Nicu Sebe
Michael S. Lew
Thomas S. Huang (Eds.)

Computer Vision in Human-Computer Interaction

LNCS 3766

ICCV 2005 Workshop on HCI
Beijing, China, October 2005
Proceedings

Springer

Nicu Sebe Michael S. Lew
Thomas S. Huang (Eds.)

Computer Vision in Human-Computer Interaction

ICCV 2005 Workshop on HCI
Beijing, China, October 21, 2005
Proceedings

Springer

Volume Editors

Nicu Sebe
University of Amsterdam, Faculty of Science
The Netherlands
E-mail: nicu@science.uva.nl

Michael S. Lew
Leiden University, LIACS Media Lab
Niels Bohrweg 1, 2333 CA Leiden
The Netherlands
E-mail: mlew@liacs.nl

Thomas S. Huang
University of Illinois at Urbana-Champaign
Beckman Institute
Urbana, IL 61801, USA
E-mail: huang@ifp.uiuc.edu

Library of Congress Control Number: 2005934480

CR Subject Classification (1998): I.4, I.5, I.3, H.5.2-3

ISSN 0302-9743
ISBN-10 3-540-29620-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-29620-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11573425 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3766

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

Human-Computer Interaction (HCI) lies at the crossroads of many scientific areas including artificial intelligence, computer vision, face recognition, motion tracking, etc. In order for HCI systems to interact seamlessly with people, they need to understand their environment through vision and auditory input. Moreover, HCI systems should learn how to adaptively respond depending on the situation.

The goal of this workshop was to bring together researchers from the field of computer vision whose work is related to human-computer interaction. The selected articles for this workshop address a wide range of theoretical and application issues in human-computer interaction ranging from human-robot interaction, gesture recognition, and body tracking, to facial features analysis and human-computer interaction systems.

This year 74 papers from 18 countries were submitted and 22 were accepted for presentation at the workshop after being reviewed by at least 3 members of the Program Committee. We had therefore a very competitive acceptance rate of less than 30% and as a consequence we had a very-high-quality workshop.

We would like to thank all members of the Program Committee for their help in ensuring the quality of the papers accepted for publication. We are grateful to Dr. Jian Wang for giving the keynote address.

In addition, we wish to thank the organizers of the 10th IEEE International Conference on Computer Vision and our sponsors, University of Amsterdam, Leiden Institute of Advanced Computer Science, and the University of Illinois at Urbana-Champaign, for support in setting up our workshop.

August 20, 2005

Nicu Sebe
Michael S. Lew
Thomas S. Huang

IEEE International Workshop on Human-Computer Interaction 2005 (HCI 2005) Organization

Organizing Committee

Nicu Sebe	University of Amsterdam, The Netherlands
Michael S. Lew	Leiden University, The Netherlands
Thomas S. Huang	University of Illinois at Urbana-Champaign, USA

Program Committee

Kiyo Aizawa	University of Tokyo, Japan
Alberto Del Bimbo	University of Florence, Italy
Nozha Boujemaa	INRIA Rocquencourt, France
Kim Boyer	Ohio State University, USA
Edward Chang	University of California, Santa Barbara, USA
Ira Cohen	HP Research Labs, USA
Jeffrey Cohn	University of Pittsburgh, USA
James Crowley	INRIA Rhônes-Alpes, France
Jonathan Foote	FXPAL, USA
Theo Gevers	University of Amsterdam, The Netherlands
Alan Hanjalic	TU Delft, The Netherlands
Thomas S. Huang	University of Illinois at Urbana-Champaign, USA
Alejandro Jaimes	FujiXerox, Japan
Brigitte Kerherve	University of Quebec, Canada
Michael S. Lew	Leiden University, The Netherlands
Frank Nack	CWI, The Netherlands
Jan Nesvadba	Philips Research, The Netherlands
Mark Nixon	University of Southampton, UK
Maja Pantic	TU Delft, The Netherlands
Ioannis Patras	University of York, UK
Vladimir Pavlovic	Rutgers University, USA
Alex Pentland	Massachusetts Institute of Technology, USA
Stan Sclaroff	Boston University, USA
Nicu Sebe	University of Amsterdam, The Netherlands
Qi Tian	University of Texas at San Antonio, USA
Guangyou Xu	Tsinghua University, China
Ming-Hsuan Yang	Honda Research Labs, USA
HongJiang Zhang	Microsoft Research Asia, China
Xiang (Sean) Zhou	Siemens Research, USA

Sponsors

Faculty of Science, University of Amsterdam
Leiden Institute of Advanced Computer Science, Leiden University
Beckman Institute, University of Illinois at Urbana-Champaign

Vol. 3704: M. De Gregorio, V. Di Maio, M. Frucci, C. Musio (Eds.), Brain, Vision, and Artificial Intelligence. XV, 556 pages. 2005.

Vol. 3703: F. Fages, S. Soliman (Eds.), Principles and Practice of Semantic Web Reasoning. VIII, 163 pages. 2005.

Vol. 3702: B. Beckert (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. XIII, 343 pages. 2005. (Subseries LNAI).

Vol. 3701: M. Coppo, E. Lodi, G. M. Pinna (Eds.), Theoretical Computer Science. XI, 411 pages. 2005.

Vol. 3699: C.S. Calude, M.J. Dinneen, G. Păun, M. J. Pérez-Jiménez, G. Rozenberg (Eds.), Unconventional Computation. XI, 267 pages. 2005.

Vol. 3698: U. Furbach (Ed.), KI 2005: Advances in Artificial Intelligence. XIII, 409 pages. 2005. (Subseries LNAI).

Vol. 3697: W. Duch, J. Kacprzyk, E. Oja, S. Zadrożny (Eds.), Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005, Part II. XXXII, 1045 pages. 2005.

Vol. 3696: W. Duch, J. Kacprzyk, E. Oja, S. Zadrożny (Eds.), Artificial Neural Networks: Biological Inspirations – ICANN 2005, Part I. XXXI, 703 pages. 2005.

Vol. 3695: M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005. (Subseries LNBI).

Vol. 3694: M. Malek, E. Nett, N. Suri (Eds.), Service Availability. VIII, 213 pages. 2005.

Vol. 3693: A.G. Cohn, D.M. Mark (Eds.), Spatial Information Theory. XII, 493 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005. (Subseries LNBI).

Vol. 3691: A. Gagalowicz, W. Philips (Eds.), Computer Analysis of Images and Patterns. XIX, 865 pages. 2005.

Vol. 3690: M. Pěchouček, P. Petta, L.Z. Varga (Eds.), Multi-Agent Systems and Applications IV. XVII, 667 pages. 2005. (Subseries LNAI).

Vol. 3689: G.G. Lee, A. Yamada, H. Meng, S.H. Myaeng (Eds.), Information Retrieval Technology. XVII, 735 pages. 2005.

Vol. 3688: R. Winther, B.A. Gran, G. Dahll (Eds.), Computer Safety, Reliability, and Security. XI, 405 pages. 2005.

Vol. 3687: S. Singh, M. Singh, C. Apte, P. Perner (Eds.), Pattern Recognition and Image Analysis, Part II. XXV, 809 pages. 2005.

Vol. 3686: S. Singh, M. Singh, C. Apte, P. Perner (Eds.), Pattern Recognition and Data Mining, Part I. XXVI, 689 pages. 2005.

Vol. 3685: V. Gorodetsky, I. Kotenko, V. Skormin (Eds.), Computer Network Security. XIV, 480 pages. 2005.

Vol. 3684: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part IV. LXXIX, 933 pages. 2005. (Subseries LNAI).

Vol. 3683: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. LXXX, 1397 pages. 2005. (Subseries LNAI).

Vol. 3682: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. LXXIX, 1371 pages. 2005. (Subseries LNAI).

Vol. 3681: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXXX, 1319 pages. 2005. (Subseries LNAI).

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005. (Subseries LNBI).

Vol. 3679: S.d.C. di Vimercati, P. Syverson, D. Gollmann (Eds.), Computer Security – ESORICS 2005. XI, 509 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005. (Subseries LNBI).

Vol. 3677: J. Dittmann, S. Katzenbeisser, A. Uhl (Eds.), Communications and Multimedia Security. XIII, 360 pages. 2005.

Vol. 3676: R. Glück, M. Lowry (Eds.), Generative Programming and Component Engineering. XI, 448 pages. 2005.

Vol. 3675: Y. Luo (Ed.), Cooperative Design, Visualization, and Engineering. XI, 264 pages. 2005.

Vol. 3674: W. Jonker, M. Petković (Eds.), Secure Data Management. X, 241 pages. 2005.

Vol. 3673: S. Bandini, S. Manzoni (Eds.), AI*IA 2005: Advances in Artificial Intelligence. XIV, 614 pages. 2005. (Subseries LNAI).

Vol. 3672: C. Hankin, I. Siveroni (Eds.), Static Analysis. X, 369 pages. 2005.

Vol. 3671: S. Bressan, S. Ceri, E. Hunt, Z.G. Ives, Z. Bellahsène, M. Rys, R. Unland (Eds.), Database and XML Technologies. X, 239 pages. 2005.

Vol. 3670: M. Bravetti, L. Kloul, G. Zavattaro (Eds.), Formal Techniques for Computer Systems and Business Processes. XIII, 349 pages. 2005.

Vol. 3669: G.S. Brodal, S. Leonardi (Eds.), Algorithms – ESA 2005. XVIII, 901 pages. 2005.

Vol. 3668: M. Gabbrielli, G. Gupta (Eds.), Logic Programming. XIV, 454 pages. 2005.

Vol. 3666: B.D. Martino, D. Kranzlmüller, J. Dongarra (Eds.), Recent Advances in Parallel Virtual Machine and Message Passing Interface. XVII, 546 pages. 2005.

Vol. 3665: K. S. Candan, A. Celentano (Eds.), Advances in Multimedia Information Systems. X, 221 pages. 2005.

Vol. 3664: C. Türker, M. Agosti, H.-J. Schek (Eds.), Peer-to-Peer, Grid, and Service-Orientation in Digital Library Architectures. X, 261 pages. 2005.

Vol. 3663: W.G. Kropatsch, R. Sablatnig, A. Hanbury (Eds.), Pattern Recognition. XIV, 512 pages. 2005.

Vol. 3662: C. Baral, G. Greco, N. Leone, G. Terracina (Eds.), Logic Programming and Nonmonotonic Reasoning. XIII, 454 pages. 2005. (Subseries LNAI).

Vol. 3661: T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, T. Rist (Eds.), Intelligent Virtual Agents. XIII, 506 pages. 2005. (Subseries LNAI).

Table of Contents

Multimodal Human Computer Interaction: A Survey 1
Alejandro Jaimes and Nicu Sebe

Tracking

Tracking Body Parts of Multiple People
for Multi-person Multimodal Interface 16
Sébastien Carbini, Jean-Emmanuel Viallet, Olivier Bernier, and Bénédicte Bascle

Articulated Body Tracking Using Dynamic Belief Propagation 26
Tony X. Han and Thomas S. Huang

Recover Human Pose from Monocular Image
Under Weak Perspective Projection 36
Minglei Tong, Yuncai Liu, and Thomas S. Huang

A Joint System for Person Tracking and Face Detection..... 47
Zhenqiu Zhang, Gerasimos Potamianos, Andrew Senior, Stephen Chu, and Thomas S. Huang

Interfacing

Perceptive User Interface, a Generic Approach 60
Michael Van den Bergh, Ward Servaes, Geert Caenen, Stefaan De Roeck, and Luc Van Gool

A Vision Based Game Control Method 70
Peng Lu, Yufeng Chen, Xiangyong Zeng, and Yangsheng Wang

Mobile Camera-Based User Interaction 79
Antonio Haro, Koichi Mori, Tolga Capin, and Stephen Wilkinson

Event Detection

Fast Head Tilt Detection for Human-Computer Interaction..... 90
Benjamin N. Waber, John J. Magee, and Margrit Betke

Attention Monitoring Based on Temporal Signal-Behavior Structures 100
Akira Utsumi, Shinjiro Kawato, and Shinji Abe

Action Recognition with Global Features..... 110
Arash Mokhber, Catherine Achard, Xingtai Qu, and Maurice Milgram

3D Human Action Recognition
Using Spatio-temporal Motion Templates 120
Fengjun Lv, Ramakant Nevatia, and Mun Wai Lee

Augmented Reality

Interactive Point-and-Click Segmentation for Object Removal
in Digital Images 131
Frank Nielsen and Richard Nock

Information Layout and Interaction Techniques
on an Augmented Round Table 141
Shintaro Kajiwara, Hideki Koike, Kentaro Fukuchi, Kenji Oka, and Yoichi Sato

On-Line Novel View Synthesis Capable
of Handling Multiple Moving Objects 150
Indra Geys and Luc Van Gool

Hand and Gesture

Resolving Hand over Face Occlusion 160
Paul Smith, Niels da Vitoria Lobo, and Mubarak Shah

Real-Time Adaptive Hand Motion Recognition
Using a Sparse Bayesian Classifier 170
Shu-Fai Wong and Roberto Cipolla

Topographic Feature Mapping for Head Pose Estimation
with Application to Facial Gesture Interfaces 180
Bisser Raytchev, Ikushi Yoda, and Katsuhiko Sakaue

Accurate and Efficient Gesture Spotting
via Pruning and Subgesture Reasoning 189
Jonathan Alon, Vassilis Athitsos, and Stan Sclaroff

Applications

A Study of Detecting Social Interaction with Sensors
in a Nursing Home Environment 199
Datong Chen, Jie Yang, and Howard Wactlar

HMM Based Falling Person Detection Using Both Audio and Video 211
B. Uğur Töreyin, Yiğithan Dedeoğlu, and A. Enis Çetin

Appearance Manifold of Facial Expression 221
Caifeng Shan, Shaogang Gong, and Peter W. McOwan

Author Index 231

Multimodal Human Computer Interaction: A Survey

Alejandro Jaimes[1] and Nicu Sebe[2]

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

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

1 Introduction

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

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

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

In this paper we highlight the main vision problems that in our view should be solved for successful MMHCI applications, and give an overview of the research areas we consider essential for MMHCI. We group vision techniques according to the human body (Figure 1). Large-scale body movement, gesture (e.g., hands), and gaze analysis are used for tasks such as emotion recognition in affective interaction, and for a variety of applications. We discuss affective computer interaction, issues in multi-modal fusion, modeling, and data collection, and a variety of emerging MMHCI

N. Sebe, M.S. Lew, and T.S. Huang (Eds.): HCI/ICCV 2005, LNCS 3766, pp. 1–15, 2005.

applications. Since MMHCI is a very dynamic and broad research area we do not intend to present a complete survey. The main contribution of this paper, therefore, is to consolidate some of the main issues and approaches, and to highlight some of the techniques and applications developed recently within the context of MMHCI.

1.1 Related Surveys

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

2 Overview of Multimodal Interaction

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

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

As depicted in Figure 1, we place input modalities in two major groups: based on human senses (*vision*, *audio*, *haptic*, *olfactory* and *touch*), and others (mouse, key-

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

board, etc.). The visual modality includes any form of interaction that can be interpreted visually, and the audio modality any form that is audible (including multilanguage input). We only discuss vision in detail, but as many new applications show (see Section 6), other modalities have gained importance for interaction (e.g., haptic [4]).

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

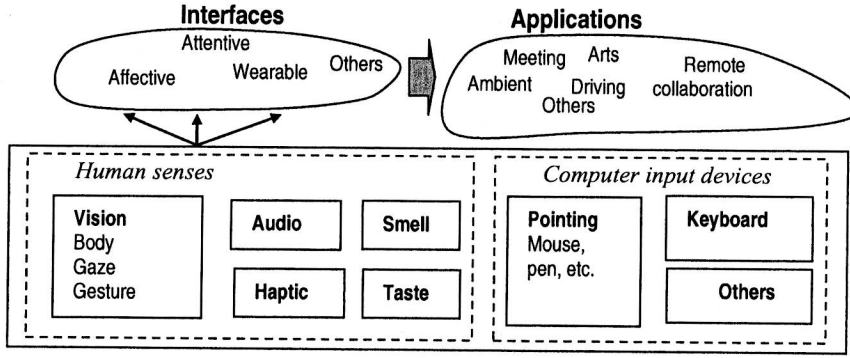

Fig. 1. Overview of multimodal interaction using a human-centered approach

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

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

3 Core Vision Techniques

We classify vision techniques for MMHCI using a human-centered approach and divide them according to how humans may interact with the system: (1) large-scale

body movements, (2) gestures, and (3) gaze. We make a distinction between *command* (actions can be used to explicitly execute commands: select menus, etc.) and *non-command* interfaces (actions or events used to indirectly tune the system to the user's needs) [45][7].

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

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

3.1 Large-Scale Body Movements

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

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

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

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

distinction between body movements, events, and behaviors, within a rule-based system framework.

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

3.2 Gesture Recognition

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

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

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

3.3 Gaze Detection

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

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

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

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

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

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

4 Affective Human-Computer Interaction

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

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

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

4.1 Facial Expression Recognition

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

4.2 Emotion in Audio

The vocal aspect of a communicative message carries various kinds of information. If we disregard the manner in which a message is spoken and consider only the textual content, we are likely to miss the important aspects of the utterance and we might