaces(MU3I)*, pp.37-42, 2004.
[86] S. Kumar and P. Cohen, "Towards a fault-tolerant multi-agent system architecture," *Int. Conf. on Autonomous Agents,* 2000.
[87] 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.
[88] P. Lang, "The emotion probe: Studies of motivation and attention," *American Psychologist*, 50(5):372–385, 1995.
[89] P. Langley, "User modeling in adaptive interfaces," *Int. Conf. on User Modeling,* 1998.
[90] A. Lanitis, C.J. Taylor, and T. Cootes, "A unified approach to coding and interpreting face images," *Int'l Confernece on Computer Vision,* pp. 368-373, 1995.
[91] V. Lauruska and P. Serafinavicius, "Smart home system for physically disabled persons with verbal communication difficulties," *Assistive Technology Research Series (AAATE)*, pp.579-583, 2003.
[92] A. Legin, A. Rudnitskaya, B. Seleznev, and Y. Vlasov, "Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie," *Analytica Chimica Acta*, 534:129-135, 2005.
[93] J. Lien, *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity,* PhD thesis, Carnegie Mellon University, 1998.
[94] 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.
[95] A.M. Malkawi and R.S. Srinivasan, "Multimodal human-computer interaction for immersive visualization: Integrating speech-gesture recognition and augmented reality for indoor environments," *Int'l Association of Science & Technology for Development Conf. on Computer Graphics and Imaging,* 2004.
[96] S. Marcel, "Gestures for multi-modal interfaces: A Review," *Technical Report IDIAP-RR 02-34*, 2002.
[97] D. Martin, A. Cheyer, and D. Moran, "The open agent architecture: A framework for building distributed software systems," *Applied Artificial Intelligence,* 13:91-128, 1999.
[98] A. Martinez, "Face image retrieval using HMMs," *IEEE worlshop on Content-based access of images and video libraries,* pp. 35-39, 1999.
[99] K. Mase, "Recognition of facial expressions from optical flow," *IEICE Trans.,* E74(10):3474-3483, 1991.
[100] M. Maybury, *Intelligent Multimedia Interfaces,* AAAI/MIT Press, 1993.
[101] D. Maynes-Aminzade, R. Pausch, and S. Seitz, "Techniques for interactive audience participation," *ICMI* 2002.
[102] 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.
[103] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought,* Univ. of Chicago Press, 1992.
[104] A. Mehrabian, "Communication without words," *Psychology Today*, 2(4):53–56, 1968.
[105] S. Meyer and A. Rakotonirainy, "A Survey of research on context-aware homes," *Australasian Information Security Workshop Conference on ACSW Frontiers*, 2003
[106] M. Minsky, "A framework for representing knowledge," *The Psychology of Computer Vision,* P. Winston, ed., McGraw-Hill, 1975.
[107] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *CVIU*, 81(3):231-258, 2001.
[108] T. Moriyama, T. Kanade, J. Xiao, and J. Cohn, "Meticulously Detailed Eye Region Model and Its Application to Analysis of Facial Images," *IEEE Trans. on PAMI,* 28(5):738-752, 2006.
[109] 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.
[110] A. Nefian and M. Hayes, "Face recognition using an embedded HMM," *IEEE Conf. on Audio and Video-based Biometric Person Authentication,* pp. 19-24, 1999.
[111] J. Nielsen, "Non-command user interfaces," *Comm. of the ACM,* 36(4):83-99, 1993.

[112] L. Nigay and J. Coutaz, "A design space for multimodal systems: Concurrent processing and data fusion," *ACM Conf. Human Factors in Computing Systems (CHI),* 1993.
[113] A. Nijholt and D. Heylen, "Multimodal communication in inhabited virtual environments," *Int. J. of Speech Technology* 5:343–354, 2002.
[114] D.A. Norman, *The Design of Everyday Things,* Doubleday, 1988.
[115] Z. Obrenovic and D. Starcevic, "Modeling multimodal human-computer interaction," *IEEE Computer*, pp. 65-72, September, 2004.
[116] N. Oliver, A. Pentland, and F. Berard, "LAFTER: A real-time face and lips tracker with facial expression recognition," *Pettern Recognition,* 33:1369-1382, 2000.
[117] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences," *IEEE Conf. on Image Processing,* pp. 546-549, 1997.
[118] T. Otsuka and J. Ohya, "A study of transformation of facial expressions based on expression recognition from temporal image sequences," *Technical Report, Institute of Electronic information, and Communications Engineers (IEICE),* 1997.
[119] 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.
[120] A. Oulasvirta and A. Salovaara, "A cognitive meta-analysis of design approaches to interruptions in intelligent environments," *ACM Conf. Human Factors in Computing Systems (CHI),* 2004.
[121] P. Qvarfordt, and S. Zhai, "Conversing with the user based on eye-gaze patterns," *ACM Conf. Human Factors in Computing Systems (CHI)*, 2005.
[122] S.L. Oviatt, R. Lunsford, and R. Coulston, "Individual differences in multimodal integration patterns: What are they and why do they exist?" *ACM Conf. Human Factors in Computing Systems (CHI)*, 2005.
[123] S.L. Oviatt, "Mutual disambiguation of recognition errors in a multimodal architecture," *ACM Conf. Human Factors in Computing Systems (CHI),* 1999.
[124] S.L. Oviatt, "Ten myths of multimodal interaction," *Comm. of the ACM,* 42(11):74-81, 1999.
[125] S.L. Oviatt, T. Darrell, and M. Flickner, eds. *Comm. of the ACM,* Special Issue on Multimodal interfaces that flex, adapt, and persist, 47(1), 2004.
[126] S.L. Oviatt and P. Cohen, "Multimodal interfaces that process what comes naturally," *Comm. of the ACM,* 43(3):45-48, 2000.
[127] S.L. Oviatt, "Multimodal interfaces," *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, chap.14, 286-304, 2003.
[128] 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 Interaction*, 15:263-322, 2000.
[129] P. Paggio and B. Jongejan, "Multimodal communication in the virtual farm of the staging Project," *Multimodal Intelligent Information Presentation,* O. Stock and M. Zancanaro, eds., pp. 27-46, Kluwer Academic Publishers, 2005.
[130] H. Pan, Z.P. Liang, T.J. Anastasio, and T.S. Huang. "Exploiting the dependencies in information fusion," *CVPR*, vol. 2:407–412, 1999.
[131] 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.
[132] M. Pantic and L.J.M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, 91(9):1370–1390, 2003.
[133] M. Pantic, N. Sebe, J. Cohn, and T.S. Huang, "Affective multimodal human-computer interaction," *ACM Multimedia,* 2005.
[134] 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.
[135] *Pattern Recognition Letters, Special Issue on Multimodal Biometrics,* Vol. 24, No. 13, September 2003.
[136] 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.
[137] J.B. Pelz, "Portable eye-tracking in natural behavior," *J. of Vision*, 4(11), 2004.
[138] A. Pentland, "Looking at people," *Comm. of the ACM,* 43(3):35-44, 2000.
[139] A. Pentland, "Socially aware computation and communication," *IEEE Computer*, 38(3), 2005.
[140] R.W. Picard, *Affective Computing*, MIT Press, 1997.

[141] R.W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. on PAMI*, 23(10):1175–1191, 2001.
[142] M. Porta, "Vision-based user interfaces: methods and applications," *Int. J. Human-Computer Studies*, 57(1):27-73, 2002.
[143] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in Visual and Audio-Visual Speech Processing*, E. Vatikiotis-Bateson and P. Perrier, eds., MIT Press, 2004.
[144] *Proceedings of the IEEE, Special Issue on Multimodal Human Computer Interface*. August, 2003.
[145] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition,* Prentice Hall, 1993.
[146] L.M. Reeves, J-C. Martin, M. McTear, T. Raman, K. Stanney, H. Su, Q. Wang, J. Lai, J. Larson, S. Oviatt, T. Balaji, S. Buisine, P. Collings, P. Cohen, and B. Kraal, "Guidelines for multimodal user interface design," *Communications of the ACM*, 47(1):57-69, 2004.
[147] J. Rehg, M. Loughlin, and K. Waters, "Vision for a Smart Kiosk," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 690-696. 1997.
[148] R. Rosales and S. Sclaroff, "Learning body pose via specialized maps," *NIPS*, vol. 14, pp 1263-1270, 2001.
[149] M. Rosenblum, Y. Yacoob, and L. Davis, "Human expression recognition from motion using radial basis function network architecture," *IEEE Trans. on Neurel Networks,* 7(5):1121-1138, 1996.
[150] A. Ross and A.K. Jain, "Information Fusion in Biometrics," *Pattern Recognition Letters, Special Issue on Multimodal Biometrics*, Vol. 24, No. 13, pp. 2115-2125, September 2003.
[151] P. Roth and T. Pun, "Design and evaluation of a multimodal system for the non-visual exploration of digital pictures," *INTERACT* 2003.
[152] R. Ruddaraju, A. Haro, K. Nagel, Q. Tran, I. Essa, G. Abowd, and E. Mynatt, "Perceptual user interfaces using vision-based eye tracking," *ICMI,* 2003.
[153] K. Salen and E. Zimmerman, *Rules of Play: Game Design Fundamentals*, MIT Press, 2003.
[154] 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.
[155] N. Sebe, I. Cohen, and T.S. Huang, "Multimodal emotion recognition," *Handbook of Pattern Recognition and Computer Vision*, World Scientific, 2005.
[156] N. Sebe, I. Cohen, A. Garg, and T.S. Huang, *Machine Learning in Computer Vision,* Springer, 2005.
[157] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues," *Int. Conf. on Pattern Recognition,* pp. 1136-1139, 2006.
[158] E. Schapira and R. Sharma, "Experimental evaluation of vision and speech based multimodal interfaces," W*orkshop on Perceptive User Interfaces*, pp. 1-9, 2001.
[159] B. Schuller, M. Lang, and G. Rigoll, "Multimodal emotion recognition in audiovisual communication," *ICME,* 2002.
[160] T. Selker, "Visual attentive interfaces," *BT Technology Journal*, 22(4):146-150, 2004.
[161] A. Shaikh, S. Juth, A. Medl, I. Marsic, C. Kulikoswki, and J. Flanagan, "An architecture for multimodal information fusion," *Workshop on Perceptual User Interfaces,* 1997.
[162] 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.
[163] B. Shneiderman, "Direct manipulation for comprehensible, predictable, and controllable user interface," *Int. Conf. on Intelligent User Interfaces,* 1997.
[164] B. Shneiderman, *Leonardo's Laptop: Human Needs and the New Computing Technologies*, MIT Press, 2002.
[165] L.E. Sibert and R.J.K. Jacob, "Evaluation of eye gaze interaction," *ACM Conf. Human Factors in Computing Systems (CHI)*, pp. 281-288, 2000.
[166] L.C. de Silva and P. Ng, "Bimodal emotion recognition," *Int. Conf. on Face and Gesture Recognition,* 2000.
[167] R. Simpson, E. LoPresti, S. Hayashi, I. Nourbakhsh, and D. Miller, "The smart wheelchair component system," *J. of Rehabilitation Research and Development*, May/June 2004.
[168] P. Smith, M. Shah, and N.d.V. Lobo, "Determining driver visual attention with one camera," *IEEE Trans. on Intelligent Transportation Systems*, 4(4), 2003.
[169] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition: A new approach," *CVPR*, 2004.
[170] F. Sparacino, "The museum wearable: Real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences," *Museums and the Web*, 2002.

[171] O. Stock, and M. Zancanaro, (eds.). *Multimodal Intelligent Information Presentation. Series Text, Speech and Language Technology*. Vol 27. Kluwer Academic. pp. 325-340, 2005.
[172] M.M. Trivedi, S.Y. Cheng, E.M.C. Childers, and 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.
[173] J. Trumbly, K. Arnett, and P. Johnson, "Productivity gains via an adaptive user interface," *J. of Human-computer Studies,* 40:63-81, 1994.
[174] M. Turk, "Gesture recognition," *Handbook of Virtual Environment Technology*, K. Stanney (ed.), 2001.
[175] M. Turk, "Computer vision in the interface," *Comm. of the ACM*, 47(1):60-67, 2004.
[176] M. Turk and G. Robertson, "Perceptual Interfaces," *Comm. of the ACM*, 43(3):32-34, 2000.
[177] M. Turk and M. Kölsch, "Perceptual interfaces," G. Medioni and S.B. Kang, eds., *Emerging Topics in Computer Vision*, Prentice Hall, 2004.
[178] M. Turk, "Multimodal human-computer interaction," *Real-time Vision for Human-computer Interaction,* B. Kisancanin, V. Pavlovic, and T.S. Huang, eds., Springer, 2005.
[179] M. Turunen and J. Hakulinen, "Jaspis2 – An architecture for supporting distributed spoken dialogs," *Eurospeech,* 2003.
[180] R. Vertegaal, ed., "Attentive user interfaces: Special issue," *Comm. of the ACM*, 46(3), 2003.
[181] J.-G. Wang, E. Sung, and R. Venkateswarlu, "Eye gaze estimation from a single image of one eye," *ICCV,* pp. 136-143, 2003.
[182] J.J.L. Wang and S. Singh, "Video analysis of human dynamics – A survey," *Real-Time Imaging,* 9(5):321-346, 2003.
[183] L. Wang, W. Hu and T. Tan "Recent developments in human motion analysis," *Pattern Recognition*, 36 (2003) 585-601
[184] 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.
[185] M. Weiser, "Some computer science issues in ubiquitous computing," *Comm. of the ACM,* 36(7):74-83, 1993.
[186] Y. Wu and T.S. Huang. "Vision-based gesture recognition: A review," *3rd Gesture Workshop*, 1999.
[187] 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.
[188] Y. Wu, G. Hua, and T. Yu, "Tracking articulated body by dynamic Markov network," *ICCV*, pp.1094-1101, 2003.
[189] Y. Yacoob and L. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. on Pattern Analysis and Machine Intell.,* 18(6):636-642, 1996.
[190] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. on PAMI*, 24(1):34–58, 2002.
[191] H. Yoshikawa, "Modeling humans in human-computer interaction," in *Human-computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications,* pp. 118-146, 2002.
[192] Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image, and thermal image of face," *Int. Workshop on Robot-human Interaction,* 2000.
[193] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensorimotor experience," *ACM Trans. on Applied Perception*, 2004.
[194] Q. Yuan, S. Sclaroff, and V. Athitsos, "Automatic 2D hand tracking in video sequences," *IEEE Workshop on Applications of Computer Vision*, 2005.
[195] 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.
[196] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, 12:399–458, 2003.