Integrating Perceptual and Cognitive Modeling for Adaptive and Intelligent Human-Computer Interaction

Z. Duric[1], W. Gray[2], R. Heishman[1], F. Li[1],
A. Rosenfeld[3], C. Schunn[4], and H. Wechsler[1]

[1]Department of Computer Science
George Mason University
Fairfax, VA 22030-4444

[1]Department of Psychology
George Mason University
Fairfax, VA 22030-4444

[3]Center for Automation Research
University of Maryland
College Park, MD 20742-3275

[4] Department of Psychology
University of Pittsburgh
Pittsburgh, PA 15260

{zduric,rheishman,fli,wechsler}@cs.gmu.edu
gray@gmu.edu,ar@cfar.umd.edu,schunn+@pitt.edu

Abstract

This paper describes technology and tools for Intelligent HCI (IHCI) where human cognitive, perceptual, motor, and affective factors are modeled and used to adapt the H - C interface. Intelligent HCI emphasizes that human behavior encompasses both apparent human behavior and the hidden mental state behind behavioral performance. IHCI expands on the interpretation of human activities, known as W4 (what, where, when, who). While W4 addresses only the apparent perceptual aspect of human behavior, the W5+ technology for IHCI described in this paper addresses also the why and how questions whose solution requires recognizing and processing around specific cognitive states. IHCI integrates parsing and interpretation of nonverbal information with a computational cognitive model of the user, which, in turn, feeds into processes that adapt the interface to enhance operator performance and provide for rational decision-making. The technology proposed is based on a general four-stage, interactive framework, which moves from parsing the raw sensory-motor input, to interpreting the user's motions and emotions, to building an understanding of the user's current cognitive state. It then diagnoses various problems in the situation and adapts the interface appropriately. The interactive component of the system improves processing at each stage. Examples of perceptual, behavioral and cognitive tools are described throughout the paper. Adaptive and Intelligent HCI are important for novel applications of computing including ubiquitous and human-centered computing.

1. Introduction

Imagine a computer interface that could predict and diagnose whether the user was fatigued, confused, frustrated or momentarily distracted by gathering a variety of *nonverbal* information (e.g., pupillary responses, eye fixations, facial expressions, upper-body posture, arm movements, keystroke force). Further imagine that the interface could adapt itself—simplify, highlight, or tutor—to improve the human-computer interaction using these diagnoses and predictions. Nonverbal information facilitates a special type of communication where the goal is to probe the inner (cognitive and affective) states of the mind before any verbal communication has been contemplated and/or expressed. This paper addresses the technology and tools required to develop novel computer interfaces suitable to handle such nonverbal information.

Assume now a private, single-engine plane wanders into a commercial flight sector. The air traffic controller does nothing. Has she noticed the plane, evaluated its flight path, and concluded that it will shortly leave the sector without posing a threat to commercial aviation? Or has the plane slipped in unnoticed and, hence, the controller has not yet considered the need to alert and reroute the five commercial flights in her sector? From the simple data that the computer gets from its operator (i.e., decisions made), it is impossible to know whether the busy controller's attention should be directed to the intruder or left alone to focus on more urgent matters. However, at the time the intruder entered the sector, the controller's upper-body was erect and tilted forward in the seat, her point-of-gaze was within 1° visual angle of the intruder, her pupils were dilated, her facial expression indicated surprise, and the force of her mouse click (on a commercial jetliner) was much less intense than normal. Imagine a computer interface that gathered such information about the operator and correctly diagnosed the operator's current cognitive state. With the above data it might decide to do nothing. With another combination of nonverbal information, it might decide to make the intruder's icon blink on and off. With a third combination, it might zoom the screen so that the intruder's icon was at the controller's point-of-gaze.

Less dramatic types of human-computer problems could also benefit by processing nonverbal information for adaptive and intelligent interfaces. For example, an operator repeatedly uses the mouse to gesture to (i.e., to point at, circle, or otherwise indicate a target) an already classified target. Is he confused, frustrated, or simply fatigued? If he is confused, the interface could be automatically simplified (since optimal display complexity is relative to the expertise of the operator), or a tutorial could be offered during the next work lull. Again, this diagnosis and remedial action could be made if the computer had access to nonverbal information about the operator. Arm movements can indicate cognitive states like surprise and fatigue in addition to being a substitute for verbal communication suitable for deaf and/or mute people, and gestures appropriate for noisy environments. Yet another example of using nonverbal information to infer the cognitive state of the user comes from pupillometry, the psychology of the pupillary response. There is general agreement that pupils dilate during increased cognitive activity and constrict (or return to some previous baseline) when the activity decreases (e.g., when a particular problem has been solved and relaxation sets in). There is also evidence to support the assertion that the constant motion of the pupil (referred to as "pupillary unrest" or "hippus") is more accentuated under conditions of fatigue or drowsiness. Knapp and Hall [37] provide further evidence regarding nonverbal information in the context of expressing emotions and its location. In particular, they note: "Rarely is the eye area tested separately from the entire face in judging emotions. Sometimes, however, a glance at the [brow and] eye area may provide us with a good deal of information about the emotion being expressed. For example, if we see tears we certainly conclude that the person is emotionally aroused, though without other cues we may not know whether the tears reflect grief, physical pain, joy, anger, or some complex blend of emotions. And, downcast or averted eyes are often associated with feelings of sadness, shame, or embarrassment."

Nonverbal information as a new communication medium is most suitable for behavior interpretation. For example, the existing work on facial processing can be now extended to task-relevant expressions rather the typical arbitrary set of expressions identified in face processing research. Moreover, the technology and tools proposed will have the added benefit of developing a framework by which one can improve our predictions of the conse-

quences of various interface decisions on behavior—an important goal in the science of HCI. In particular, this paper emphasizes that human behavior encompasses both apparent performance and the hidden mental state behind performance. Towards that end we suggest an integrated system approach that can measure the corresponding perceptual and cognitive states of the user, and then can adapt in an intelligent fashion the HCI for enhanced human performance and satisfaction. The outline of the paper is as follows. Section 2provides the conceptual and intellectual framework needed to address issues related to adaptive and intelligent nonverbal interfaces. Section 3 describes recent research related to the interpretation of human activities, also known as W4 (*what, where, when, who*). The shortcomings of W4, since it is dealing only with the apparent perceptual aspect, are discussed and provide the motivation for our novel proposed methodology, W5+ (*what, where, when, who, why, how*); the *why* and *how* questions are directly related to recognize and process around specific cognitive states. Section 4 describes in detail the W5+ methodology and motivates the choices made. In addition to migration from W4 to W5+, emphasis is placed on the fact that performance needs to be monitored in terms of both apparent (external) and internal behavior. We also contrast our framework with the more traditional use of Bayesian networks (Pearl, 1998 and Hurwitz, 1996) and Dynamic Belief Networks (DBNs), which ”meter” only the external behavior. Section 5 describes the tools required to implement the perceptual processing module, focusing on the interpretation of lower arm movements, facial expressions, pupil size, and eye-gaze location. Section 6 describes the technology required for behavioral processing, focusing on the novel area of mouse gesture interpretation. Section 7 overviews the components of embodied models of cognition and how they can be extended to include affect. Section 8 elaborates on how user interfaces can be adapted dynamically using the embodied model of cognition, and what additional issues need to be considered. We conclude the paper in Section 9 with a summary of the novel W5+ technology and recommendations for further research and tool development.

2. Background

HCI has developed mostly along two competing methodologies [51]: direct manipulation and intelligent agents (also known as delegation). These approaches can be contrasted as the computer sitting passively waiting for input from the human versus the computer taking over from the human. Another dimension for HCI is that of affective computing [46]. Affective computing is concerned with the means to recognize ”emotional intelligence”. Whereas emotional intelligence includes both bodily (physical) and mental (cognitive) events, affective computing presently focuses mainly on the apparent characteristic of verbal and non-verbal communication, as most HCI studies elicit emotions in relatively simple settings [47]. Specifically, recognition of affective states focuses on their physical form (e.g., blinking or face distortions underlying human emotions) rather than implicit behavior and function (their impact on how the user employs the interface). In contrast to the established paradigms of direct manipulation and intelligent agents, Intelligent Human Computer Interaction (IHCI) uses computer intelligence to increase the bandwidth through which humans interact with computers [33, 44]. Nonverbal information such as facial expressions, posture, point-of-gaze, and the speed or force with which a mouse is moved or clicked can be parsed and interpreted by the computer to iteratively construct and refine a model of the human’s cognitive and affective states. The availability of such users’ models can be then used in an adaptive fashion to enhance human-computer interactions and to make them appear intelligent, i.e., causal, to an outside observer.

It is not only computer technology that needs to change to make such novel interfaces a reality. People have to change too and adapt to the interface the computer presents them with. In the end both people and the computer have to understand each other’s intentions and/or motivations, provide feedback as necessary to each other, and eventually adapt to each other. W5+ systems are examples of novel intelligent interfaces and they make the transition from HCI to Intelligent HCI where people and human can augment each other capabilities and display characteristics of team behavior. The methodology we propose for IHCI (see Fig. 1) integrates parsing and interpretation of nonverbal information with a computational cognitive model of the user that, in turn, feeds into processes that adapt the interface to enhance operator performance and provide for rational decision-making. Adaptive and intel-

ligent HCI combines advanced work in perceptual recognition, machine learning, affective computing, as well as the computational modeling of embodied cognition. Our methodology is based on a general four-stage, interactive framework. The system moves from parsing the raw sensory-motor input, to interpreting the user's motions and emotions, to building an understanding of the user's current cognitive state. It then diagnoses various problems in the situation and adapts the interface appropriately. The interactive component of the system improves processing at each stage. For example, knowledge of the user's current cognitive state helps predict changes in eye and head location, which in turn improves image parsing. We expect that our approach will have potential benefits for a broad class of human-computer interactions. Moreover, our integrated methodology will also advance many areas of basic research (e.g., computer vision and facial processing, perception, cognition, human learning and adaptation). We view our proposal as the necessary steps in developing intelligent HCI systems where human cognitive, perceptual, motor, and affective factors are (fully) modeled and used to adapt the interface.

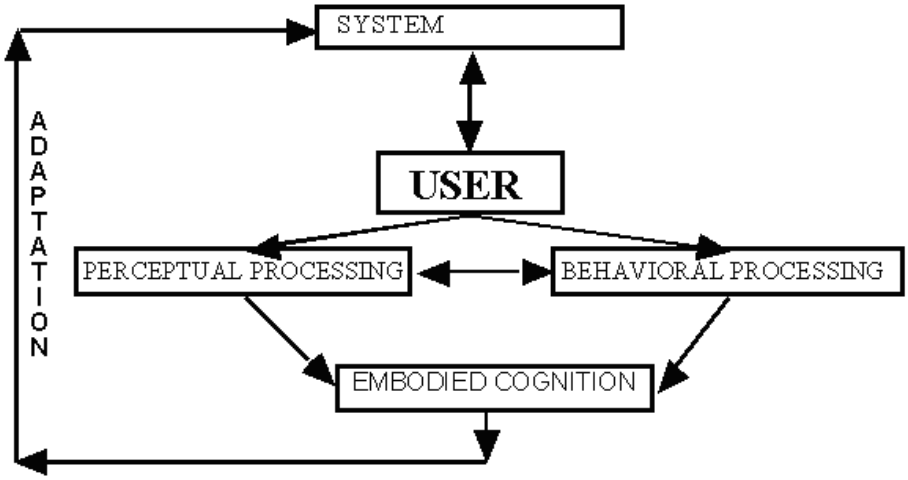

Figure 1: System architecture for adaptive and intelligent HCI.

Intelligent HCI (IHCI) promotes also human activity and creativity. As part of emerging intelligent synthesis environments (ISE), intelligent HCI supports human-centered and immersive computing, the infrastructure for distributed collaboration, rapid synthesis and simulation tools, and life-cycle system integration and validation. Intelligent HCI combines the (computational) ability to perceive mixed affordance (input) patterns, reasoning and abstraction, learning and adaptation, and finally, communication, language and visualization. These concepts echo those of Kant, for whom perception without abstraction is blind, while abstraction without perception is empty; and of Confucius, for whom learning without thought is useless, and thought without learning is dangerous.

Ubiquitous or pervasive computing, a new metaphor for computing, provides users constant access to information and computation in the form of mobile and wireless computing devices—Personal Digital Assistants (PDAs), such as cell phones, wearable computers and appliances—that are endowed with intuitive user interfaces. Ubiquitous computing maintains computing at its core but removes it from our central focus. Computing is embedded into a surrounding, but almost invisible and friendly world, in order to facilitate collaborative work and the creation and dissemination of knowledge. Ubiquitous computing is more than virtual reality, which puts people inside a computer-generated world, and is more than personal digital assistants. Ubiquitous computing involves explicit representations of oneself and other humans, possibly using avatars, and representation of cognitive, affective, social, and organizational aspects of the human behind the avatar (e.g., natural and expressive faces and gestures, representing and reasoning about others' places in organizational systems, and social relationships). The boundaries between the 'real world', augmented reality, and virtual environments are blurred to create a mixed reality.

Unlike virtual reality, mixed reality seeks to enhance the real environment, not to replace it. That is, while an interface agent or PDA will alert a pilot of impeding collision, ubiquitous computing will display for the pilot airspace information that provides continuous spatial awareness of surrounding objects. The mixed reality metaphor avoids making systems appear too human-like in cases where they have very limited intelligence and are brittle in their interaction.

Intelligent HCI makes ubiquitous computing possible by continuously adapting the interface medium to meet specific users needs and demands. The emergence of human-centered interaction with intelligent systems buttresses the utilization of both verbal and nonverbal communication to create a richer, more versatile and effective environment for human activity. Human-centered design is problem-driven, activity- centered, and context-bound, and employs computing technology as a tool for the user, not as a substitute. Thus, the emphasis is on supporting human activity using adaptive and intelligent interfaces rather than on building (fully) autonomous systems that mimic humans. One approach to a human-centered use of intelligent system technology seeks to make such systems "team players" in the context of human activity, where people and computer technology interact to achieve a common purpose. Another possible approach focuses on building effective computational tools for modeling, interpreting, fusing and analyzing cognitive and social interactions such as speech, vision, gesture, haptic inputs, and/or affective state expressed using body language. The goal for IHCI is to expand on the human perceptual, intellectual, and motor activities.

People tend to misjudge the bounds of systems' capability, ranging from over reliance on system performance—in cases where it is inappropriate to do so—and to loss of trust and lack of acceptance in situations where the system performs well. Ubiquitous computing seeks to derive benefits from a truly complementary relationship with their human partner, forcing computing to thrive in an integrated (virtual and physical) world with people; and to become predictable, comprehensible, and informative for their human partner. To forge a trusted partnership and expand human intellect and abilities, computing has to become compliant with our demands, communicative regarding their processes, and cooperative with our endeavors.

Ubiquitous computing emphasizes distributed affordances instead of focused expertise. People are most effective when they are fully engaged, mind and body, in the world. What becomes apparent is situated (grounded) computing, leading to practical intelligence facilitated by shared context acquired at the intersection of perception (senses, affective and physiological), language and communication, thought and reason, and action (purposive and functional). Shared context leads to recoordination of human behavior and subsequent reconceptualization. We will soon reach the point when computing power and storage are cheap commodities relative to the challenge of making them readily and effectively available to everyone, everywhere. The vision of ubiquitous computing can become reality only with the emergence of several components, and intelligent HCI is one of them. Wearable devices need to supply correct and just-in-time information and reduce the wearer's cognitive load. Such devices combine communication, computation, and context sensitivity, while also supporting increased collaboration. Users must be able to leave their desktops and continue their daily tasks, possibly engaging in hands-free tasks, while still remaining connected to computing and communication resources. With wearable computers, interaction with the real world is the primary task, making current WIMP (windows-icons-menu-pointers) interfaces obsolete. An emerging augmented system that interfaces people and computing devices, promotes mutual understanding—including their background, goals, motivations and plans—and optimal sharing of the computational load.

3. Review of W4: where, what, when, why

We briefly review here recent research related to the interpretation of human activities (W4). As it will become apparent from our discussion, W4 only deals with the apparent perceptual aspect and makes no reference to the cognitive element responsible for that aspect. Extensive work on analyzing images of humans' activities began in the 1990s. Many references are listed in [94]. A large number of papers on face and gesture recognition were

presented in the four international conferences on the subject [65, 66, 67, 68]. A very good review of early work on motion understanding approaches and applications was prepared by Cedras and Shah [18]. A review of research papers on hand gesture recognition for human-computer interaction was done by Pavlovic et al. [44], and a broader review of research papers on visual analysis of human motion was done by Gavrila [63]. A review of papers on nonrigid motion analysis, in particular on articulated and elastic motion, has been done by Aggarwal et al. [1]. A comprehensive review of various methods for computer vision-based capture of human motions was recently done by Moeslund and Granum [89]. In the remainder of this section we review recent work on motion analysis and understanding because it is the primary force behind human activities. Three main criteria can be used to classify research on human motion analysis. First, the research can be classified in terms of the tasks that it focuses on: detection, tracking, or recognition. Second, it can be classified in terms of the models used to represent objects and humans. Third, it can be classified in terms of the control mechanisms used.

Detection of humans in static or video images has mostly been addressed through background subtraction and matching. A background subtraction method that uses colors and edges was described in [77]. Some authors have used background subtraction as a part of a system combining detection, body labeling, and tracking of humans [57, 62, 71, 103]. In some cases cues such as skin color have been used to detect humans in images [76]. Other authors have used motion from single or multiple cameras to detect, label, and track humans or their body parts in video images [81, 83, 93, 96, 98, 103, 102]. Some authors have approached this problem as one of matching. Humans or their parts have been detected and tracked as configurations of points such as light displays, markers, and image features [100]; as configurations of edges [64, 84, 85, 86, 90, 92, 97]; and as collections of particularly shaped strips [75], cylinders or superquadrics [58, 62, 96, 99, 102]. For tracking some authors have focused on using motions of image points and edges. Human models have been initialized by hand in the first frame of each sequence [15, 53]. Some authors have considered the problem of action/activity/gesture recognition for humans using shape and/or motion information [11, 13, 56, 69, 70, 60, 61, 79, 80, 95, 101, 104, 105]. Dynamic recognition, most appropriate for interpreting video sequences, is done using recursive and neural networks, deformable templates, spatio-temporal templates [43], and graphical models [13] because they offer dynamic time warping and a clear Bayesian semantics for both individual (HMM) and interacting or coupled (CHMMs) generative processes [91]. Finally, some authors have implemented systems that combine detection, tracking, and recognition [2, 16, 56, 71, 73, 87, 88, 93].

A second set of criteria that can be used for classifying research on human motions is based on how to model humans. Humans have been modeled as elongated, blob-like shapes either implicitly [56, 71, 73, 87, 88, 93] or explicitly [64, 90, 92]. Deformable models have been utilized for body part (hands) and facial feature tracking [13, 97, 104]. Some authors have modeled humans as articulated stick figures [2, 84, 85, 86, 95, 100]; this approach has been particularly effective for moving light display analysis. Finally, humans have been modeled as articulated objects, where parts correspond to blobs [103], strips [15, 75, 81], tapered superquadrics [62, 83], or cylinders [58, 96, 98, 99, 102].

A third set of criteria that can be used for classifying research on human motions is based on the mechanisms used to control search in detection, tracking and recognition. Kalman filtering has been used frequently; examples include [53, 83, 98, 96, 102]. More recently, Bayesian inference has been used [58, 92, 97, 99]; these methods are also known as Condensation. Other strategies that have been used include search algorithms such as best-first [62] and/or "winner take all" [53, 88, 105].

Bobick [55] recently proposed a taxonomy of movement, activity, and action. In his taxonomy, movements are primitives, requiring no contextual or sequence knowledge in order to be recognized. Activities are sequences of movements or states, where the only knowledge required to recognize them involves statistics of the sequence. According to Bobick, most of the recent work in gesture understanding falls within this category. Actions are larger-scale events which typically include interactions with the environment and causal relationships. An important distinction between these levels is the degree to which time must be explicitly represented and manipulated, ranging from simple linear scaling of speed to constraint-based reasoning about temporal intervals.

Other related work includes biomechanics and human modeling in computer graphics and movement notations in choreography. In biomechanics, researchers are interested in modeling forces and torques applied to human bodies and tissues during various physical activities [106]; these models provide tools for analyzing the relationship between movements and actions. In computer graphics, researchers are interested in producing realistic images for virtual reality applications, human factor analysis and computer animation [11, 12, 10, 72]. Formalisms for describing the motions of humans include movement notations [10, 11] such as Labanotation [74], which is mostly used for dance, and Eshkol-Wachmann notation [59, 74], which is also used for sign languages [54].

4. Adaptive and Intelligent HCI – Methodology

Our methodology is quite general and it has been outlined in Fig. 1. The main modules are perceptual processing, behavioral processing, embodied cognition, and adaptive system interface. The user is interacting with an adaptive system interface, which changes as a function of the current task state and the cognitive or mental state of the user. The nonverbal front-end includes the perceptual and behavioral processing modules, and its input consists of raw-sensory information about the user. The perceptual module processes images of the face, the eye (gaze-location and pupil size), and the upper body and analyzes their relative motion; the behavioral module processes information about actions done to the computer interface directly, a such as keystroke choices, the strength of key strokes, and mouse gestures. Both the perceptual and behavioral modules provide streams of elementary features that are then grouped, parsed, tracked, and converted eventually to *sub-symbolic, summative affective representations* of the information in each processing modality. In other words, one output of the perceptual and behavioral processing modules is a stream of affective states at each point in time. States that could be recognized include confusion, fatigue, stress, and other task relevant affective states. The quest for sub-symbolic and summative affective representations is motivated by abstraction and generalization, communication, reasoning, in particular, and the Perception-Control-Action (PCA) cycle [113]. Furthermore, "Signal and [sub-]symbol integration and transformation is an old but difficult problem. It comes about because the world surrounding us is a mixture of continuous space time functions with discontinuities. Recognition of these discontinuities in the world leads to representations of different states of the world, which in turn place demands on behavioral strategies. Similarly, agent's (biological or artificial) closed loop interactions with the world/environment can be modeled as a continuous process, where as switching between behaviors is naturally discrete. Furthermore, the tasks that are either externally given to the agents or internally self-imposed prespecify and, hence, discretize an otherwise continuous behavior. Thus, we have three sources for discretization of the agent-world behavioral space : 1. Natural space-time discontinuities of the world; 2. The model of agent-world dynamics during execution of a given task; and 3. Furthermore, in computer vision, symbols served mainly as a data reduction mechanism, while in AI the following was missing : 1. Explicit acknowledgment that the transformation from signal to symbols results in the loss of information; 2. Self-correction and updating mechanisms of the obtained symbolic information; and 3. Explicit models of the dynamic interaction between an agent and its world. Symbols not only provide nice abstractions for low-level strategies, but also allow us to move one level up the modeling hierarchy and observe the properties of the systems and their interactions between each other and their environment at a more macroscopic level. Symbolic representation mediates reasoning about the sequential and repetitive nature of various tasks" [110]. The Adaptive and Intelligent HCI methodology proposed in this paper addresses the problems raised above using embodied cognition to connect the apparent perceptual and behavioral sub-symbolic affective representations and symbolic mental states, and in the process adaptively derive the summative sub-symbolic states from raw signals and also adapt the user/system interface for enhanced performance and human satisfaction. The task description language chosen for manipulation tasks using such an adaptive and Intelligent HCI is that of ACT-R/PM cognitive architecture (see Section 7).

These affective sub-symbols are fed into the embodied cognition module and mediate fusion and reasoning about possible cognitive states. While the sub-symbols correspond to external manifestations of affective states,

the cognitive states are hidden and not directly observable. The embodied cognition module generates hypotheses about possible task-relevant cognitive states, resolves any resulting ambiguity by drawing from contextual and temporal information, and optimally adapts the interface in order to enhance human performance. Knowledge of the user's state and the system's state are used to diagnose potential problems, and these diagnoses trigger adaptive and compensatory changes in the computer interface. While this process has been described as a linear process, in fact it is an interactive process in which information from later phases can feedback to augment processing in earlier phases. For example, knowledge of the current cognitive activities can be used to improve recognition of affective states.

The development of the Perceptual and Behavioral Processing and Embodied Cognition modules and their interactions is the key to making progress on constructing adaptive intelligent interfaces. The embodied cognition module is a novel addition to the more traditional HCI approaches as (i) it bridges between direct manipulation and intelligent agents through physical modeling of the users, and (ii) it augments emotional intelligence through cognitive modeling of user behavior. The capability of modeling the effects of the affective state on cognitive performance will have an impact on the choice of models as well as the computational techniques employed.

The experimental platform envisions continuous operation of the system over extended periods of time with a single user in a particular task environment. While it is possible to personalize the user interface and be able to detect characteristics of the users automatically, one can also assume that the system will be properly initialized for a particular user. This would comprise initialization of hardware and software by acquiring both a physical (perceptual and behavioral) and cognitive model of the user, calibration of the video cameras and eye tracker, and detection of the face and upper body. Initial localization of facial landmarks is crucial for successfully tracking features over time. For a first time user, generic information concerning the user's expertise and experience is used to initialize the cognitive model. For return users, the cognitive model from the user's last session is retrieved and instantiated in the current setting.

The raw video data for the perceptual processing module include color and intensity. Standard low-level computer vision techniques are used to extract additional features such as corners, edge information and motion feature maps. The feature extraction process is guided by the initialized model, for which the spatial locations of head, upper body, eyes and mouth have been determined. Our previous research on perceptual processing includes detecting faces and the upper body [32, 77, 87, 88], and automatic detection of facial landmarks [52]. The measurements acquired in the data acquisition (color, intensity, edge information and motion) can be considered as first order features. These features are combined together in order to acquire reliable estimates of the shape (contour) and motion fields of the eye, mouth facial regions, and /or arms. In the next immediate level of the hierarchy, lower-order parametric descriptions of the shapes and motion fields associated with smaller spatial regions corresponding to eyes and mouth are sought. The modes of these parametric descriptions accumulated over time are processed using Learned Vector Quantization (LVQ) and yield sub-symbolic indicators contributing to the assessment of the affective state. A variety of eye movements including pupils, irises, eyelids and eyebrows is captured as different modes. For example, the movement of the eyelids can reveal information about the stress level and tension of the user. At this level we also capture lower arm and hand movements; Section 5 provides additional description of this module. The last level of hierarchy uses the upper body shape and motion information. One can estimate independently the 3D pose for the head and shoulders, which can undergo independent or combined motions. Information on the two 3D poses is then abstracted using modal analysis and then fed into the embodied cognition module.

The processing of raw eye data (pupil location and size) requires additional perceptual processing, and is currently processed most effectively using special purpose hardware. The eye tracker data includes time stamped *x-y* coordinates of the point-of-gaze (POG) and the diameter of the pupil (at 60 samples or more per second). The particular state information corresponds to a spatial location where fixation occurred and the transition between the states (events) correspond to saccadic eye movements. Changes in recorded eye positions are parsed as fixations or saccades according to their magnitude and direction. Eye blinks are calculated from the raw data by

detecting consecutive samples with zero pupil dilation. An eye blink is indicated if these samples span a time of 30 to 300 milliseconds. Like fixation data, the number of eye blinks and rate of eye blinks between mouse clicks is calculated.

The *behavioral processing* module processes keystroke (choice and rate) and mouse data (clicks and movement). The keystroke data includes key choice and timing of keystrokes. The mouse data include the time stamped x-y coordinates of the pointer, the force and the frequency of mouse clicks, and the acceleration applied to mouse movements. The keystroke data is the primary method by which the cognitive model of the user is updated, through the process of model tracking (see below). Raw mouse data is collected at the same time that a raw eye data sample is collected. For the mouse, motion could be a movement from one position on the screen to another, and the dynamics would describe the applied force and the duration of the movement.

Parsing and interpreting the mouse data deserves additional notes, as they represent very novel uses of mouse data. The mouse data provides more than the obvious performance data of how fast and how accurately users make choices. We analyze the force data (how hard individuals click the mouse) and the trajectory data (how users move the mouse). The force data is divided into two dimensions: average force of a click and duration of the click. The trajectory data is treated as a form of gesture data. In other domains, we have found that people will gesture to various aspects of the screen using the mouse and these mouse gestures are indicators of preliminary cognitions [50]. For example, people sometimes circle objects, trace trajectories, or move rapidly between objects. Informally, we have seen the same behavior in the Argus domain (to be described below). To recognize the mouse gestures, we will use a technique that we developed in the context of sign language recognition [70]. There, hand gestures corresponding to American Sign Language are first located using projection analysis and then normalized in size, while recognition takes place using hybrid mixture of (connectionist and symbolic) experts consisting of Ensembles of Radial Basis Functions (ERBFs) and Decision Trees (DTs).

The mouse data will be used in two different ways. First, the sub-symbols corresponding to mouse gestures will be passed on to the embodied cognition module, with the goal of providing additional information about the aspects of the screen that are being attended. Second, the addition of sub-symbols corresponding to combined mouse gestures and force data will allow for the possibility of recognizing combinations of affective states in the user. That is, a person's face may reflect fatigue and their mouse gestures may reflect confusion. Research on hand gestures has often found that people can reflect different information about their internal cognitions in speech than they do in co-occurring gestures [22]. Similarly, real affective states are often a combination of basic states, and our hypothesis is that the components of the combinations may be externalized simultaneously but in different external forms (e.g., fatigue in mouse-movements, disgust in facial expressions).

One simple alternative approach to our own would be to try to go directly from these diagnoses of affective states to adaptations of the interface (e.g., confusion = simplify interface). However, such a simple method is not likely to work because it does not take into account the cognitive state of the individual with respect to the task being performed. How to best adapt the interface will usually depend upon what cognitive operations are currently being performed. For example, simplifying an interface by removing information will only work if that information is not needed in the computations currently being performed by the individual. Moreover, affective states are often directed at particular aspects of the current task/interface. For example, a particular object on the screen or aspect of an interface is often a source of confusion, and it is better to clarify or simplify the offending object/aspect than to simplify random aspects of the interface or the entire interface, which would cause more confusion. The embodied cognition module uses the process of model tracing to understand the user's behavior, thereby making intelligent interface adaptation possible. In model tracing, a cognitive model is built that is capable of solving the human tasks in the same way as it is solved by the humans. The model is then aligned with the task and behavioral choice data (i.e., what the state of the world is and what the human chose to do) such that one can see which internal cognitive steps the human must have taken in order to produce the observed behavioral actions. Towards that end, the embodied cognition model also uses the affective sub-symbols and their degree of belief, derived earlier by the perceptual and behavioral processing modules. The embodied cognition module is described

further in Section 7.

The *adaptive system interface* module adapts the interface to the current needs of the human participant. Different affective and cognitive diagnoses include confusion, fatigue, stress, momentary lapses of attention, and misunderstanding of procedures. Different adaptations include simplifying the interface, highlighting critical information, and tutoring on selected misunderstandings. For instance, in one of the examples described earlier on, if the force of the controller's mouse click and parsing of facial expressions concur in suggesting that the participant's visual attention is totally consumed by a commercial airliner, the system will intervene to alert the controller to the intruder's presence. Similarly, if later in her work shift, the controller's facial expressions and a wandering POG indicate a waning of attention, and the cognitive model interprets this as resulting from a decrease in cognitive resources (due to fatigue) then steps may be taken to off-load parts of the tasks, to increase the salience of the most safety-critical components, or to relieve the controller. The types of interface adaptations that one can consider include: 1) addition and deletion of task details; 2) addition and deletion of help/feedback windows; 3) changing the formatting/organization of information; and 4) addition and removal of automation of simple subtasks. Further details will be described in Section 8.

5. Perceptual Processing

We describe here tools for perceptual processing, including lower arm movements, facial data processing, eye-gaze tracking, and mouse gestures. Additional tools are possible, including upper body posture, head and shoulders.

5.1. Interpretation of Lower Arm Movements

We describe next a method for detection, tracking and interpretation of lower arm hand movements from color video sequences. This method is relevant to parsing the raw sensory-motor input and in particular to interpreting the user's hand motions. It corresponds to perceptual processing (see Figure 1) and its role is to transform signals to sub-symbols expressing some affective state and suitable for the embodied cognition component. The method works as follows. The moving arm is detected automatically, without manual initialization, foreground or background modeling. The dominant motion region is detected using normal flow. EM (expectation maximization), uniform sampling, and Dijkstra's shorter path algorithm are used to find the bounding contour for the moving arm. An affine motion model is fit to the arm region; residual analysis and outlier rejection are used for robust parameter estimation. The estimated parameters are used for both the prediction of the location of the moving arm and the motion representation. In analogy to linguistic analysis, the processed sensory information is made compact and suitable for interpretation using Learning Vector Quantization (LVQ), whose task is to abstract motion information. LVQ maps the affine motion parameters into a discrete set of codes $\{$A, B, G, J, C, E, D, I, H$\}$. The final transition from signals to a hierarchical and sub-symbolic representation is enabled by clustering. In particular, clustering will map the discrete set of codes generated by LVQ into more abstract "subactivities" codes up, down, circle first, and finally into specific "activities" pounding, swirling sub-symbols. Each activity or the expression of some cognitive state corresponds now to its own sequence of sub-symbols and can be properly distinguished from other activities or affective states of mind. Figures 2,3,4show some of the steps responsible for parsing the raw signal to generate a sub-symbolic description.

5.2. Processing of Facial Data

The cognitive and emotional states of a person can be correlated with visual features derived from images of the mouth, eye and eye region [37]. Figure 5 illustrates pilot data of facial expressions in a complex simulated radar classification task—expressions of the same subject in a baseline condition (9 targets to track), a low load condition (5 targets), and a high load condition (30 targets). There are detectable differences of the type that one would expect, but the differences are subtle; in particular, the mouth and eye regions display increase of tension for the difficult task.

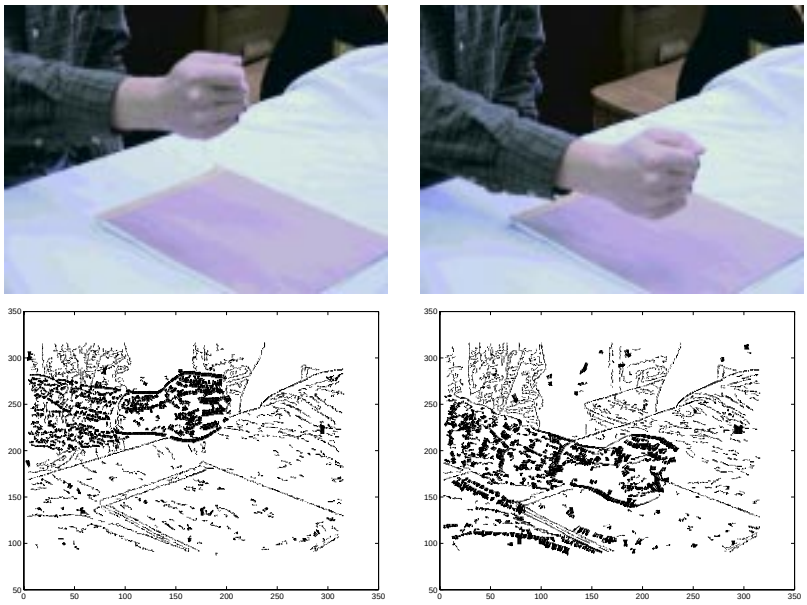

Figure 2: The first frames from two color image sequences and their corresponding normal flow. Left: the first frame from a "pounding" sequence of 400 frames. Right: the first frame from a "swirling" sequence of 100 frames. Upper row: color images. Lower row: normal flows. These images were collected using a Sony DFW-VL500 progressive color scan camera; the frame rate was thirty frames per second and the resolution was 320×240 pixels per frame.

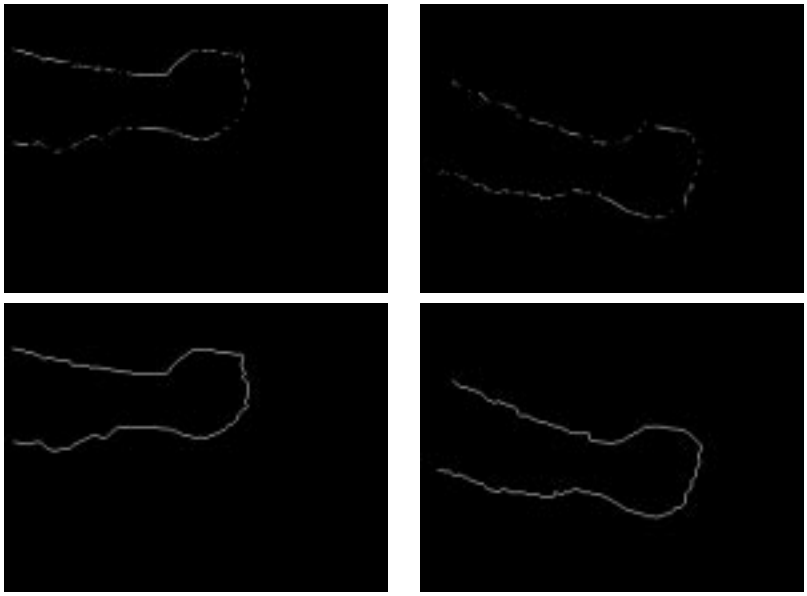

Figure 3: Delineating the foreground objects for images in Figure 2. Upper row: points with high normal flow values and high gradient magnitudes. Lower row: foreground object outlines.

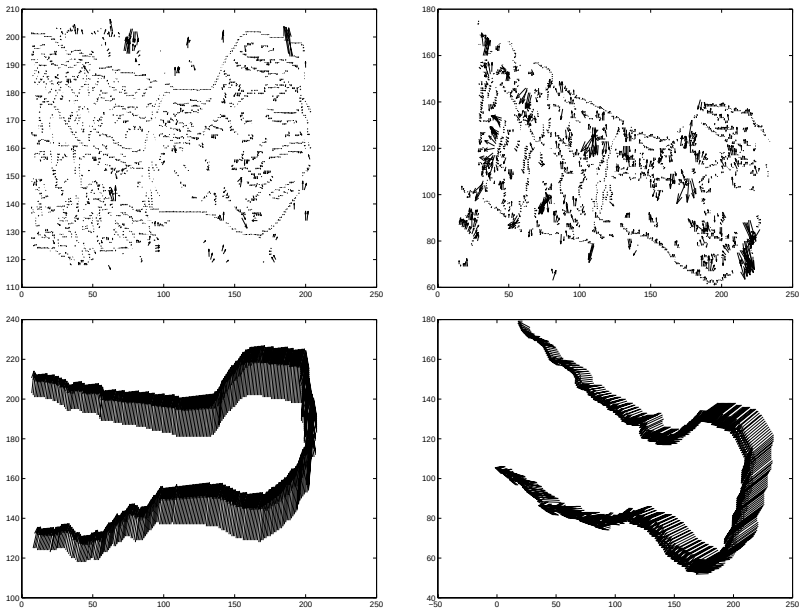

Figure 4: Residual and reestimated flows for the detected arms in Figure 2. Upper row: residual flow computed as the difference between the computed normal flow and the estimated affine normal motion field. Lower row: reestimated affine flow after outlier rejection.

Baseline condition Easy load Difficult load

Figure 5: Facial expressions of an individual in during baseline, easy, and difficult load conditions.

For the eye region, the visual features related to cognitive states of a person include the gaze direction (position of the irises relative to the eyes), the pupil dilation, and the degree of occlusion of the iris and the pupil by the eyelids. For example, the pupil dilation indicates arousal or heightened cognitive activity; while the averted gaze may indicate increased mental activity associated with the processing of data. The visual features related to emotional states include the degree of the openness of the eyes, the position of the eyelids relative to the irises, the position and the shape of the eyebrows (arched, raised, drawn together, etc.), and the existence and shape of lines in particular eye regions (eye corners, between the eyebrows, below lower eyelid). For example, surprise is indicated by wide open eyes with the lower eyelids drawn down, and raised and arched eyebrows; fear is indicated wide open eyes with the upper eyelids raised (exposing the white of the eye) and the lower eyelids tensed and drawn up, and with the eyebrows raised and drawn together; happiness shows primarily in the lower eyelids, which show wrinkles below them and may be raised but are not tense.

Figure 6 illustrates some of those parameters measured on eyes. Note that we expect that the parameters P_1 through P_5 will always be positive; P_3 and/or P_4 can only become small when the eyes are almost closed. Note that other parameters, such as density of lines around eyes and curvature of the eyelids and the eyebrows can be added to complement these parameters.

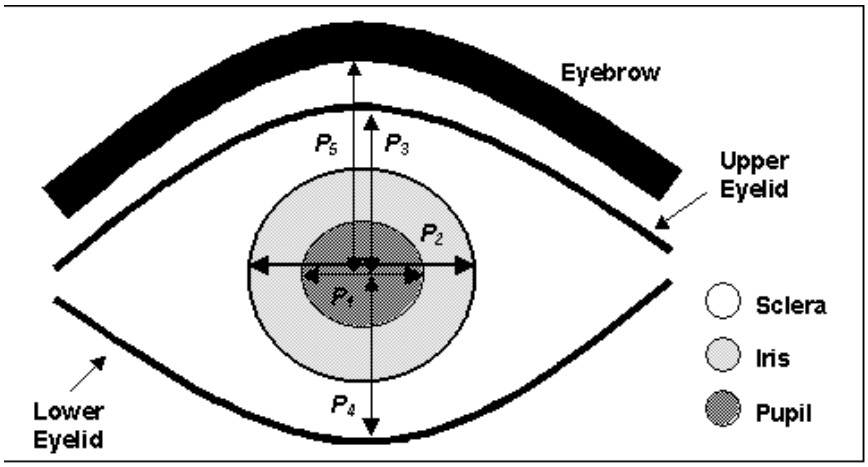

Figure 6: Eye parameters: pupil diameter (P_1), iris diameter (P_2), distance from center of the iris to the upper eyelid (P_3), distance from center of the iris to the lower eyelid (P_4), and distance from center of the iris to the eyebrow (P_5).

The eye parameters are acquired from high-resolution color images. The color edges are detected by combining the gradients in red, green, and blue color bands [78]; the edges are thinned using a non-maxima suppression. Irises and pupils are located via Generalized Hough Transform using multiple size templates; it is assumed that their shapes are always circular. Eyelids and eyebrows are detected using a variation of the method described in [52]; the edges in the vicinity of irises are labeled as the candidates for the eyelids and the eyebrows. Left column of Figure 7 shows eye regions corresponding to anger, surprise and happiness; the right column of the figure shows the results of detecting irises, the eyelids, and the eyebrows for the corresponding images on the left.

Numerical results for the examples shown in Figure 7 are as follows. For the images showing anger (top row) the irises were detected at positions (114,127) and (499,103) with the radii $P_2 = 32$; for the left eye (in the image) the distance from the upper eyelid to the iris was $P_3 = 20$, the distance from the lower eyelid to the iris center $P_4 = 39$, and the distance from the iris center to the eyebrow $P_5 = 38$; for the right eye (in the image) the distance from the upper eyelid to the iris was $P_3 = 11$, the distance from the lower eyelid to the iris center $P_4 = 35$, and the distance from the iris center to the eyebrow $P_5 = 26$. For the images showing surprise (middle row) the irises were detected at positions (100,173) and (483,149) with the radii $P_2 = 32$; for the left eye (in the image) the distance

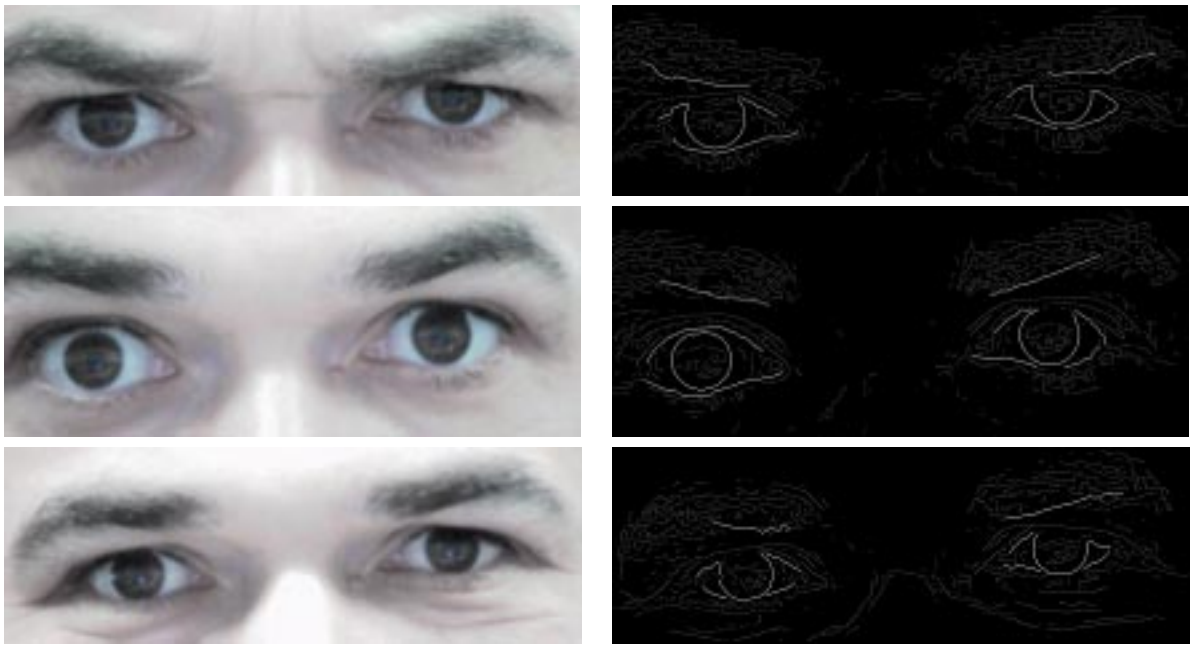

Figure 7: Left column: eye regions displaying anger, surprise, and happiness. Right column: processed eye regions.

from the upper eyelid to the iris was $P_3 = 36$, the distance from the lower eyelid to the iris center $P_4 = 34$, and the distance from the iris center to the eyebrow $P_5 = 71$; for the right eye (in the image) the distance from the upper eyelid to the iris was $P_3 = 36$, the distance from the lower eyelid to the iris center $P_4 = 39$, and the distance from the iris center to the eyebrow $P_5 = 77$. For the images showing happiness (bottom row) the irises were detected at positions (148,139) and (497,115) with the radii $P_2 = 30$; for the left eye (in the image) the distance from the upper eyelid to the iris was $P_3 = 14$, the distance from the lower eyelid to the iris center $P_4 = 27$, and the distance from the iris center to the eyebrow $P_5 = 49$; for the right eye (in the image) the distance from the upper eyelid to the iris was $P_3 = 17$, the distance from the lower eyelid to the iris center $P_4 = 33$, and the distance from the iris center to the eyebrow $P_5 = 48$. Computing the ratios P_3/P_2, P_4/P_2, P_5/P_2 we obtain the following results for the left and right eye pairs: for anger $(0.34, 1.09, 0.81)$ and $(0.62, 1.22, 1.19)$; for surprise $(1.13, 1.06, 2.22)$ and $(1.13, 1.22, 2.4)$; and for happiness $(0.47, 0.9, 1.63)$ and $(0.57, 1.1, 1.6)$.

5.3. Eye-Gaze Tracking

Because people can consider and discard various aspects of a task rather quickly (i.e., in less than 200 ms), eye movements can provide detailed estimate of what information an individual considers. Eye tracking is becoming an increasingly popular on-line measure of high-level cognitive processing (e.g. [40]). By gathering data on the location and duration of eye fixations, psychologists are able to make many inferences about the microstructure of cognition. The use of eye tracking in estimating cognitive states rests on the immediacy assumption (people process information as it is seen) and the eye-mind assumption (the eye remains fixated on an object while the object is being processed). As long as the visual information requires fine discrimination, these assumptions are generally considered valid—but when the visual information is very coarse scale, people can process the information without fixating on it.

In order to reliably separate eye fixations from saccades, one needs to sample gaze-data at least 60 times per second with accuracy of at least two degrees of visual angle. A variety of eye-tracking methods exist. In terms

of the data collected from the eye, two popular methods are 1) shining a light at the eye and detecting corneal reflection and 2) simply taking visual images of the eye and then locating the dark iris area. Which method is best to use depends upon the external lighting conditions.

To compute where in the world the person is fixating, there are three popular methods. The first method simplifies the calculations by having fixed geometries by forcing the person to hold still by biting on a bar or putting the head in a restraint. The second method has the person wear a head sensor that tracks the head orientation and location in 3-dimensions, and then combines this information together with eye-direction information. The third method places the eye-tracking apparatus on the person's head along with a scene camera such that a visual image is displayed showing what the person is currently looking at with a point on the image indicating the object being fixated. To achieve the high levels of accuracy required, all three methods require recalibration for each individual being tracked in a given session, however methods exist for automatic recalibration. While the first method of computing the point-of-fixations is the most accurate, it is purely a research method with no applicability to IHCI for obvious practical reasons. If one ones to track upper-body and facial gestures at the same time, a head-mounted camera is not practical either. The remote camera is the least accurate method, but extreme levels of precision are probably not needed for IHCI.

To separate out fixations from the raw point-of-regard data, the most popular method is to use a movement/time threshold: whenever the distance between consecutive points-of-regard is below a threshold for a sufficient length of time, then a fixation is assumed. A more sophisticated an accurate approach uses a centroid submodel tracing methodology developed by Salvucci and Anderson [**?**]. The methodology involves categorizing eye movements using hidden Markov models (HMMs) and model tracing. Raw eye-data is first categorized into fixations and saccades using a two-state HMM given velocity information alone. Centroids of each fixation are then determined. One could examine which object on the screen is closest to this centroid and simply assume the person was looking at that object. However, because of the noise in the eye data, this would frequently miscategorize the fixation. Instead, another HMM fitting process is used which takes into account the closeness of each fixation to objects on the screen AND the context of which other objects were just fixated. This model fitting process is done by comparing the sequence of fixations with all plausible sequences of fixation, and selecting the sequence with the highest probability (best overall fit).

Figure 8 presents an example of point-of-regard data extracted while a user interacts with a complex computer display. From the location of fixations, one can determine which objects were likely encoded at a detailed level. From the duration of the fixations, one can determine which objects were most likely involved in more detailed computations.

6. Behavioral Processing

Behavioral processing focuses on two kinds of data input: keyboard and mouse. The both keyboard and mouse data are first used as primary input into the computer interface. We are not proposing that these perceptual processing sources of information replace mouse and keyboard. In addition to serving as direct interaction with the interface, keyboard and mouse input will have an additional function of providing insights into the cognitive and affective states of the user. Keystroke data will provide information about the cognitive state of the user through the process of model tracing, which will be described in Section 7. Mouse data will provide information about user cognition and user affect, and this process is described next.

Mouse data can be divided into two primary types: mouse gestures and mouse move-and-clicks. Mouse gestures are movements of the mouse that do not result in mouse clicks. In most HCI environments, these movements are nonfunctional, although some systems provide rollover information. For the moment, we treat rollover cases as if they involve a mouse click. The mouse-gestures, these nonfunctional mouse movements, can be viewed as windows into the mind. That is, they indicate what objects are currently being reasoned about [50]. Common mouse gestures include circling objects, linking objects, circling groups of objects, and shape tracing of large

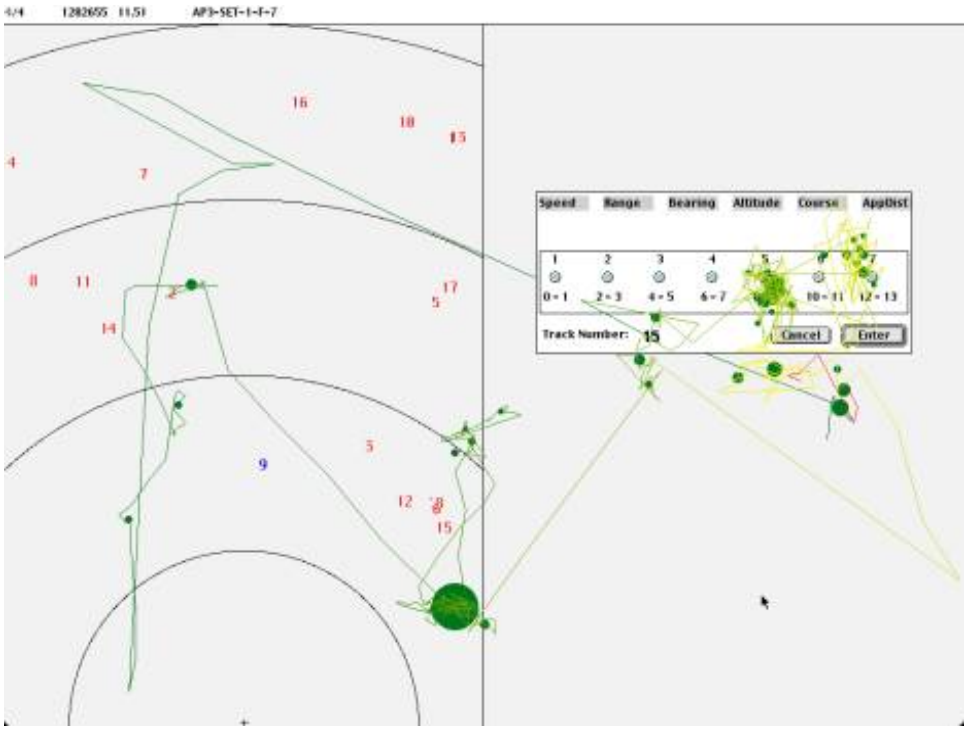

Figure 8: Example eye-tracking data from a complex simulated radar classification task. The moving targets to be identified are represented by the numbers in the left half of the screen. The details of target track 15 are in the table on the right half of the screen. Overlaid on the interface image is 10 seconds of Point-of-Gaze data (changing from dark to light over time). Identified eye fixations (minimum = 100ms) are indicated with disks (larger disks for longer fixations).

objects or groups. Object circling represents indecision about an object. Repetitive movements between objects is a linking gesture, and like circling groups of objects, represents a categorical decision that the linked objects are similar along an important dimension. Shape tracing of larger objects represents reasoning about the shape of the object.

Move-and-click movements have three important dimensions: speed of movement, force of click, and directness of movement to clicked object. These dimensions are indicators of the general level of arousal and of indecision and confusion. Slower movement and weaker clicks combined with direct movements indicate low levels of arousal (i.e., fatigue). Slower movements and weaker clicks combined with indirect movement indicate confusion. Fast movements with strong clicks combined with indirect movements indicate frustration.

To recognize the mouse gestures, one can use a technique that we developed in the context of sign language recognition [70]. There, hand gestures corresponding to American Sign Language are first located using projection analysis and then normalized in size, while recognition takes place using hybrid mixture of (connectionist and symbolic) experts consisting of Ensembles of Radial Basis Functions (ERBFs) and Decision Trees (DTs). ERBFs display robustness facing the variability of the data acquisition process using for training both original data and their distortions caused by geometrical changes and blurring, while preserving the topology that is characteristic of raw data. In addition, ERBFs are similar to boosting as each of their RBF components is trained to capture a subregion of the perceptual or behavioral landscape and can then properly recognize it. Inductive learning, using DTs, calls for numeric to sub-symbolic data conversion, suitable for embodied cognition. The ERBF output vectors chosen for training are tagged as 'CORRECT' (positive example) or 'INCORRECT' (negative examples) and properly quantized. The input to DT implemented using C4.5 [112] consists of a string of learning (positive and negative) events, each of them described as a vector of discrete attribute values. To further parse the move-and-click movements, one can use Hidden Markov Models (HMM) because the user is likely to stay in a given affective state for several movements. The states of the Markov model are the diagnosed affective states (alertness, fatigue, confusions, etc).

7. Embodied Cognition

The embodied cognition module has at its core an embodied cognitive model and a model tracing function. A cognitive model is capable of solving the tasks solved by the humans using the same cognitive steps as humans use to solve the tasks. An embodied cognitive model also has affective states to match the user's states and the ability to perceive an interact with an external world in a similar way as the user does. In model tracing, the model is aligned with the task and behavioral choice data (i.e., what the state of the world is and what the human chose to do) such that one can see which internal cognitive steps the human must have taken in order to produce the observed behavioral actions. Towards that end, the embodied cognition model also uses the affective sub-symbols and their degree of belief, derived earlier by the perceptual and behavioral processing modules.

Currently the best way to build models of embodied cognition is to use a cognitive architecture (e.g., ACT-R, Soar, EPIC, 3CAPS) that has a relative complete and well-validated framework for describing basic cognitive activities at a fine grain-size. The currently most developed framework that works well for building models of embodied cognition is ACT-R/PM [17], a system that combines the ACT-R cognitive architecture [8] with a modal theory of visual attention [9] and motor movements [34]. ACT-R is a hybrid production system architecture, representing knowledge at both a symbolic level (declarative memory elements and productions) and sub-symbolic level (the activation of memory elements, the degree of association among elements, the probability of firing productions, etc). ACT-R/PM contains precise (and successful) methods for predicting reactions times and probabilities of responses that take into account the details of and regularities in motor movements, shifts of visual attention, and capabilities of human vision. The task for the embodied cognition module is to build a detailed mapping of the interpretations (i.e., motion/affective state) of the parsed sensory-motor data onto the ACT-R/PM model.

One can extend ACT-R/PM to make it a true model of embodied cognition by incorporating the effects of

affect on performance. For example, in addition to handling the interactions among and between memory, vision, and motor movements, the model becomes fatigued over time and distracted when there is too much to attend to. Better than merely *becoming* fatigued and distracted, such an extended ACT-R/PM can *model the effects of fatigue* and distraction on memory, vision, and motor behavior and thereby on performance. Similar to people, as the model becomes fatigued several changes should occur. First, the model slows down (increasing the interval between physical actions and shifts in visual attention, as well as increasing the time needed to store or retrieve information from memory). Second, the accuracy of its responses decreases (this includes the physical accuracy due to increased noise in eye-hand coordination and mental accuracy due to increased noise in memory retrieval — e.g., retrieving the target's old, rather than current, flight information). Third, it the model becomes distracted, losing its focus of attention (running the risk of applying the "right" response to the "wrong" object or the "wrong" response to the "right" object). Fourth, it becomes narrower in what it chooses to encode [41].

This incorporation of affective sub-symbols into models of embodied cognition is a product in its own right. Such a capability can be applied to other task environments (simulated or prototypes of real systems) to determine changes in human performance over time. As such, models of embodied cognition could become an important tool for designers of real-time, safety-critical systems (see, e.g., [29]). One novelty here lies in using a broader range of nonverbal data in guiding the model tracing process. Recent work in cognitive science suggests that nonverbal information, such as gestures, provides important insights into an individual's cognition [3, 50]. Mouse gestures, the eye data, and affective states are important tools to improve this model tracing process.

One can explore three qualitatively different methods of incorporating affect into the cognitive model. First, affect can be thought of as directly modifying parameters in the cognitive model to produce relatively simple changes in behavior. For example, fatigue may affect processing speed (i.e., how fast someone thinks) as well as working memory capacity (i.e., how much information they can keep in mind). Similarly, confusion or frustration may influence the noise parameter in the decision process (i.e., the likelihood of making non- optimal choices) or the threshold amount of effort a person will expend on a task (which influences the probability of giving up). Parameters controlling processing speed, working memory capacity, noise, and effort expended are formally defined within the ACT-R architecture. Second, affect can also change more structural or strategic aspects of the cognitive model. For example, when a person becomes confused, fatigued, or frustrated, they may adopt an entirely different way about thinking about the task and making choices (i.e., alternative strategies). Thus, the performance parameters of the cognitive model may be held constant but, instead, action and decision rules themselves change as the affect changes. A third possibility is some combination of these two types of changes in the model with changing affect. Individuals use a wide variety of qualitatively different strategies to solve any given type of problem [49], and changes in model performance parameters are likely to produce changes in strategy choice. For example, in a classic decision making study, Payne, Bettman, and Johnson [45] showed that as the cognitive effort required for task performance increased (thereby placing greater demands on a limited capacity working memory system) the decision-making strategies that people adopted changed. Lohse and Johnson [40] showed that changes in decision-making strategies were also induced by tradeoffs between perceptual-motor versus cognitive effort. Hence, it may well be that changes in strategies induced by changes in affective state are mediated by changes in underlying cognitive parameters. ACT-R contains clear predictions for how certain parameter changes will influence strategy choice, assuming a good characterization of the features of each strategy.

The process of model tracing keeps the model aligned with the user. It takes as primary input the behavioral interactions with the interface (i.e., the keystroke and mouse click data), and tries to match symbolic steps in the model. In a production system model, this amounts to matching to sequences of production firings. There are three factors that shape the model tracing process. First, any realistic model of human cognition acknowledges some stochastic variability in human choices. That is, at many points in time, the model on its own must choose randomly (although typically with particular biases) among a set of alternative actions. Model tracing examines the behavioral data and identifies which among all the possible alternative paths the model could have taken best fits the behavioral data observed. The second factor shaping model tracing is that typically there are several

internal steps for every external step. Thus, the model must be run for several steps, with each step potentially having alternative choices, to produce the full set of possible matches to the behavioral data. If there are many internal steps between external behaviors, then there may be a large set of internal paths that need to be generated. The third factor is that the behavioral data may not uniquely distinguish among different model paths. In such circumstances, one must select the currently most probably path. With the addition of eye-tracking data, the density of observable data points goes up significantly, making it easier to match model to data.

8. Adaptation of User/System Interface

A system based on IHCI can adapt the interface based on current needs of the human participant as found in the embodied model of cognition. As stated earlier, different affective and cognitive diagnoses include confusion, fatigue, stress, momentary lapses of attention, and misunderstanding of procedures. Different adaptations include simplifying the interface, highlighting critical information, and tutoring on selected misunderstandings. The types of interface adaptations that one can consider include: 1) addition and deletion of task details; 2) addition and deletion of help/feedback windows; 3) changing the formatting/organization of information; and 4) addition and removal of automation of simple subtasks. These changes are described generically here for generality.

With respect to the addition and deletion of task details, the important insight is that modern interfaces contain details relevant to many subtasks. When an operator becomes confused or distracted it may well be because details relevant to subtask A interfere with his or her attention to details needed to accomplish subtask B. One general strategy is to identify the currently critical subtask with the goal of eliminating details relevant to other subtasks or enhancing details relevant to the critical subtask. Interface details relevant to other subtasks can be replaced when the user appears able to handle them. The combination of the point- of-gaze data (via eye-tracking) and the affective response data (via facial expressions) provides important information regarding which aspects of the interface to change and in what manner to change them. For example, if an important aspect of the screen is not attended and the individual appears fatigued, then that aspect should be highlighted. By contrast, if an aspect of the screen is attended for an unusually long period of time and is coupled with a look of confusion, then a situation-relevant help window will be displayed. All of the possible interface structures that are possible will have advantages and disadvantages that are likely to vary with the cognitive and affective state of the user. Thus, different interface structures will be optimal for different points in time—if a particular structure is generally suboptimal, there is no reason to ever use it. For example, having a help window display help messages may be useful for a confused individual, but may be distracting for a nonconfused individual. Alternatively, having less information on the screen may be helpful to a fatigued individual, but harmful to fully attentive individual (because they could make appropriate use of the extra information to handle more subtasks).

Because no particular interface structure is better than another one across all situations, one can avoid strange feedback loops in which a user becomes trained (either implicitly or explicitly) to always look frustrated because that makes the task easier. Instead, users will be trained to correctly externalize their internal states. For example, when frustrated, look frustrated because that will produce a change that is useful for dealing with this particular source of frustration; but when not frustrated, do not look frustrated because that will produce chances that reduce optimal performance (of someone who is not frustrated).

A model of embodied cognition that is continuously being updated to reflect the individual's cognitive, perceptual, motor, and affective states makes it possible to have two different methods of adapting the interface: reactive and proactive. In reactive adaptation, the system waits for external evidence of some cognitive or affective change before adapting the interface. For example, the user becomes confused and this confusion is manifested by a confused look, longer choice latencies, and longer fixations across a broader range of entities. A reactive system adapts the interface only after the confusion is manifested. Alternatively, a proactive system applies the model of embodied cognition (which is capable of performing the task) to the correct task state and predicts what kinds of problems the user is likely to encounter. These predictions are used to adapt the interface; that is, the interface

changes before the user becomes confused, frustrated, or bored (or at least before it can be diagnosed from outward performance changes). Once the model tracing has approached a high-level of accuracy that we believe it can using this broadened set of inputs to it, then one can explore including proactive interface adaptation in the system.

For either proactive or reactive adaptation, the adaptation will have to be conservative (i.e., relatively infrequent with relatively small changes at any one time). An interface that is constantly changing is a source of frustration in itself. Moreover, there should be a relatively small set of possible changes to the interface, and the set needs to be introduced during initial training. The embodied model provides insights into how conservative to be (i.e., to predict how disruptive various interfaces changes will be), in addition to providing insights into what interface adaptations are likely to be helpful.

9. Conclusions

This paper has described a W5+ methodology for Intelligent HCI that extends on the current method of interpreting human activities. Our approach to Intelligent HCI has four central pieces. First, the behavioral interactions between the user and interface are processed in a very rich fashion, including the novel use of mouse gestures, to provide a more rich understanding of the user's cognitive and affective state. Second, additional nonverbal information is gathered through perceptual processing of eye-gaze, pupil size, facial expressions, and arm movements to further enrich the understanding of the user's cognitive and affective state. Third, an embodied model of cognition is synchronized with the behavioral and perceptual data to produce a deep understanding of the user's state. Fourth, the computer interface is adapted in reaction to problems diagnosed in the user's cognitive and affective state in a way that is task sensitive.

While our complete methodology has not yet been implemented in a running system, we have described the tools that would be required to implement the system. Further, these tools appear to be well within current computational capabilities. We are currently engaging in research to further flesh out the computer science and cognitive psychology underlying these tools.

There are some subtle components in our methodology that require further comment. In particular, the process of building an embodied cognitive model has three separate advantages for intelligent adaptation of an interface. First, it forces one to develop highly detailed models of the precise cognitive and affective problems a user might experience because the model must be designed to perform the task in the same way that the user does. This enforced detail allows for more precise diagnosis of sources of problems. Second, the embodied cognitive model allows one to test the consequences of different changes to the interface so that one can have a good understanding of which interface changes are likely to help and why they will help. Without the embodied cognitive model, one must simply rely on simple rules of thumb and past experiences to determine which changes will be effective. Third, the embodied cognitive model has a predictive component that allows one to predict what problems a user is likely to have in the future and so can warn the user in advance of problems (e.g., when fatigue is likely to begin to occur given the recent load and current arousal levels).

In developing a running system that implements our methodology, we recommend a strategy of using a simulated task environment. In field research, there is often too much complexity to allow for any more definite conclusions, and in laboratory research, there is usually too little complexity to allow for any interesting conclusions [14]. Those who study complex situations as well as those who wish to generalize their results to complex situations have often faced the dilemma so succinctly framed by Brehmer and Dörner. Simulated task environments are the solution to this dilemma. The term, simulated task environment, is meant to be both restrictive and inclusive. There are many types of simulations; however, the term is restricted to those that are intended as simulations of task environments. At the same time, the term includes the range of task simulations from high fidelity ones that are intended as a substitute for the real thing, all the way to microworlds that enable the performance of tasks that do exist (Gray, in press). The common denominator in these simulated task environments is

the researcher's desire to study complex behavior. The task environment must be complex enough to challenge the current state of the art, but malleable enough so that task complexity and interface adaptivity can be controlled and increased as the research progresses. These requirements can be met by using simulated task environments. We are working on the Intelligent HCI approach in the context of human operators interacting with ARGUS, a simulated task environment for radar operator tasks [24]. This radar operator task represents a real-time, safety-critical environment in which improving human-computer interaction is of utmost importance. Similar issues relating to image processing, cognitive modeling, and intelligent interface adaptation can be found in the HCI of a wide variety of domains (e.g., medical, educational, business). We can collect and time stamp every mouse click made by the subject, every system response, and every mouse movement with 17 msec accuracy and to interleave this record with point-of-gaze data collected 60 times per second. ACT-R/PM models currently interact with both Argus. In addition, because we own the ARGUS code and it is written in Lisp, the simulated task environment is easy to modify.

Perception in general, and form and behavior analysis in particular are important not because they can merely describe things, but as Aristotle realized long ago, because they make us know and bring to light many difference between things so we can categorize them and properly respond to their affordances. Once both forms and behavior are represented their most important functionality is to serve for discrimination and classification. Recognition is thus based on both perceptual form and behavior, together with their associated functionality. Form and behavior analysis consider things like average prototypes and/or the similarity holding between them, while functionality carves the perceptual and behavioral layout according to innate physical and geometrical constraints, sensory motor affordances, and their corresponding cognitive mental states. According to this view functional and purposive recognition takes precedence over perceptual and behavioral reconstruction. As things are always changing and constancy is an illusion, form and behavioral recognition require generalization and motivate learning and adaptation. What form and behavioral analysis do not require, however, is to take them to bits in ways that destroy the very relations that may be of essence; as Lewontin [111] would say "one murders to dissect". Embodies cognition, as described and advocated in this paper, provides the glue connecting the apparent visual form and behavior with hidden mental models, which bear on both functionality and performance. To further emphasize the important role functionality plays in perception, it is instructive to recall Oliver Sack's well-known story, "The Man Who Mistook his Wife for a Hat". The story describes someone who can see, but not interpret what he sees: shown a glove, the man calls it a "receptacle with five protuberances". The moral of that story is that people see not only with the eyes, but with the brain too. In other words, perception involves a whole and purposive cognitive process, and this is what this paper advocates in terms of technology and tools for intelligent HCI.

References

[1] J.K. Aggarwal, Q. Cai, and B. Sabata. Nonrigid motion analysis: Articulated and elastic motion. Computer Vision and Image Understanding, 70:142?156, 1998.

[2] K. Akita. Image sequence analysis of real world human motion. Pattern Recognition, 17:73?83, 1984.

[3] Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. Cognitive Psychology, 25(4), 468-523.

[4] Altmann, E. M., & Gray, W. D. (1999a). Functional decay in serial attention. Proceedings of the Sixth ACT-R Workshop (pp.). Fairfax, VA: ARCH Lab.

[5] Altmann, E. M., & Gray, W. D. (1999b). Preparing to forget: Memory and functional decay in serial attention. Manuscript submitted for publication.

[6] Altmann, E. M., & Gray, W. D. (2000). Managing attention by preparing to forget. Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting (pp.). Santa Monica, CA: Human Factors and Ergonomics Society.

[7] Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. Artificial Intelligence, 42, 7-49.

[8] Anderson, J. R., & Lebiére, C. (Eds.). (1998). Atomic components of thought. Hillsdale, NJ: Erlbaum.

[9] Anderson, J. R., Matessa, M.,& Lebiére, C. (1997). ACT-R: A theory of higher-level cognition and its relation to visual attention. Human-Computer Interaction, 12(4), 439-462.

[10] N.I. Badler, C.B. Phillips, and B.L. Webber. Simulating Humans. Oxford University Press, New York, 1993.

[11] N.I. Badler and S.W. Smoliar. Digital representations of human movement. Computing Surveys, 11:19-38, 1979.

[12] R. Barzel. Physically-Based Modeling for Computer Graphics. Academic Press, Boston, MA, 1992.

[13] C. Bregler. Learning and recognizing human dynamics in video sequences. In Proc. Computer Vision and Pattern Recognition, pages 568-574, 1997.

[14] Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. Computers in Human Behavior, 9(2-3), 171-184.

[15] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In Proc. Computer Vision and Pattern Recognition, pages 8-15, 1998.

[16] H. Buxton and R. Howarth. Watching behaviour: The role of context and learning. In Proc. International Conference on Image Processing, volume 2, pages 797-800, 1996.

[17] Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson & C. Lebióre (Eds.), The atomic components of thought (pp. 167-200). Hillsdale, NJ: Erlbaum.

[18] C. Cedras and M. Shah. Motion-based recognition: A survey. Image and Vision Computing, 13:129-155, 1995.

[19] Fu, W.-T., & Gray, W. D. (1999a). ACT-PRO: Action protocol tracer – a tool for analyzing simple, rule-based tasks. Proceedings of the Sixth ACT-R Workshop (pp.). Fairfax, VA: ARCH Lab.

[20] Fu, W.-T., & Gray, W. D. (1999b). Redirecting direct-manipulation, or, what happens when the goal is in front of you but the interface says to turn left? Proceedings of the CHI 99 Extended Abstracts (pp. 226-227). New York: ACM Press.

[21] Fu, W.-t., & Gray, W. D. (2000). Memory versus Perceptual-Motor Tradeoffs in a Blocks World Task. Manuscript submitted for publication.

[22] Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. Psychological Review, 100(2), 279-297.

[23] Gray, W. D. (2000). The nature and processing of errors in interactive behavior. Cognitive Science, 24(2).

[24] Gray, W. D. (in press). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. In R. Mahan, D. Serfaty, S. Kirschenbaum,

[25] M. McNeese,& L. Elliott (Eds.), Scaled Worlds (working title) . Hillsdale, NJ: Erlbaum.

[26] Gray, W. D., & Altmann, E. M. (in press). Cognitive modeling and human-computer interaction. In W. Karwowski (Ed.), International encyclopedia of ergonomics and human factors . New York: Taylor & Francis, Ltd.

[27] Gray, W. D., & Boehm-Davis, D. A. (in press). Milliseconds Matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. Journal of Experiment Psychology: Applied.

[28] Gray, W. D., & Fu, W.-t. (2000). The influence of source and cost of information access on correct and errorful interactive behavior, Manuscript submitted for publication .

[29] Gray, W. D., Palanque, P., & Paternò, F. (in press). Introduction to the Special Issue on: Interface Issues and Designs for Safety-critical Interactive Systems. ACM Transactions on Computer-Human Interaction.

[30] Gray, W. D., & Salzman, M. C. (1998). Repairing damaged merchandise: A rejoinder. Human-Computer Interaction, 13(3), 325-335.

[31] Gutta, S., Huang, J., Phillips, P. J., & Wechsler, H. (in press). Mixtures of experts for classification of gender, ethnic origin and pose of human faces. IEEE on Neural Networks.

[32] Huang, J., Gutta, S., & Wechsler, H. (1996). Detection of Human Faces Using Decision Trees. Paper presented at the 2nd Int. Conf. on Automatic Face and Gesture Recognition, Killington, VT.

[33] Huang, T. S. (1997). Workshop on Human Computer Intelligent Interaction and Human Centered Systems. Paper presented at the ISGW '97 NSF Interactive Systems Grantees Workshop,.

[34] Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Human-Computer Interaction, 12(4), 391-438.

[35] Krash, R., & Breitenbach, F. W. (1983). Looking at looking: The amorphous fixation measure. In R. Groner, C. Menz,

[36] D. Fisher, & R. Monty (Eds.), Eye Movements and Psychological Functions: International Views . Hillsdale, NJ: Erlbaum.

[37] M.L. Knapp and J.A. Hall. Nonverbal Communication in Human Interaction. Harcourt Brace, New York, 1997.

[38] Liu, C., & Wechsler, H. (2000). Robust coding schemes for indexing and retrieval from large face databases. IEEE Trans. on Image Processing, 9(1), 132-137.

[39] Liu, C., & Wechsler, H. (in press). Evolutionary pursuit and its application to face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence.

[40] Lohse, G. L., & Johnson, E. J. (1996). A comparison of two process tracing methods for choice tasks. Organizational Behavior and Human Decision Processes, 68(1), 28-43.

[41] Lovett, M. C., & Schunn, C. D. (1999). Task representations, strategy variability and base-rate neglect. Journal of Experimental Psychology: General, 128(2), 107-130.

[42] Newell, A., & Card, S. K. (1985). The prospects for psychological science in human-computer interaction. Human- Computer Interaction, 1(3), 209-242.

[43] S.A. Niyogi & E.H. Adelson. Analyzing and recognizing walking figures in XYT, In *Proc. Computer Vision and Pattern Recognition*, pages 469–474, 1994.

[44] V.I. Pavlovic, R. Sharma and T.S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19:677-695, 1997.

[45] Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The adaptive decision maker. New York: Cambridge University Press.

[46] Picard, R. (1995). Affective computing (Technical Report 321). Cambridge, MA: MIT Media Laboratory.

[47] Picard, R. (1998). Toward agents that recognize emotion (Technical Report 515). Cambridge, MA: MIT Media Laboratory.

[48] Schoelles, M. J., & Gray, W. D. (2000). Argus Prime: Modeling emergent microstrategies in a complex simulated task environment. Proceedings of the Third International Conference on Cognitive Modeling (pp.). Groningen, NL: .

[49] Schunn, C. D., & Reder, L. M. (2001). Another source of individual differences: Strategy adaptivity to changing rates of success. Journal of Experimental Psychology: General.

[50] Schunn, C. D., Trickett, S., & Trafton, J. G. (1999). What Gestures Reveal about the Scientist's Mind: Data Analyses of Data Analysis. Paper presented at the Krasnow Institute Brown Bag Series, George Mason University.

[51] Shneiderman, B., & Maes, P. (1997). Direct manipulation vs interface agents. interactions, 4(Nov-Dec), 643-661.

[52] Sirohey, S., Rosenfeld, A., & Duric, Z. (2001). Eye tracking. Pattern Recognition. In press.

[53] T.-J. Cham and J.M. Rehg. A multiple hypothesis approach to figure tracking. In Proc. Computer Vision and Pattern Recognition, volume 2, pages 239-245, 1999.

[54] E. Cohen, L. Namir, and I.M. Schlesinger. A New Dictionary of Sign Language. Mouton, The Hague, The Netherlands, 1977.

[55] A.F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. In Proc. Royal Society Workshop in Knowledge-based Vision in Man and Machine, London, England, 1997.

[56] J.W. Davis and A.F. Bobick. The representation and recognition of human movement using temporal templates. In Proc. Computer Vision and Pattern Recognition, pages 928-934, 1997.

[57] L.S. Davis, D. Harwood, and I. Haritaoglu. Ghost: A human body part labeling system using silhouettes. In Proc. ARPA Image Understanding Workshop, pages229-235, 1998.

[58] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In Proc. Computer Vision and Pattern Recognition, pages II:126-133, 2000.

[59] N. Eshkol and A. Wachmann. Movement Notation. Weidenfeld & Nicholson, London, England, 1958.

[60] I.A. Essa and A.P. Pentland. Facial expression recognition using a dynamic model and motion energy. In Proc. International Conference on Computer Vision, pages 360-367, 1995.

[61] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19:757-763, 1997.

[62] D.M. Gavrila and L.S. Davis. 3D model-based tracking of humans in action: A multi-view approach. In Proc. Computer Vision and Pattern Recognition, pages 73-80, 1996.

[63] D.M. Gavrila. The visual analysis of human movement: A survey. Computer Vision and Image Understanding, 73:82-98, 1999.

[64] D.M. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In Proc. International Conference on Computer Vision, pages 87-93, 1999.

[65] Proc. International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, 1995.

[66] Proc. International Conference on Automatic Face and Gesture Recognition, Killington, VT, 1996.

[67] Proc. International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.

[68] Proc. International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000.

[69] K. Gould and M. Shah. The trajectory primal sketch: A multi-scale scheme for representing motion characteristics. In Proc. Computer Vision and Pattern Recognition, pages 79-85, 1989.

[70] S. Gutta, I.F. Imam, and H. Wechsler. Hand gesture recognition using ensemble of radial basis function (RBF) networks and decision trees. International Journal of Pattern Recognition and Artificial Intelligence, 11:845-872, 1997.

[71] I. Haritaoglu, D. Harwood, and L.S. Davis. W4S: A real-time system for detecting and tracking people. In Proc. Computer Vision and Pattern Recognition, pages 962-968, 1998.

[72] J.K. Hodgins and N.S. Pollard. Adapting simulated behaviors for new characters. In Proc. SIGGRAPH, pages 153-162, 1997.

[73] D. Hogg. Model based vision: A program to see a walking person. Image and Vision Computing, 1:5-20, 1983.

[74] A. Hutchinson Guest. Choreo-graphics: A Comparison of Dance Notation Systems from theFifteenh Century to the Present. Gordon and Breach, New York, 1989.

[75] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient matching of pictorial structures. In Proc. Computer Vision and Pattern Recognition, pages II:66-73, 2000.

[76] S. Ioffe and D.A. Forsyth. Finding people by sampling. In Proc. International Conference on Computer Vision, pages 1092-1097, 1999.

[77] S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Detection and location of people in video images using adaptive fusion of color and edge information. In Proc. International Conference on Pattern Recognition, 2000.

[78] B. Jähne. *Digital Image Processing*. Springer-Verlag, Berlin, Germany, 1997.

[79] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14:210-211, 1973.

[80] G. Johansson. Spatio-temporal differentiation and integration in visual motion perception. Psychological Research, 38:379-393, 1976.

[81] S.X. Ju, M.J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In [67], pages 38-44, 1995.

[82] S.K. Jung. Motion analysis of articulated object for optical motion capture. Tech Memo 97-8 (PhD thesis), Korea Advanced Institute of Science and Technology, Taejon, Korea, 1996.

[83] I. Kakadiaris and D. Metaxas. 3D human body model acquisition from multiple views. In Proc. International Conference on Computer Vision, pages 618-623, 1995.

[84] H.-J. Lee and Z. Chen. Determination of 3D human body posture from a single view. Computer Vision, Graphics, and Image Processing, 30:148-168, 1985.

[85] M.K. Leung and Y.-H. Yang. Human body segmentation in a complex scene. Pattern Recognition, 20:55- 64, 1987.

[86] M.K. Leung and Y.-H. Yang. First sight: A human body outline labeling system. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17:359-377, 1995.

[87] S.J. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. In [68], pages 348-353, 2000.

[88] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking Groups of People. Computer Vision and Image Understanding, 80:42-56, 2000.

[89] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. Computer Vision and Image Understanding, 81:231-268, 2001.

[90] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In Proc. Computer Vision and Pattern Recognition, pages 193-199, 1997.

[91] N.M. Oliver, B. Rosario, and A.P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:831–843, 2000.

[92] V. Philomin, R. Duraiswami, and L.S. Davis. Quasi-random sampling for condensation. In Proc. European Conference on Computer Vision, 2000.

[93] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. International Journal of Computer Vision, 23:261-282, 1997.

[94] K.E. Price. Annotated computer vision bibliography. http://iris.usc.edu/Vision-Notes/bibliography/contents.html

[95] R.F. Rashid. Towards a system for the interpretation of moving light displays. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2:574-581, 1980.

[96] J.M. Regh and T. Kanade. Model-based tracking of self occluding articulated objects. In Proc. International Conference on Computer Vision, pages 612-617, 1995.

[97] J. Rittscher and A. Blake. Classification of human body motion. In Proc. International Conference on Computer Vision, pages 634-639, 1999.

[98] K. Rohr. Towards model-based recognition of human movements in image sequences. CVGIP: Image Understanding, 59:94-115, 1994.

[99] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In Proc. European Conference on Computer Vision, 2000.

[100] Y. Song, L. Goncalves, E. di Bernardo, and P. Perona. Monocular perception of biological motion: Detection and labeling. In Proc. International Conference on Computer Vision, pages 805-813, 1999.

[101] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In [65], pages 189-194, 1995.

[102] S. Wachter and H.H. Nagel. Tracking persons in monocular image sequences. Computer Vision and Image Understanding, 74:174-192, 1999.

[103] C.R Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19:780-785, 1997.

[104] Y. Yacoob and L.S. Davis. Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18:636-642, 1996.

[105] M.-H. Yang and N. Ahuja. Recognizing hand gestures using motion trajectories. In Proc. Computer Vision and Pattern Recognition, volume 1, pages 466-472, 1999.

[106] V.M. Zatsiorsky. Kinematics of Human Motion. Human Kinetics, 1997.

[107] H. Hess and J.M. Polt, ”Pupil Size in Relation to Mental Activity during Simple Problem-Solving”, Science, Vol. 143, Issue 3611, March 13, 1964, pp. 1190-1194.

[108] B.C. Goldwater, ”Psychological Significance of Pupillary Movements”, Psychological Bulletin, Vol. 77, No. 5, 1972, pp. 340-355.

[109] M.P. Janisse, Pupillometry, Hemisphere Publishing Company, Washington, DC, 1977.

[110] Bajcsy, R. and J. Kosecka (1995), The Problem of Signal and Symbol Integartion: A Study of Cooperative Mobile Autonomous Agent Behavior, DAGM Symposium, Bielefield, Germany.

[111] R. C. Lewontin, The Science of Metamorphosis, The New York Reviews of Books XXXVI (7), 1989.

[112] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[113] H. Wechsler, Computational Vision, Academic Press, 1990.