Available online at www.sciencedirect.com
SCIENCE DIRECT
Int. J. Human-Computer Studies 64 (2006) 847–873
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Toward a decision-theoretic framework for affect recognition and user assistance

Wenhui Liao[a], Weihong Zhang[a], Zhiwei Zhu[a], Qiang Ji[a,*], Wayne D. Gray[b]

[a]*Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA*
[b]*Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA*

Received 13 May 2005; received in revised form 23 March 2006; accepted 2 April 2006
Communicated by S. Brave
Available online 19 May 2006

Abstract

There is an increasing interest in developing intelligent human–computer interaction systems that can fulfill two functions—recognizing user affective states and providing the user with timely and appropriate assistance. In this paper, we present a general unified decision-theoretic framework based on influence diagrams for simultaneously modeling user affect recognition and assistance. Affective state recognition is achieved through active probabilistic inference from the available multi modality sensory data. User assistance is automatically accomplished through a decision-making process that balances the benefits of keeping the user in productive affective states and the costs of performing user assistance. We discuss three theoretical issues within the framework, namely, user affect recognition, active sensory action selection, and user assistance. Validation of the proposed framework via a simulation study demonstrates its capability in efficient user affect recognition as well as timely and appropriate user assistance. Besides the theoretical contributions, we build a non-invasive real-time prototype system to recognize different user affective states (stress and fatigue) from four-modality user measurements, namely physical appearance features, physiological measures, user performance, and behavioral data. The affect recognition component of the prototype system is subsequently validated through a real-world study involving human subjects.

Keywords: Affective computing; Human–computer interaction; Influence diagrams; Active sensing; Stress modeling; Fatigue recognition

1. Introduction

The field of human–computer interaction (HCI) has moved from a focus on user-friendly graphical user interfaces (GUIs) to systems that bring to bear powerful representations and inferential machinery (Maes and Schneiderman, 1997) in understanding, explaining, justifying, or augmenting user actions. An important example of this new wave in HCI is the design of user assistance systems that enhance users' daily performance (Horvitz, 1999). Although progress is being made in user-modeling (Bauer et al., 2001), augmented cognition, and adaptive user interfaces (Hass and Hettinger, 2001), the majority of existing systems continue to assume normative performance, and all existing systems fail to adapt to user affect. A constellation of recent findings, from neuroscience, psychology, and cognitive science, suggests that emotion plays surprisingly critical roles in users' rational, functional, and intelligent behaviors (Picard et al., 2001). In fact, the situations where affective considerations are most critical are precisely the types of situations where the consequences of the human–machine interaction failures are most severe. It is especially important to recognize dangerous affect in the increasing numbers of HCI systems in critical, typically high-stress, applications such as air traffic control, process control in nuclear power plants, emergency vehicle dispatchers, pilots and drivers, and a variety of military operational contexts. In fact, the increasing frequency of accidents and incidents attributed

*Corresponding author. Tel.: +1 518 276 6440; fax: +1 518 276 6261.
E-mail addresses: liaow@rpi.edu (W. Liao), zhangw9@rpi.edu (W. Zhang), zhuz@rpi.edu (Z. Zhu), jiq@rpi.edu (Q. Ji), grayw@rpi.edu (W.D. Gray).

1071-5819/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2006.04.001

to the broad area of "human error" in a variety of settings could be reduced by considering the user affective states in system design, particularly such negative states as stress, fatigue, anxiety, and frustration. Therefore, recognizing such negative user affect and providing appropriate interventions to mitigate their effects is critical for the successful completion of a task, for avoiding (often disastrous) errors, for achieving optimal performance, for improving HCI experience, and for improving learning and decision-making capability.

The causes and manifesting features of various user affective states have been extensively investigated in psychology, computer vision, physiology, behavioral science, ergonomics and human factor engineering (Beatty, 1982; Gardell, 1982; Ortony et al., 1988; Breazeal, 1999; Mindtools, 2004). In spite of the findings from diverse disciplines, it is still a rather challenging task to develop an intelligent user affect recognition and assistance system. First, the expression and the measurements of user affect are very much person-dependent and even time or context dependent for the same person. Second, the sensory observations are often ambiguous, uncertain, and incomplete. Third, users' affective states are dynamic and evolve over time. Fourth, both affect recognition and user assistance must be accomplished in a timely and appropriate manner. Finally, lack of a clear criterion for ground-truthing affective states greatly increases the difficulty of validating affect recognition approaches and user assistance systems.

In this paper, we propose a general dynamic probabilistic decision-theoretic model based on Influence Diagrams (IDs) (Howard and Matheson, 1981) for unifying affect recognition with user assistance. We are interested in recognizing negative task-dependent affective states (e.g. stress, fatigue, anxiety, confusion, frustration, etc.) and providing assistance to mitigate their effects in order to maintain user in a productive state. Such an ID explicitly includes random variables that represent affective states and sensory observations, decision variables that represent user assistance and sensory actions, and utility functions that represent the benefits and costs associated with user assistance. Within the proposed framework, efficient user affect recognition can be achieved by an active inference based on selecting and integrating a subset of most informative observations; meanwhile, timely and appropriate assistance can be achieved by a decision choice balancing between the benefit of the assistance and their operational and interruption costs. Compared with other existing mathematical tools, ID enjoys several unique advantages. First, it provides a coherent and fully unified hierarchical probabilistic framework for representing and modeling the uncertain knowledge about user affect and assistance determination at different levels of abstraction. Second, within the unified framework, affect recognition is cast as a standard probabilistic inference procedure, while user assistance is formulated as a decision-making procedure. Third, it naturally incorporates the evolution of user affect and accounts for the temporal aspect of decision-making with the dynamic structure. Thus, such a model is an ideal candidate to accommodate the aforementioned challenges.

This paper intends to make contributions in both theory and applications. Theoretically, we provide a formal treatment of ID-based user modeling that addresses the theoretical foundations of affect recognition and automatic user assistance determination within a unified framework. In addition, an active sensing strategy is proposed to decide an optimal sensory action set to collect the best user measurements for efficient affect recognition and for timely decision-making for user assistance. Practically, based on the theoretical model, we develop a *non-invasive* and *real-time* prototype system that monitors two affective states—stress and fatigue. The system collects sensory data from four modalities: physical appearance, behavior, physiological measures, and performance. The system is *non-invasive* in that all the measurements are collected in a non-intrusive manner without interrupting the user.

The remainder of this paper is organized as follows. A brief literature review is presented in Section 2. Section 3 proposes the dynamic ID framework and Section 4 describes how this framework enables us to bridge between inferring affective states and deciding appropriate user assistance. Section 5 illustrates a simulation system to validate the proposed framework and Section 6 discusses a real-world user affect monitoring system and its validation. Finally, Section 7 concludes the paper with several future research directions.

2. Related work

In this section, we first review the related work in user affect modeling and recognition. This is then followed by a review of current work in user assistance.

2.1. User affect modeling and recognition

2.1.1. General approaches

In predicting and recognizing user affect, the methods can be classified as predictive inference (top-down), diagnostic inference (bottom-up), or a hybrid combining both predictive and diagnostic inference. For predictive inference, affect is recognized based on prediction using factors that influence or cause affect. A predictive approach usually rests itself on the established psychological theories. For instance, Ortony et al. (1988) defines emotions as valenced (positive or negative) reaction to situations consisting of events, actors, and objects. The valence of one's emotional reaction depends on the desirability of the situation, which, in turn, is defined by one's goals and preferences. The theory defines 22 emotions as a result of situation appraisal. If a person's goals and perception of relevant events are known, they are used to predict the person's emotions.

In contrast to a predictive approach, diagnostic approaches infer affect from physiological or behavioral measurements of the user. A rich body of literature has revealed the use of various features to infer user affect. In Kaapor et al. (2001), the authors discuss how to monitor eyebrow movements and body posture to provide evidence of students' engagement while interacting with a computer-based tutor. Heishman et al. (2004) propose to use eye region biometrics (including eyebrow, pupil, iris, upper/lower fold, and upper/lower eyelid) to reveal user affective (fatigue) and cognitive (engagement) states. In Ji et al. (2004), physical appearance features extracted from real-time videos are used to assess users' fatigue status. The work by Berthold and Jameson (1999) studies the effects of cognitive workload on two speech symptoms—sentence fragments and articulation rate. Cowie and co-workers (Cowie et al., 2001) develop a hybrid system capable of using information from faces and voices to recognize people's emotions.

To improve recognition accuracy, diagnostic and predictive methods may be combined. For example, affect-influencing factors such as task, environment, time of day, or user traits or physical conditions are often combined with physiological or behavioral data to provide a more consistent and accurate affect characterization. Most probabilistic approaches, to be surveyed later, belong to hybrid ones (Conati, 2002; Ji et al., 2004).

We can also classify the approaches in affect recognition based on the mathematical tools they used. The first group uses traditional classification methods in pattern recognition. The approaches include rule-based systems (Pantic et al., 2002), discriminate analysis (Ark et al., 1999), fuzzy rules (Elliott et al., 1999; Massaro, 2000; Hudlicka and McNeese, 2002), case-based and instance-based learning (Scherer, 1993; Petrushin, 2000), linear and nonlinear regression (Moriyama et al., 1997), neural networks (Petrushin, 1999), Bayesian learning (Qi et al., 2001; Qi and Picard, 2002; Kapoor et al., 2004) and other learning techniques (Heishman et al., 2004). Most of these research efforts focus on the low-level mapping between certain sensory data and the underlying affect. The mapping is often performed statically and independently, ignoring the history or current context that might influence the interpretation of user affective states. In fact, a common criticism of these approaches is their inadequacy in systematically representing prior knowledge, the dependencies among affect variables, the dynamics of affect, and in accounting for the uncertainties in both user affect and its measurements.

To overcome these limitations, the second group of approaches uses probabilistic graphical models such as Hidden Markov Models (HMMs), Bayesian networks (BNs), etc. With the aid of causal and uncertainty representation structure, these methods maintain a balance between global and local representations as well as provide powerful capabilities for handling complex situations in practical systems. HMMs have been used as a framework for recognizing the affective states (hidden) from observational data. For example, Picard (1997) uses HMMs to model the transitions among three affective states, namely, interest, joy, and distress. She also discusses the utility of HMMs for capturing environmental, cultural, or social context. Cohen et al. (2000) propose an HMM approach to recognize facial expressions and then classify user affect based on the recognized expressions. For each of the affective states studied, an HMM corresponding to an affective state is constructed. The features, based on the Facial Action Coding System (Ekman and Friesen, 1978), are extracted from the real-time videos. These measures are used to compute the posterior probability of a particular user affect. One problematic assumption made in the paper is that facial expression always reflects emotion. This assumption is unrealistic as facial expressions are ambiguous, therefore unable to uniquely characterize emotions. Yeasin et al. (2004) exploits HMMs to learn the underlying models for each universal expression. It is shown that HMMs can be used to accurately recognize six basic emotional facial expressions—surprise, happiness, sadness, fear, anger and disgust. The average recognition rate of the proposed facial expression classifier is 90.9%. The assumption behind this work is that facial expressions have a systematic, coherent, and meaningful structure that can be mapped to affective dimensions (Breazeal, 1999; Machleit and Enoglu, 2000). HMMs, however, lack the capability to represent dependencies and semantics at different levels of abstraction between emotion and the factors that cause emotion as well as the various observations that reflect emotion.

As a generalization to HMMs, BNs use graphical models to represent, at different levels of abstraction, the prior knowledge of affective states and the dependencies among user affect, the factors influencing affect and the observations reflecting affect. In Ball and Breeze (2000), a two-layer BN is created to model valence and arousal of users' affect during the interaction with an embodied conversational agent. The model uses measurements from linguistic behavior, vocal expression, posture and body movements. HMMs and BNs can be also combined within one framework. For example, Kaliouby and Robinson (2004) develop a real-time system for inferring six mental states, including agreement, concentrating, disagreement, interested, thinking, and unsure, from facial expressions and head gestures. The system consists of three levels: action unit analysis, facial and head display recognition, and mental state inference. The first two levels are implemented via an HMM approach. The output of HMMs is fed to a dynamic Bayesian network (DBN) for inferring user mental states. In summary, BNs are expressive in modeling conditional dependency and have been used for affect recognition. However, they do not explicitly model decisional choices and their utilities. Within the BN framework, decisions must be made separately and often in an ad hoc manner.

2.1.2. User stress and fatigue recognition

In this section, we especially review related work in recognizing human stress and fatigue since they are usually the significant factors causing a variety of human–machine interaction failures. Especially, these are the two affective states that we have experimented in our real-world system.

Human stress is a state of tension that is created when a person responds to the demands and pressures that arise from work, family, and other external sources, as well as those that are internally generated from self-imposed demands, obligations, and self-criticism. Although some stress is beneficial in certain circumstances, due to the adverse effects of excessive stress in our daily life, it is important to detect stress in a timely manner and treat it properly. In the past, researchers from different disciplines have developed inference approaches or pragmatic systems to recognize user stress level. The approaches or systems differ from each other in either the sensory modalities, or inference techniques, or both. In Healy and Picard (2000), a sequential forward floating algorithm (SFFS) is used to find an optimal set of features from the physiological measures (electrocardiogram, electromyogram, respiration, and skin conductance) and then the k-NN (nearest neighbor) classifier is applied to classify the stress into four levels. In Rani et al. (2003), after extracting physiological parameters from the measures of cardiac activity, electrodermal activity, electromyographic activity, and temperature, regression tree and fuzzy logic methodologies are used to classify human anxiety into 10 levels. A non-contact skin temperature measuring system is developed to evaluate stress in Kataoka et al. (1998), where only the skin temperatures on nose and forehead are measured. Rimini-Doering et al. (2001) combines several physiological signals and visual features (eye closure, head movement) to monitor driver drowsiness and stress in a driver simulator.

Over the years, many efforts have been made in the field of fatigue modeling and monitoring and the results are reviewed by Ji (2002) and Hartley et al. (2000). Traditionally, physiological measures have been widely used for fatigue detection. The popular physiological measures include the electroencephalograph (EEG) (Empson, 1986) and the multiple sleep latency test (MSLT) (Carskadon and Dement, 1982). EEG is found to be useful in determining the presence of ongoing brain activity and its measures have been used as the reference point for calibrating other measures of sleep and fatigue. MSLT measures the amount of time a test subject falls asleep in a comfortable sleep-inducing environment. Unfortunately, most of these physiological parameters are obtained intrusively, making them unacceptable in real-world applications. Thus, in recent years, there has been increasing research activity focused on developing systems that detect the visual facial feature changes associated with fatigue using a video camera. These facial features include eyes, head position, face or mouth. This approach is non-intrusive and becomes more and more practical with the rapid development of camera and computer vision technology. Several studies have demonstrated their feasibility and some of them claimed that their systems perform as effectively as the systems detecting physiological signals do (Saito et al., 1994; Ueno et al., 1994; Boverie et al., 1998; Grace, 2001). However, efforts in this direction are often directed to detecting a single visual cue such as eyelid movement. Since a single visual cue is often ambiguous, varies with time, environment or subjects, its validity is questioned (Heitmann et al., 2001). To overcome this limitation, it is necessary to combine multiple measures to produce more accurate fatigue-related performance estimation. Our real-world system works towards this goal as will be detailed later.

2.2. User assistance

Appropriate and timely user assistance is crucial for a HCI system. Ensuring that the intervention will be as welcomed as it will be valuable and timely is an important research issue. Here, we briefly review current methodologies in user assistance.

In Li and Ji (2004), a DBN is proposed to recognize user affect by modeling multiple visual appearance variables. If the user affect level exceeds a pre-determined threshold, certain assistance is provided. The reported experiments show that the framework works well with synthetic data. Unfortunately, a limitation of this approach is that the threshold is manually set and therefore needs human intervention. This is different from the proposed work where the user assistance is automatically generated, based on the utility of assistance.

Murray et al. (2004) describe a decision-theoretical approach based on dynamic decision network for selecting tutorial actions while helping a student with a task. The theoretical schema is specifically designed for two application domains: calculus-related rate problems and elementary reading. These applications exploit a rich model of the tutorial state, including attributes such as the students' knowledge, focus of attention, affective state, and next action(s), along with task progress and the discourse state. Via an extensive simulation, their work focuses on evaluating whether the selected tutorial action is rational and fast enough under a variety of tutorial situations. Both Murray's and our work decide optimal actions with the maximal expected utility and exploit the temporal properties, although Murray's work looks multiple steps ahead while ours looks one step ahead. Murray's work and ours are different in several ways. First, Murray's work is more concerned with how an optimal tutorial action can be selected given the tutorial state instead of studying how to recognize the tutorial state, such as the student's affective states, focus of attention, etc.; while our work focuses on both how to automatically and efficiently recognize users' affective states from multiple-modality measurements and on how user assistance can be timely and appropriately applied within the integrated framework. More specifically,

instead of developing a component of user affect recognition system, we develop an integrated system. This involves development of both the sensing system and the inference engine as well as their systematic integration into a prototype system. Second, Murray et al. systematically evaluate the effectiveness of the selected tutorial actions through simulations; we also evaluate the effectiveness of the user assistance through simulations, but the evaluation is limited to the effect of the assistance on the users' affective states. For our studies, a real-time and non-invasive prototype system is built based on the proposed framework for, respectively, recognizing human fatigue and stress using sensors of different modalities. In addition, real-world experiments are conducted to validate the affect recognition part of the system.

In Conati (2002), a dynamic decision network based on ID is used in pedagogical agents to monitor a user's emotions and to generate interventions aimed at achieving the best tradeoff between the user's learning and engagement during their interaction with educational games. Like ours, in inferring user affective states, the work accounts for both the possible causes of the user's emotional arousal and its effects such as bodily expressions. Also, the intervention is decided by maximizing expected utility. However, there are apparent differences between their work and ours. First, our method integrates affect recognition and intervention in a unified framework, allowing affect recognition and intervention to be performed simultaneously. Second, we consider a larger evidence set. Their work uses only bodily expression-related features, while our work utilizes physical appearance, physiological, behavioral measures, and performance data. Third, their work assumes user measurements that are already provided (such as facial expressions, etc.), while we develop methods to obtain various measurements in our system.

In Horvitz et al. (2003), a decision-theoretic model is proposed to help a user in choosing actions by inferring her/his attention, a concept closely related to user affective states. The attention focus is inferred from the observed states of several sensors such as microphones, cameras, accelerometers and location sensing facilities. The attentional focus is also reflected by the user's interactions with software and devices, background information about the history of the user's interests, and prior patterns of activities. The work differs from our work in that it emphasizes performing attention inference mostly from desktop activities, while ours emphasizes performing affect inference from sensory evidence of different modalities.

In summary, there are numerous efforts in developing systems for user affect recognition and for user assistance. They tend to focus on either affect recognition or user assistance. In addition, the sensory modality that is used for user state measurements is often limited and the sensory measurement acquisition is often intrusive. Our research intends to build an integrated system that can simultaneously fulfill the two functions: affect recognition and user assistance. Compared with the cited ones, the significance of our framework is that it employs dynamic inference and sequential decision-making techniques to unify affect recognition with user assistance, utilizes evidence from multiple modalities and acquires them in real-time and in a non-intrusive manner, and applies active sensing strategies for selecting most informative measurements for efficient and timely affect recognition. In addition to validating the proposed framework with a simulation system, the affect recognition component is validated with a real-world study.

3. A unified framework for modeling affect and user assistance

3.1. Influence diagrams

Since IDs were introduced by Howard and Matheson in 1981 (Howard and Matheson, 1981), it has been widely used as a knowledge representation framework to facilitate decision and probabilistic inference problems under uncertainty. An ID is a directed acyclic graph consisting of nodes and the directed links between nodes. Nodes are grouped into decision nodes, chance (random) nodes, and utility nodes. Decision nodes, usually drawn as rectangles, indicate the decisions to be made and their set of possible alternatives. Chance nodes, usually drawn as circles, represent uncertain variables that are relevant to the decision problem. Utility nodes, usually drawn as diamonds, are associated with utility functions to represent the utility of each decision. The arcs connecting different types of nodes have different meanings, based on their destinations. Arcs among chance nodes represent the probabilistic dependencies among the connected nodes while arcs between decision nodes represent time precedence. Arcs connecting to utility nodes represent the value influence. Fig. 1 gives an ID example.

The top node Θ indicates the target hypothesis variable, for example, it could be affective states. Each bottom node, $E_1, \ldots, E_n$, indicates the possible observations of the target

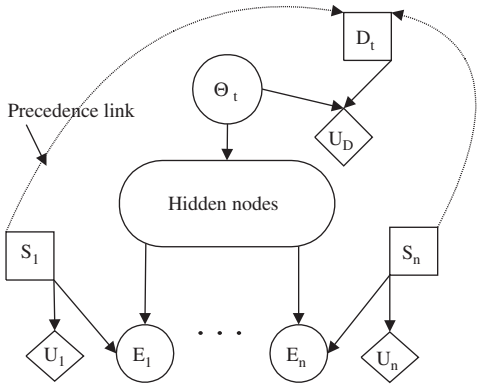

Fig. 1. An ID example. Chance nodes are represented by ellipse, decision nodes are represented by rectangles, and utility nodes are represented by diamonds.

variable. The big ellipse indicates all the chance nodes between the E nodes and Θ node. These nodes are collectively called hidden nodes. They model the probabilistic relationships between the Θ node and E_i nodes with different abstraction levels. The decision node D indicates the possible actions associated with the hypothesis node Θ while each decision node S_i controls whether to obtain observations from an information source or not. Each utility node U connected with S_i defines the cost of obtaining data from the information source. The utility node connected with both nodes Θ and D indicates the benefit (penalty) of taking appropriate (inappropriate) actions with respect to a particular target hypothesis state. In addition to the semantic relationships, an ID need be parameterized. This involves quantifying each chance node with a probability matrix describing the conditional probability of the node given all possible outcomes for its parent(s), and each utility node a utility function. Overall, such an ID can be used to decide a best evidence set to reduce the uncertainty of the hypothesis Θ as well as to decide a best decision D associated with such a hypothesis.

In summary, IDs use an acyclic directed graph representation to capture the three diverse sources of knowledge in decision-making: conditional relationships about how events influence each other in the decision domain, informational relationship about what action sequences are feasible in any given set of circumstances, and functional relationships about how desirable the consequences are (Pearl, 1988). The goal of ID modeling is to choose a decision alternative maximizing the expected utilities. We call this decision alternative as an optimal policy and call the maximized utility as optimal expected utility. Evaluating an ID is to find such an optimal policy as well as compute the optimal expected utility (Shachter, 1986).

3.2. Modeling affect recognition and user assistance with influence diagrams

In this section, we present our framework based on ID for simultaneously modeling affective states and user assistance. The framework actually has a similar structure to the example ID in Fig. 1. We discuss the details of the proposed framework in both qualitative part (the structure, the various nodes and their links) and quantitative part (conditional probability distributions and utility functions).

Central to user affect modeling are affective states, the measurements used to infer user affect, and the user assistance that can be provided. In addition, a complete model ought to include the factors that influence user affect, and the sensory nodes that enable evidence collection. In our model, the following components constitute the affect detection and the user assistance system: a set of user affective states, a set of external assistance that may alter user affective states, a set of user state measurements (also called evidence), a set of conditional probability distributions that characterize the conditional dependencies among related variables, and a set of utility functions that characterize the benefits or costs of performing assistance/sensory actions. An ID implementation of these components is shown in Fig. 2.

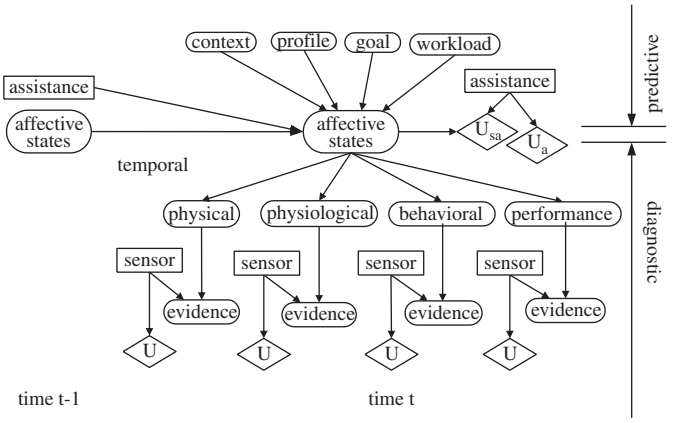

Fig. 2. A generic influence diagram for user assistance and affective state recognition. For simplicity, we show the dynamic ID at time t but only draw the "affective states" and "assistance" nodes at time $t-1$. Ellipses denote chance nodes, rectangles denote decision nodes, and diamonds denote utility nodes.

The schematic diagram captures the information necessary for two purposes: providing user assistance and recognizing affective states. The upper, *predictive portion*, of the diagram depicts contextual factors that can alter affective states. Such elements include environmental context, user profile, goal that the user is pursuing, workload, etc. The lower, *diagnostic portion*, depicts the observable features that reflect user affective states. These features may include quantifiable measures on physical appearance, physiology, behaviors, and performance. The left, *temporal portion*, models the temporal evolution of user affect and sequential decision-makings on user assistance. The inference of user affect and the determination of the appropriate assistance based on integrating the predictive, diagnostic, and temporal inference is more accurate and consistent than any of them alone.

3.3. Model description

This subsection describes the qualitative part of the model, namely the affective states, evidence, actions and utilities.

3.3.1. Affective states

An affective state is an integration of subjective experience, expressive behavior, and neurochemical activity. We focus on negative affective states such as stress, fatigue, and confusion since they can adversely alter users' productivity and negatively affect their decision-making and learning abilities. An affective state has a set of possible values. For instance, stress may vary from low to normal and to high. Naturally, affective states are not

mutually exclusive. For example, a subject can be both fatigued and stressed. Accordingly, if multiple affective states are of interest, each of them needs to be represented by a separate random node in the model.

3.3.2. Factors influencing affective states

User affective states could be affected by a variety of factors. These factors may include the environmental context, the user profile (or personal traits), the workload, and importance of the goal the user is pursuing. The environmental context reflects the exterior impact from the outside situation on user affective states. The workload and the importance of the goal lead to interior influence on a user (Karasek, 1979; Ortony et al., 1988; Jones and Bright, 2001). And the profile information may include age, experience, skills, health, etc., which plays an important role in adapting the model to individual differences among users. Thus, these factors are represented as parent variables of the "affective states" node and form the *predictive* portion of the model.

3.3.3. Sensory measurements

An evidence is an observable feature that is capable of providing clues about the user's internal affective state. We consider four classes of measurable evidence: the user's physical appearance features, physiological measures, behavioral characteristics, and performance measures.

Physical appearance evidence includes the visual features that characterize user's eyelid movement, pupil movement (eye gaze movement, pupillary response, etc.), facial expression, and head movement. These features have a systematic, coherent, and meaningful structure that can be mapped to affective states (Beatty, 1982; Breazeal, 1999; Machleit and Enoglu, 2000; Partala and Surakka, 2003). Specifically, eyelid movement can be characterized by average eye closure speed, blinking frequency, etc.; pupil movement can be characterized by gaze spatial distribution and fixation, pupil dilation, pupil size variation, etc.; facial expression can be happy, sad, angry, scary, etc.; and head movement can be characterized by head pose, tilting frequency, etc. The importance of these features may vary with different affective states. For example, head tilting frequency is a useful feature for identifying fatigue, while it may not be effective for recognizing stress. An ID diagram modeling the physical appearance evidence for estimating human affective states is shown in Fig. 3. In the figure, the "physical" node is added as an intermediate node between "affective states" and the physical appearance evidence in order to model the correlations among the evidence. The intuition is that the user affect influences his physical status, which, in turn, influences the physical appearances such as eyelid, gaze, etc. The "physiological" and "behavioral;; nodes are added in Figs. 4 and 5, respectively, for similar reasons.

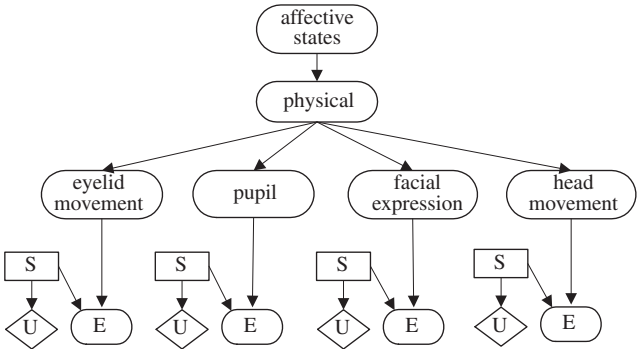

Fig. 3. ID modeling for physical appearance evidence. *E* represents an evidence node; *S* represents a sensor node or a sensing algorithm; and *U* represents a utility node.

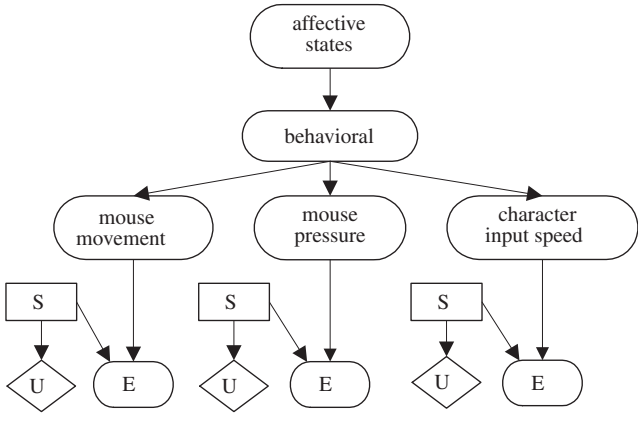

Fig. 5. ID modeling for behavioral evidence.

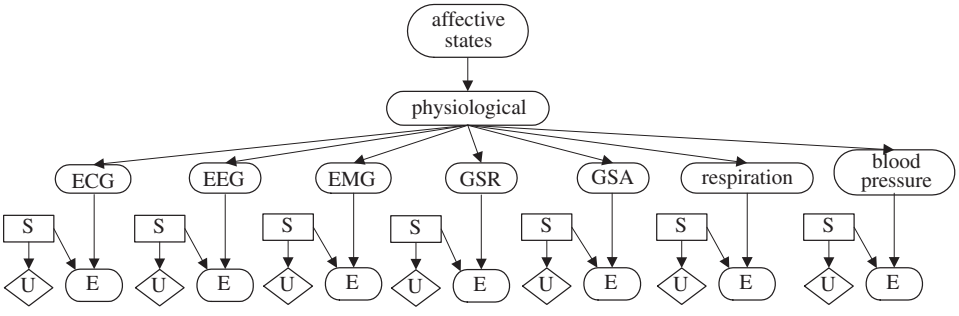

Fig. 4. ID modeling for physiological evidence.

The physiological variables provide physiological evidence about user affect (Gardell, 1982; Picard, 1997; Jones and Bright, 2001). The physiological evidence can be electromyography (EMG) measures that assess the electrical signal generated by muscles when they are being contracted, electrocardiograph (ECG) measures that assess the electrical pulse of the heart, galvanic skin response (GSR) measures that assess the electrical properties of the skin in response to different kinds of stimuli, general somatic activity (GSA) measures that assess the minute movement of human body, electroencephalography (EEG) measures that assess the electrical activity produced by the brain and many others such as respiration and blood pressure. A typical approach to modeling the physiological evidence is shown in Fig. 4.

User behavior may be influenced by user affective states. For example, an emotional user might press the keyboard heavily and irrationally. The behaviors in question may be patterns of interaction between user and computer, e.g. the interaction between user and the mouse or the keyboard. For this research, we monitor several user behaviors including mouse/keyboard pressing pressure, mouse click frequency and speed, character input speed from the keyboard, and cursor movement speed and pattern, etc. An ID that models behavioral evidence is shown in Fig. 5.

Performance can vary with user affective states and therefore may indicate user affect. In a task specific environment, the performance may be accuracy rate, user response time, or other measures derived from these two. As an instance of affective state influencing user performance, Karasek (1979) demonstrated that occupational stress was affected by the tasks presented to a user.

3.3.4. Assistance, sensory actions, and utilities

In addition to random nodes, an ID has decision nodes and utility nodes. Two types of decision nodes are embedded in the proposed framework. The first type is the assistance node associated with the affective state node. Assistance actions may have different degrees of intrusiveness to a user. For example, in one extreme, the assistance can be null if the user is at positive affective states; in the other extreme, the user may be interrupted and forced to quit if he is in a dangerous level of negative affective states. Some typical assistance could be "warning" (friendly inform the user his negative affective states), "alleviating" (simplify user interface, decrease task difficulty, etc.), "intervening" (stop user from work), and etc. How to design appropriate assistance should depend on the applications. Another type of decision node is the sensory action node. It controls whether to activate a sensor for collecting evidence or not.

Corresponding to the decision nodes, there are three types of utility nodes. The utility node associated with the assistance node denotes the physical cost of performing assistance. The utility node associated with both affective states and assistance node indicates the benefit (penalty) of taking appropriate (inappropriate) assistance with respect to a particular user state with respect to a particular user state. The utility node associated with a sensory node denotes the cost of operating the sensor for evidence collection. More details will be given in Section 3.4.

3.3.5. Dynamics

While a static ID only models the static aspect of affect states and user affect evolves over time, it is important to model the temporal evolution of affect and the sequential decision-making on user assistance with a dynamic ID. In general, a dynamic ID is made up of interconnected time slices of static IDs, and the relationships between two neighboring time slices are modeled by an HMM, i.e. random variables at time t are affected by variables at time t, as well as by the corresponding random variables at time $t-1$. Each time slice is used to represent the snapshot of an evolving temporal process at a time instant. These time slices are interconnected by the arcs joining particular temporal variables from two consecutive slices.

For modeling affect and assistance, the temporal nodes include the affect node and the assistance node. The temporal links between the temporal affect nodes in two consecutive slices represent temporal evolution of the user state over time, with the nodes at time $t-1$ providing a diagnostic support for the corresponding nodes at time t. Specifically, the temporal state node Θ_{t-1} at time $t-1$ provides a diagnostic support for the affect node Θ_t at time t. And the temporal link from the assistance node at $t-1$, D_{t-1}, to current state Θ_t indicates the assistance applied in the previous step may influence affective state in the current step.

3.3.6. The complete model

We are now ready to combine the above components into a complete model for affect recognition and user assistance. Combining physical appearance (Fig. 3), physiological features (Fig. 4), behavioral features (Fig. 5), performance features, and the dynamic, and the dynamics into the schematic graph (Fig. 2), we obtain a complete graph in Fig. 6. Overall, the complete model consists of the predictive portion that models the contextual factors that can alter/cause user affective states, the diagnostic portion that models the observable features that reflect user affective states, and the temporal portion that models the evolution of affective states and sequential decision-making on user assistance. In addition, it consists of two types of decision nodes, the assistance node for providing appropriate assistance to maintain user in positive affective states, and the sensory action nodes for controlling whether to activate sensors for collecting valuable evidence. Such a model demonstrates a coherent and fully unified hierarchical structure for representing and integrating various variables that are closely related with user affective states and assistance.

In addition, the generic model is flexible to allow variables to be inserted and modified. For example, the random variables under the behavioral node may vary with

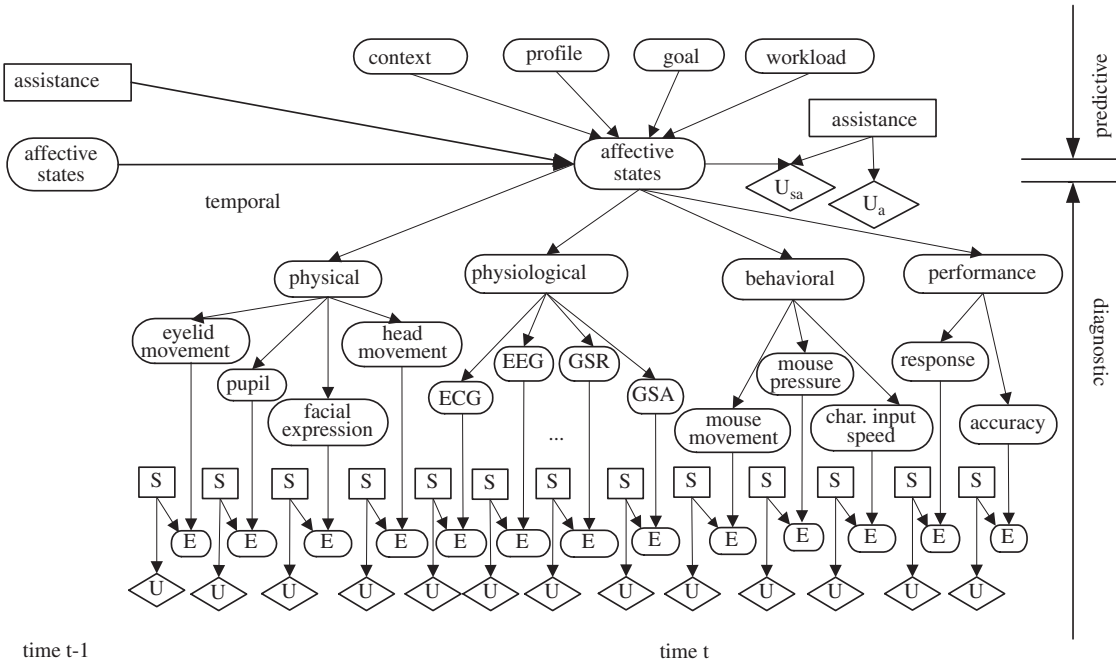

Fig. 6. The complete influence diagram for modeling user assistance and affect. Note that some physiological features are omitted due to space limit.

different applications, the random variables under the physiological node may change, depending on the availability of the required measuring devices. Furthermore, variables can be inserted between the evidence (E) nodes and their parent nodes to model the reliability of the corresponding sensors. Fig. 7 summarizes some of the discrete variables and their states in the model.

3.4. Model parameterization

This subsection introduces the quantitative part of the model, namely the conditional probability tables (CPTs) and utility functions.

3.4.1. CPTs

Once the topology of the dynamic ID is specified, the next step is to quantify the relationships between connected nodes—this is done by specifying a conditional probability distribution for each chance node and a utility function for each utility node. As we are only considering discrete variables, the conditional dependencies among the variables are characterized by CPTs. For each random node, given a value assignment of its parents, a CPT is a probability distribution over all possible values of the random node. If the node has no parents, a CPT degenerates to a priori distribution. For example, for the node "workload", $p(workload)$ denotes the prior distribution of the variable "workload" over the values "low", "normal", and "high". For an evidence node (say E linking to EEG), the CPT $p(E|EEG, Sensor = on)$ is a distribution of the variable E over the measurable range of EEG given the EEG and that *sensor* is turned on.

Similarly, if a node has a temporal link, its CPT characterizes the conditional dependence of the node on its parent nodes, some of which come from the previous time step. Let us take the affective state node as an example. Its CPT $p(affect|previous\_affect, previous\_assistance, context, profile, goal, workload)$ denotes a probability distribution over the current affective states for each value assignment of the parent nodes. In particular, if the assistance is null, the CPT describes how the affective state evolves over time given the values of other parent variables; if the user is provided with an assistance, the CPT describes how the assistance alters the user affective state given the values of other parent variables. Fig. 8 gives some CPT examples.

In general, CPTs are obtained from statistical analysis of a large amount of training data. For this research, the CPT parameters come from two sources. First, we refer to several large-scale subjective surveys and domain experts (Ortony et al., 1988; Picard, 1997; Rosekind et al., 2000; Sherry, 2000; Picard et al., 2001; Zhang and Ji, 2005a) to obtain initial CPTs. An experienced expert can often produce rather accurate estimates of local probabilities. For the case of a random node that has multiple parent nodes, certain causal-independence assumptions (e.g. noisy-or or generalized noisy-or principle, Diez, 1993; Zhang and Poole, 1996; Lemmer and Gossink, 2004) may be used to simplify the parameterization. Second, we obtain training data from the human subjects study we conducted. These data are then used to train the ID model with the existing learning algorithms (Buntine, 1994; Lauritzen, 1995; Jordan, 1999) such as the EM method (Lauritzen, 1995). With the learning algorithm, the initial

Component	Variables	States	Description
context	context	complex/simple	The surrounding environment of the subject
profile	health age skill …	good/normal/bad old /young strong/weak …	Personal information of the subject
goal	goal	important/not important	The importance of successfully finishing the tasks to the subject
workload	workload	high/normal/low	
affective states	stress fatigue nervous …	positive/negative positive/negative positive/negative …	
physical	physical	high/normal/low	Physical state of the subject
physiological	physiological	high/normal/low	Physiological state of the subject
behavior	behavioral	normal/abnormal	
performance	performance	good/normal/bad	
eyelid movement	BF AECS	high/normal/low fast/normal/slow	Blinking frequency Average eye closure speed
pupil	PerSac GazeDis PerLPD PRV	large/normal/small normal/abnormal large/normal/small large/normal/small	Percentage of saccadic eye movement Gaze spatial distribution Percentage of large pupil dilation Pupil ratio variation
facial expression	facial expression	neutral, happiness, sadness, anger, surprise, disgust, fear	
head movement	head movement	normal/abnormal	
ECG	ECG	normal/abnormal	Electrical pulse of the heart
mouse pressure	mouse pressure	high/normal/low	
response	response	fast/normal/slow	
accuracy	accuracy	good/normal/bad	
assistance	assistance	null/warning/alleviating /intervening	
sensor	S	on/off	on: evidence is collected off: no evidence is collected

Fig. 7. Some variables and their states in the ID model.

CPTs are automatically refined to match each individual subject.

3.4.2. Utility

When deciding upon the user assistance and sensory actions, we need to consider our preferences among the different possible outcomes of the available actions. The ID uses utility functions associated with utility nodes to provide a way for representing and reasoning with preferences. A utility function quantifies preferences, reflecting the "usefulness" (or "desirability") of the outcomes, by mapping them to real numbers (Kevin and Ann, 2003). Specifically, for the proposed framework, utility functions are defined according to the three types of utility nodes. A utility node associated with the assistance node only denotes the physical and interruption cost of performing that assistance. We would assign a higher cost to actions that require more resources and time, or interrupt the user more. The cost would map to negative values in the utility function. A utility node associated with both assistance and affective state nodes denotes the benefit (penalty) of providing appropriate (inappropriate) assistance with respect to user affective state. For example, if a user is very stressed, an appropriate assistance may be to reduce workload, while an inappropriate assistance would be to let the user continue his work or increase task difficulty. Thus the former should be assigned a high positive value while the latter should be given a low negative value in the utility function. A utility node associated with a sensory node denotes the cost of operating the sensor for evidence collection. The cost includes the computational cost, physical cost, etc. For example, when a device is used to measure EEG, there is a cost for operating the device as well as analysing the data. We want to emphasize that all the utilities need to be calibrated with a certain calibration standard so that they can be appropriately used in the same ID framework.

4. Affect recognition and user assistance

Given the parameterized dynamic ID framework, this section focuses on the techniques for recognizing affective states and determining user assistance.

4.1. Overview

Fig. 9 outlines the main steps in applying the dynamic ID to affect recognition and user assistance. This is achieved

Node Name	Parent Node	Parent Node State	Child Node State	Condition Probability
physical	stress	negative	low	0.25
		negative	normal	0.60
		negative	high	0.15
		positive	low	0.15
		positive	normal	0.20
		positive	high	0.65
performance	stress	negative	bad	0.20
		negative	normal	0.30
		negative	good	0.50
		positive	bad	0.70
		positive	normal	0.20
		positive	good	0.10
AECS	eyelid movement	slow	slow	0.80
		slow	normal	0.15
		slow	fast	0.05
		normal	slow	0.35
		normal	normal	0.45
		normal	fast	0.20
		fast	slow	0.15
		fast	normal	0.20
		fast	fast	0.65
PerLPD	pupil	normal	large	0.15
		normal	normal	0.82
		normal	small	0.03
		abnormal	large	0.68
		abnormal	normal	0.10
		abnormal	small	0.22
mouse pressure	behavioral	normal	low	0.35
		normal	normal	0.55
		normal	high	0.10
		abnormal	low	0.10
		abnormal	normal	0.20
		abnormal	high	0.70

Fig. 8. Some examples of CPTs. Note for different subjects, the CPTs may be different.

through ID evaluation. ID evaluation starts from the current user affect, which is influenced by three streams of information: one from the previous user state estimate, one from the available contextual information, and one from the performed user assistance at the previous step. Then, an active sensing strategy is employed to decide an optimal sensor set S_t^* as well as whether the sensors in S_t^* are worth activating or not. If the sensors in S_t^* will benefit the decision-making on user assistance and affect recognition, these sensors are activated and new evidence E_t^* are collected. The collected evidence is then propagated to update user affect and the posterior probabilities of user stress p_t is computed with the dynamic inference technique. If no new sensors are activated, the user affect is updated with only the temporal information from the affective states at the previous time step. Based on the updated estimate of user state, the system determines the optimal assistance d_t^* that maximizes the overall expected utility. After the assistance is performed, the user state may change and new evidence may need to be collected. Thus the system goes to the next step and repeats the ID evaluation procedure. With this systematic procedure, our model is capable of deliberately choosing optimal sensors and avoiding unnecessary sensory actions, efficiently recognizing user affective states, and providing timely and appropriate assistance.

The procedure of deciding optimal sensors and user assistance is actually to find an optimal policy for the ID model. A policy in the model is $\Delta = (s_1, \ldots, s_n, d)$ consisting of one rule (value) for each decision node, where s_i is a value for each sensory node S_i, and d is a value for the assistance node. An optimal policy Δ^* is the one that can maximize the overall expected utility EU:

$$\Delta^* = \arg\max_d EU_\Delta, \tag{1}$$

$$EU_\Delta = \sum_{\text{aff}} P_\Delta(\text{aff}) g_{u_{\text{sa}}}(\text{aff}, d) + P_\Delta(d) g_{u_{\text{a}}}(d) + \sum_i P_\Delta(s_i) g_{u_i}(s_i), \tag{2}$$

where aff indicates the affective state node, $g_{u_{\text{a}}}$, $g_{u_{\text{sa}}}$, g_{u_i} are the utility functions for the utility nodes associated with the assistance node, both the affective state and

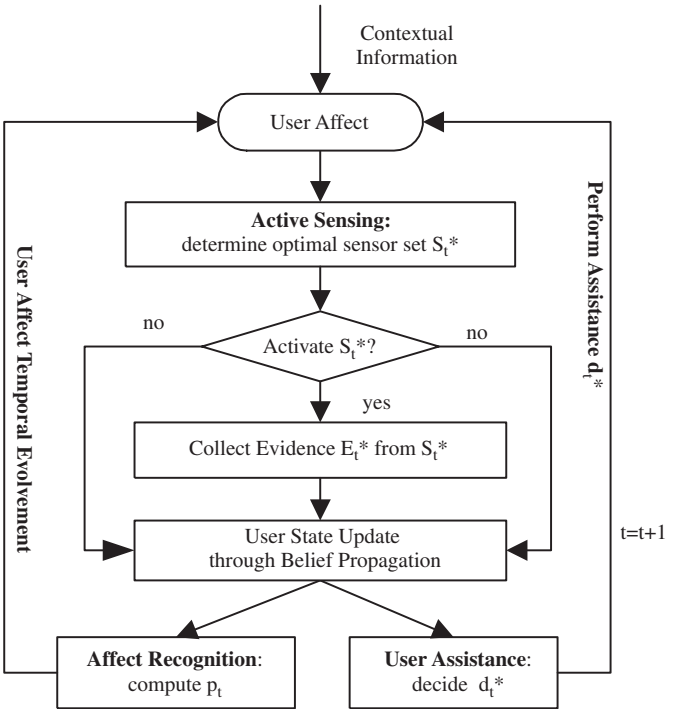

Fig. 9. The flowchart for affect recognition and user assistance.

assistance node, and the sensor nodes, respectively. Considering the special structure of the ID model and time precedence of the decision nodes, finding the optimal policy is to decide the optimal sensory actions first with the active sensing strategy, and then decide the optimal decision for user assistance. Such an optimal policy decides a timely and appropriate assistance using a sensor set that has the best tradeoff between cost and benefit.

4.2. Active sensor selection

Although a lot of sensors are available in the system, the usage of more sensors incurs more cost for acquiring information. For making a timely and efficient decision on user assistance as well as efficiently recognizing user affect, it is important to avoid unnecessary or unproductive sensory actions. This is accomplished with active sensing. By deliberately choosing sensors, active sensing attempts to identify a set of sensors that achieve the optimal tradeoff between the benefit and cost of the sensors. The benefit is that the evidence collected by the sensors may provide informative knowledge for decision-making on user assistance and user affect recognition, e.g. it may change the choice of optimal assistance, reduce the uncertainty of user affect, etc. The cost includes both physical and computational cost. Both the cost and benefit are quantified by the utility functions in the ID model. Thus, in what follows, we exploit utility theory to derive a criterion for active sensing.

We use value-of-information (VOI) to guide the selection of sensors. The VOI of a sensor set is defined as the difference in the maximum expected utilities with and without the information collected from this set. It evaluates the valuableness of a sensor set by considering both the benefit and cost of using the sensors. Let S be a sensor set, the VOI of S, $VOI(S)$, for the specific ID, can be defined as

$$VOI(S) = EU(S, Co) - EU(\bar{S}) \tag{3}$$

$$= EU(S) - Co - EU(\bar{S}), \tag{4}$$

$$EU(S) = \sum_{E} P(E) \max_{d} \left[\sum_{\text{aff}} P(\text{aff}|E) g_{u_{\text{sa}}}(\text{aff}, d) + g_{u_{\text{a}}}(d) \right], \tag{5}$$

$$EU(\bar{S}) = \max_{d} \left[\sum_{\text{aff}} P(\text{aff}) g_{u_{\text{sa}}}(\text{aff}, d) + g_{u_{\text{a}}}(d) \right], \tag{6}$$

where $Co = \sum_{S_i \in S} g_{u_i}(S_i = 1)$ is the total cost activating the set of sensors and E is the evidence collected from the sensor set S. $EU(S, Co)$ denotes the expected utility to the decision maker should the sensors in S be activated. We assume the delta property holds. The delta property states that an increase in value of all outcomes in a lottery by an amount Δ increases the certain equivalent of that lottery by Δ (Howard, 1967). Thus, $EU(S, Co) = EU(S) - Co$. $EU(S)$ denotes the expected utility to the decision maker, with its cost set to zero. And $EU(\bar{S})$ denotes the expected utility to the decision maker, without activating the sensor set S. If the VOI of a sensor set is positive, it means the expected benefit caused by the sensor set is larger than the cost. Also, it means the information collected by activating the sensors may change the choice of optimal assistance. An optimal sensor set S^* is the one that has the maximum VOI. However, if the maximum VOI is negative or zero, no sensors will be activated. The benefit of such a criterion is that sensors are activated and used only if they could benefit current decision-making on user assistance, therefore avoiding unnecessary sensory actions.

To get an optimal sensor set, we can enumerate all the sensor combinations and compare their VOIs. However, it is impractical to find the optimal set in this way since the number of sensor combinations is increasing exponentially as the number of sensors. In practice, we use a greedy approach by ranking each individual sensor based on its VOI and then pick the first m sensors, where m is empirically decided. Our experiments show that the selected sensor set with this approach is usually the one with maximum VOI among all the sensor sets whose size is not larger than m. We are working on developing more sophisticated active sensing approaches in more general case (Zhang and Ji, 2005b).

4.3. Optimal user assistance

After deciding the optimal sensor set, the values of the policy $\Delta = (s_1, \ldots, s_n, d)$ are fixed except the d part. Deciding the optimal user assistance is to find d_t^* achieving the optimal tradeoff between the cost and benefit of the assistance given the evidence collected from the optimal sensor set S_t^*. The cost of an assistance may include operational cost, interruption cost, and the cost of delaying or not providing the assistance. The benefit of an assistance is characterized by its potential to return the user to a productive affect state. Let e_t^* be the evidence collected after activating the sensors in the optimal sensor set, the optimal assistance d_t^* can be decided as follows:

$$d_t^* = \arg\max_d EU_{d_t}, \tag{7}$$

$$EU_{d_t} = \sum_{\text{aff}_t} P(\text{aff}_t|e_t^*, d_{t-1}^*) g_{u_{\text{sa}}}(\text{aff}_t, d_t) + g_{u_{\text{a}}}(d_t), \tag{8}$$

where the sum is taken over every possible value aff of the user state. The quantity EU_{d_t} balances the benefit/cost of taking appropriate/inappropriate assistance (the first term), and the cost (the second term) of performing assistance. The optimal assistance d_t^* is the one that maximizes EU_{d_t} among all available assistance. Please note that one of alternatives of d is null assistance. Hence no assistance will be provided if the null assistance alternative has the maximum expected utility.

Once d_t^* is performed, it will alter user affect state unless $d_t^* = \text{Null}$; in this case, the user affective state will evolve naturally. The steps repeat as shown in Fig. 9.

4.4. Affect recognition

Affect recognition is to estimate user affect from the evidence collected from the selected sensor sets using the dynamic inference technique. The system tries to estimate the user affect at each time step t. We first introduce the notations and then define the problem. We shall use the first one or several characters of a node name to refer to the node, i.e. w referring to *workload*, aff referring to *affective states*. In addition, we subscript a variable by a step t to refer to the variable at time t, i.e. aff_t for affective states node at time t. Under these notations, the ID model specifies two probabilistic relationships: the user affect transition model $p(\text{aff}_t|\text{aff}_{t-1}, w_t, c_t, pro_t, g_t, d_{t-1}^*)$ (d_{t-1}^* denotes the assistance at time $t-1$) and the evidence generation model $p(e_t^*|\text{aff}_t)$, where e_t^* is the set of evidence observed at step t. The inference at step t is to calculate the probability $p(\text{aff}_t|e_{1:t}^*, d_{t-1}^*)$. In case $t = 0$, $p(\text{aff}_t|e_{1:t}^*, d_{t-1}^*)$ degenerates to the prior $p(\text{aff}_0)$.

From a Bayesian point of view, the task is to compute $p(\text{aff}_t|e_{1:t}^*, d_{t-1}^*)$ from $p(\text{aff}_{t-1}|e_{1:t-1}^*, d_{t-2}^*)$ recursively. The task can be accomplished in two stages: prediction using the predictive portion of the ID and correction using the diagnostic portion. In the prediction stage, the prior probability $p(\text{aff}_t|e_{1:t-1}^*, d_{t-1}^*)$ of user affect at step t is calculated as follows:

$$\begin{aligned} p(\text{aff}_t|e_{1:t-1}^*, d_{t-1}^*) = \sum_{\text{aff}_{t-1}, w_t, c_t, pro_t, g_t} & \{p(w_t)p(c_t)p(pro_t)p(g_t) \\ & \times p(\text{aff}_{t-1}|e_{1:t-1}^*, d_{t-2}^*) \\ & \times p(\text{aff}_t|\text{aff}_{t-1}, w_t, c_t, pro_t, g_t, d_{t-1}^*)\} \end{aligned} \tag{9}$$

In the correction stage, the evidence set e_t^* is used to update the prior $p(\text{aff}_t|e_{1:t-1}^*, d_{t-1}^*)$ by Bayes' rule:

$$\begin{aligned} p(\text{aff}_t|e_{1:t}^*, d_{t-1}^*) &= \frac{p(e_t^*|\text{aff}_t)p(\text{aff}_t|e_{1:t-1}^*, d_{t-1}^*)}{p(e_t^*|e_{1:t-1}^*, d_{t-1}^*)} \\ &= \frac{p(e_t^*|\text{aff}_t)p(\text{aff}_t|e_{1:t-1}^*, d_{t-1}^*)}{\sum_{\text{aff}_t} p(e_t^*|\text{aff}_t)p(\text{aff}_t|e_{1:t-1}^*, d_{t-1}^*)}. \end{aligned} \tag{10}$$

5. Experiments with synthetic data

In order to validate the proposed framework, we have built a simulation system. We first report experiments with the focus on one affective state, and then extend it to the multiple affective states case.

5.1. Simulation system

We develop a simulation system that simulates a so-called *truthful user*, represented by a *source model*, and an *observed user*, represented by a *working model*. Both the working model and the source model have the same structure and parameters as the ID model in Fig. 6, while they perform different functions. The source model produces evidence reflecting the true affective states and accepts user assistance, whereas the working model accepts the perturbed evidence, uses the evidence to estimate user affect, and determines what assistance to provide.

Fig. 10 illustrates how the simulator works. Initially, the source model begins with a probability distribution for the affective state representing *true user affect*. Next, the source model generates a set of evidence through a top-down inference. Based on the VOI criterion (Section 4.2), the working model determines a set of sensors, collects a set of perturbed evidence generated from the source model, decides optimal user assistance d^* (Section 4.3), and estimates the observed user affect (Section 4.4). In the meanwhile, the source model performs the user assistance d^* to update the true affective state. The simulation procedure repeats at the next time step.

Actually, the relationship between the source model and working model is similar to the relationship between a patient and a doctor. A doctor (a working model) decides what tests (sensors) to perform on a patient, makes a diagnosis (estimates user affect), and treats the patient (decides user assistance). A patient (a source model) shows symptoms (evidence) because of some illness (true user affect), and accepts treatment (receives user assistance).

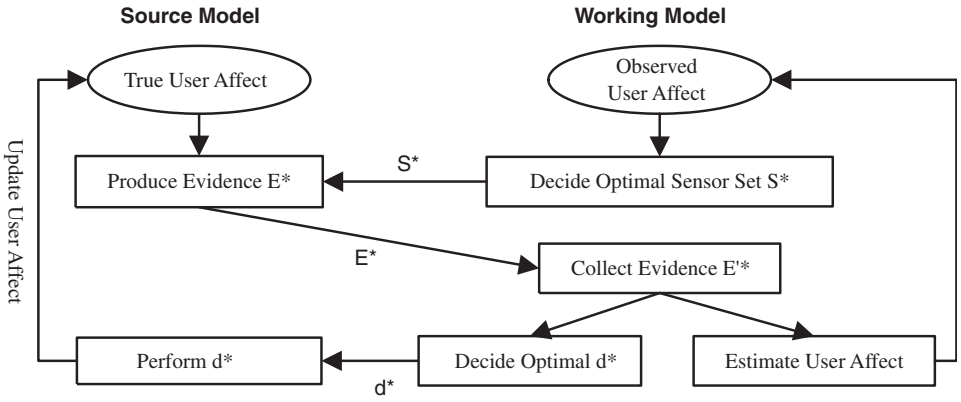

Fig. 10. A simulation system to validate the ID framework.

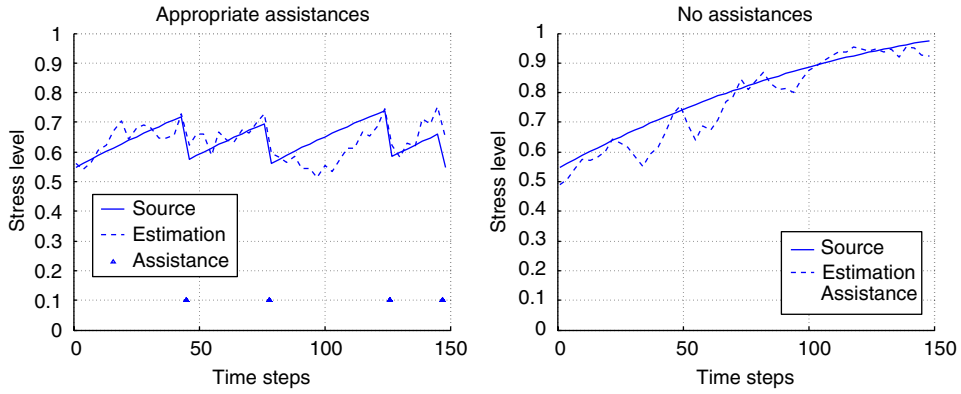

Fig. 11. Appropriate and timely user assistance vs. no assistance. In the left chart, the system provides timely and appropriate assistance. In the right, the system does not perform any assistance and degenerates to a recognizer of affective states. Solid (dashed) curve indicates the true (estimated) stress level from the source (working) model; triangle denotes assistance. The initial stress level in both the source model and working model is .5.

A good doctor should make correct diagnosis and treat a patient correctly; thus a working model should estimate user affect correctly and provide appropriate and timely user assistance, which is also the criterion we will use to evaluate our framework.

5.2. Simulation results

We have conducted experiments to show that the ID framework is capable of providing accurate estimation of affective states as well as offering timely and appropriate assistance. The experiments were conducted with the ID model shown in Fig. 6. First, the "affective states" node only models one affective state as stress, and it will be extended to the multiple affective states case later. We note that the specific design of assistance may vary with the applications. In the experiments, four types of assistance are designed: *null*, *warning*, *alleviating*, and *intervening*. Null means there is no assistance; warning is conveyed by playing stirring music or displaying an exhilarating scene; alleviating is to reduce workload on the user; and intervening may entail removing the user from control.

5.2.1. Appropriate and timely user assistance

The goal of our system is to provide timely and appropriate user assistance. Timely assistance means that the assistance is provided at time that the user is in a negative emotional state. Appropriate assistance optimizes the tradeoff between the benefits of bringing the user to a productive state and the cost of performing an assistance and interrupting the user.

Fig. 11 compares changes in the user's simulated emotional state when no assistance vs. appropriate assistance is provided. The left chart confirms that the system provides assistance only when the stress level is relatively high; at most times the system does not provide assistance since the user can maintain a productive state. Moreover, the performed assistance effectively reduces user stress level. Consequently, over all steps, the user maintains a reasonable stress level in the source model. Although we have designed four kinds of assistance in the system, only the first two (null, warning) are performed because the user never has a chance to reach a high stress level[1] due to the timely and appropriate assistance.

[1]We refer any stress level that is greater than .8 as high.

In contrast, as shown by the right chart in Fig. 11, the user assistance system degenerates to an affect recognizer if we set the costs of performing user assistance to be infinite. In this case, the system will never provide any assistance. Without any assistance, user stress level increases as time goes on. However, as shown by the chart, the system can still recognize user affect. Note that the estimated stress level from the working model is quite close to the stress level in the source model. The two curves track each other very well.

5.2.2. *Affective states recognition*

This subsection shows that the working model can recognize the affective states even without knowing the initial state precisely. The results are demonstrated in Fig. 12. The two charts show that the system is effective in recognizing affective states when the working model has no prior knowledge of the user affective states at all. The left chart is the recognition results for the case in which the source model starts with a highly stressed state, whereas the right chart is the recognition results for the case in which the source model starts with non-stressed state.

5.2.3. *Active sensor selection*

Fig. 13 shows how recognition performance varies with different sensor selection strategies. The left chart demonstrates the performance using our active sensing strategy, whereas the right chart shows the performance when sensors are randomly selected. Obviously, the active sensing leads to a better recognition performance. The average recognition error of using active sensing strategy is around 0.02, while the average error of random selection is around 0.07.

5.2.4. *Multiple affective states recognition and user assistance*

We also applied our methodologies to the case of recognizing multiple affective states and providing user assistance. We simulated three affective states—stress, fatigue and frustration. The results are shown in Fig. 14. From the top to the bottom, the results are shown, respectively, for stress, fatigue and frustration. In each chart, the solid curve denotes the truthful user affect, while the dashed curve denotes their estimations. For each affective state, it can be seen that the dashed curve is close to the solid one. This indicates that the ID works well in recognizing multiple affective states. In addition, the assistance is generated only when the user is at a high level of negative affective states; further, the user is occasionally provided only with `warning` assistance, which is the least intrusive assistance. Consequently, the system can provide timely and appropriate assistance.

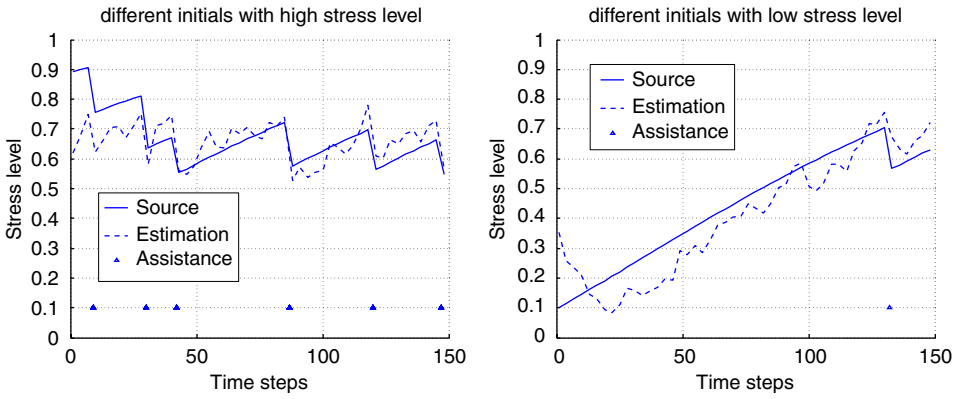

Fig. 12. User assistance varies with different initial affective states.

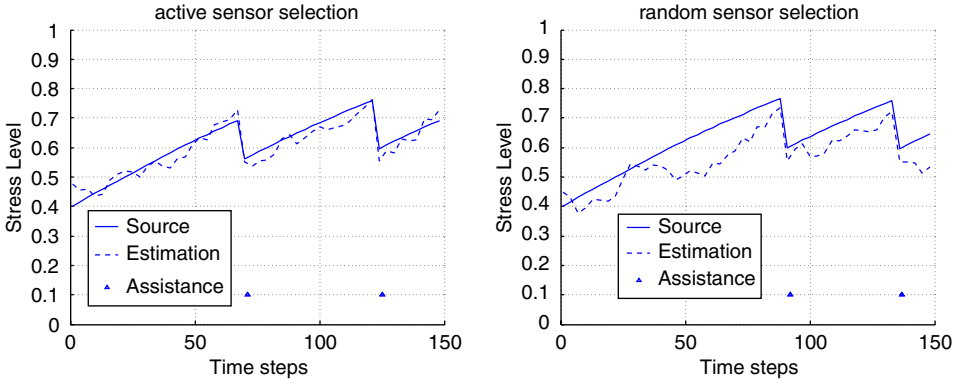

Fig. 13. Performance varies with different sensor selection strategies.

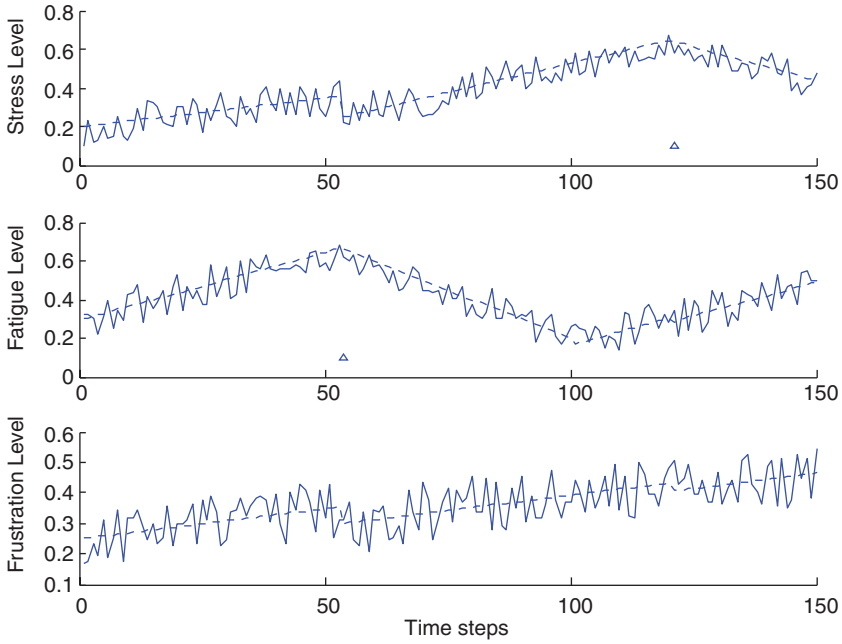

Fig. 14. Recognizing multiple affective states and user assistance.

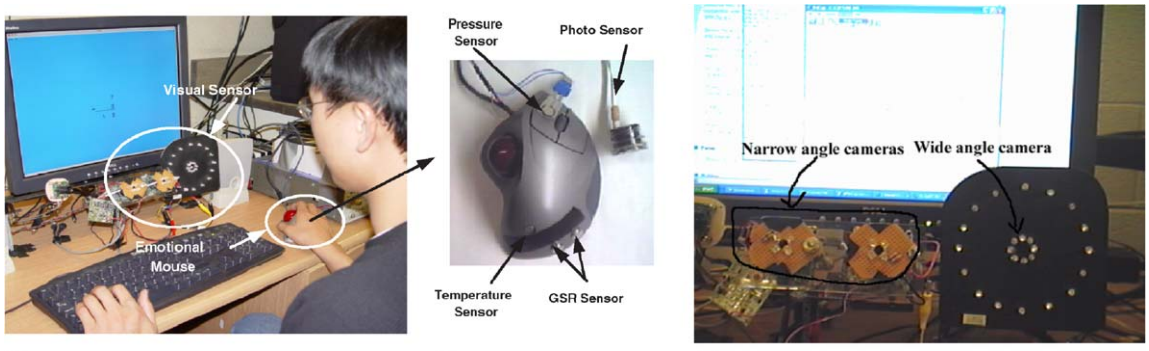

Fig. 15. The human–computer interaction environment: (a) overall hardware set-up; (b) emotional mouse; (c) visual sensor.

6. Human affect recognition validation with real data

6.1. System overview

We present our experimental results for recognition of two affective states—stress and fatigue. We begin by introducing our implemented system, the experimental environments, as well as the protocol to validate our framework. This is then followed by a report of our analysis results.

The HCI environment is shown in Fig. 15. During experiments, a user sits in front of a computer screen and responds to the tasks presented in the screen. A visual sensor suite, which is shown in Fig. 15(c), is used to monitor the user in real-time. It consists of three cameras: one wide-angle camera focusing on the face and two narrow-angle cameras focusing on the eyes. In addition, an emotional mouse (see Fig. 15(b)), which is built from a regular tracking-ball mouse by equipping it with physiological sensors, is used to collect physiological and behavioral data. Furthermore, a log file is created in the computer to record the user's performance data on the tasks that the user is working on. Under such a system set-up, various user state measurements characterizing the user can be extracted simultaneously and non-intrusively in real-time.

Fig. 16 gives an overview of the user affect monitoring system. It consists of three conceptual components. First, visual, physiological, behavioral and performance measures are extracted from corresponding sensors. Second, two statistical methods—correlation analysis and analysis of variance (ANOVA), are used to select the most discriminative features regarding user affect. Third, a dynamic influence diagram (DID) is constructed to infer user affect, which consists of a parameterization procedure

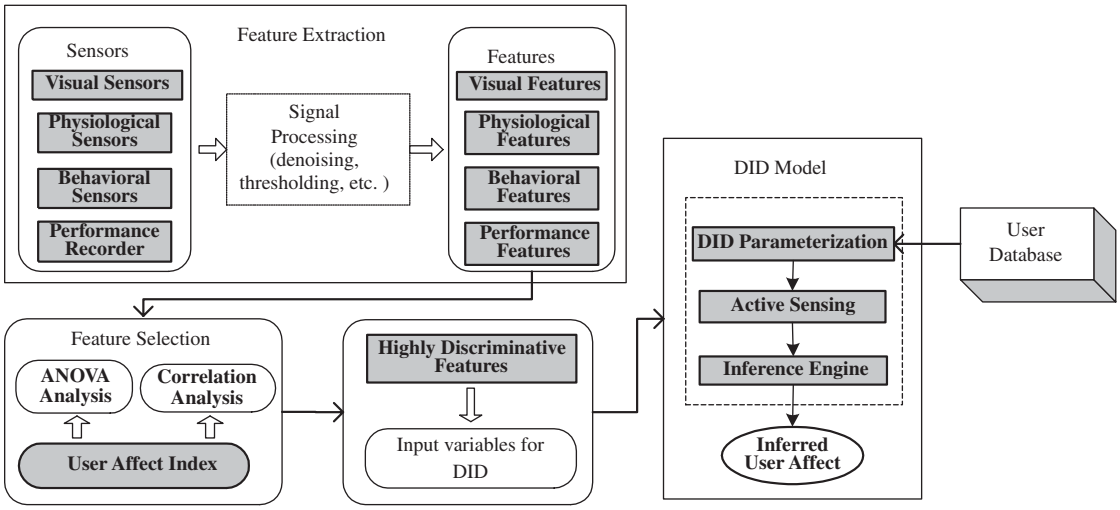

Fig. 16. The conceptual components of the user affect monitoring system.

to customize the DID to individual users with machine learning techniques, an active sensing strategy to select an optimal set of features for purposive and sufficing information integration, and a dynamic inference technique for efficient user affect recognition.

6.2. Feature extraction

6.2.1. Physical appearance features

We have developed a set of non-intrusive computer vision techniques for monitoring eyelid movement, eye gaze, head movement and facial expression in real-time (Gu et al., 2002; Ji, 2002; Zhu et al., 2002; Ji et al., 2004; Zhu and Ji, 2004, 2005). A number of visual features that can characterize a person's affective states are extracted. Our visual measurements consist of 10 features extracted from the real-time video: Blinking Frequency (BF), Average Eye Closure Speed (AECS), Percentage of Saccadic Eye Movement (PerSac), Gaze Spatial Distribution (GazeDis), Percentage of Large Pupil Dilation (PerLPD), Pupil Ratio Variation (PRV), Head Movement (HeadMove), Tilting Frequency (TiltFreq), Yawning Frequency (YawnFreq), and Mouth Openness (MouthOpen). The entire extraction procedure is divided into four relatively separate components—eye detection and tracking (extracting BF and AECS), gaze estimation (extracting PerSac, GazeDis, PerLPD and PRV), facial feature tracking (extracting Mouth Openness, Yawning Frequency) and face-pose estimation (extracting Head Movement and Tilting Frequency).

Visual feature extraction starts with eye detection and tracking, which serves as the basis for subsequent eyelid movement monitoring, gaze determination, face orientation estimation and facial expression analysis. A robust eye detection and tracking approach is developed via the combination of the appearance-based mean-shift tracking technique and bright-pupil effect under infrared light illumination (Zhu et al., 2002). Thanks to this combination, the eyes can be tracked under various face orientations and variable lighting conditions. Even though the eyes are completely closed or partially occluded due to the oblique face orientations, our eye tracker can still track them accurately. After tracking the eyes successfully, the eyelid movement can be subsequently monitored and the relevant eyelid movement parameters can be computed accurately.

In order to estimate the eye gaze under natural head movement and minimize the personal calibration, a computational dynamic head compensation model is developed (Zhang and Ji, 2005). The model can automatically update the gaze mapping function to accommodate the 3D head position changes when the head moves. Consequently, the gaze tracking technique allows free head movements in front of the camera but still achieves high gaze accuracy; meanwhile, the technique reduces the number of gaze calibration procedure to one time for each user. After estimating the eye gaze successfully, the gaze movement can be monitored and the relevant gaze parameters can be computed accurately.

To analyse the facial expressions, 28 facial features around eyes and mouth are selected for tracking (Zhang and Ji, 2005a). Each facial feature is represented by a set of multi scale and multi orientation Gabor wavelet. At each frame, based on the possible region for each facial feature as constrained by the detected pupils, the initial positions of each facial feature can be located via Gabor wavelet matching. In order to achieve a robust and accurate detection, the initial feature positions are then refined by a flexible global shape model based on active shape model (ASM) that constrains the spatial relationships between the detected facial features. To account for face poses, the global face shape model, which is learned under frontal

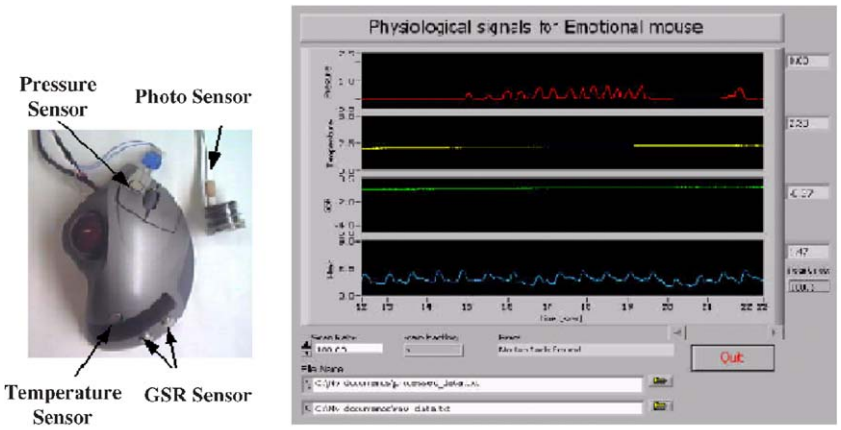

Fig. 17. A sample screen showing the measured physiological signals.

faces, is dynamically deformed via the previously estimated face-pose parameters to accommodate the face geometry changes. Thus, the correct global spatial constraints can always be imposed over the facial features so that they can be still tracked robustly under varying face orientations. Moreover, we also introduce a multi state face shape model in order to handle different facial expressions. Finally, a confidence verification procedure is carried out as a post-processing to handle cases of mis-tracking or self-occlusion. As a result, our technique is robust and insensitive to the variations in lighting, head motion, and facial expression.

Given the tracked facial features, the 3D non-rigid facial motion caused by the facial expression is estimated using a proposed motion extraction method. It will automatically eliminate the 3D head motion from the tracked facial features, therefore, the 3D non-rigid facial motion caused by the facial expression can be extracted under arbitrary face orientations. Then based on the extracted 3D non-rigid facial motion, a probabilistic framework is utilized to recognize the facial expressions by integrating the DBNs with the facial action units (AUs) from psychological view. Because of the successful modeling of the spatial and temporal behaviors of the facial expression via the proposed framework, the facial expressions can be recognized robustly and accurately under various face orientations. Six standard facial expressions can be recognized in the system (Zhang and Ji, 2005a).

In the face-pose estimation, we developed a technique that automatically estimates the 3D face pose based on the discovered facial features (Zhu and Ji, 2004). First, in order to minimize the effect of the facial expressions, our approach only chooses a set of rigid facial features that will not move, or move slightly, under various facial expressions for the face-pose estimation. Second, these rigid facial features are used to build a face shape model, whose 3D information is first initialized from a 3D generic face model. With the use of a frontal face image, the generic 3D face shape model is individualized automatically for each person. Third, based on the personalized 3D face shape model and its corresponding tracked facial features, our approach exploits a robust random sample consensus (RANSAC)-based method to estimate the face-pose parameters. Since this method automatically removes the inaccurate facial features, face-pose parameters can be always estimated from a set of facial features that are accurately detected in the face image.

6.2.2. Physiological, behavioral and performance evidence

To collect physiological features, an emotional mouse was built from a regular tracking-ball mouse by equipping it with physiological sensors. The emotional mouse measures heart rate, skin temperature, GSR and finger pressure. The mouse is designed to be non-intrusive. One sample measuring screen is shown in Fig. 17.

For behavioral evidence, we monitor a user's interactions with the computer, e.g. the mouse pressure from the finger (MousePre) each time when the user clicks the emotional mouse. Performance data is extracted from a log file that keeps track of user's response to the tasks, e.g. for the tasks in stress recognition, math/audio error rate (MathError, AudioError) and math/audio response time (MathRes, AudioRes) are extracted; for the tasks in fatigue recognition, response time, omission errors, and commission errors are extracted.

6.3. Stress recognition

6.3.1. Experiment design

Although the system is flexible enough to recognize multiple affective states, currently, we use the system to recognize only stress and fatigue. One fundamental difficulty in validating a stress monitoring system is the absence of ground-truth stress data. Some experiments have shown that even user self-reports are erroneous and unreliable. Fortunately, the existing results from

psychological studies show that occupational stress is affected by two job characteristics: demand and control (Karasek, 1979; Searle et al., 1999). Demand refers to the amount of attention and effort required for a user to carry out a job. We will interchangeably use *demand* and *workload*. Control primarily refers to the decision-making freedom presented in a job. It is predicted and confirmed that a user becomes more stressed when workload is high or when control is low (Searle et al., 1999).

In our study, we work on designing the experiments that are able to manipulate a subject's stress level by deliberately varying the task workload while fixing the control during the task trials. While task workload causes stress, they are not the same. To model their relationships and to infer stress from workload, we construct a BN as shown in Fig. 18. The directed links in the BN represent the casual relationships between the connected nodes. Specifically, for the stress study, the subject performs math and audio tasks.

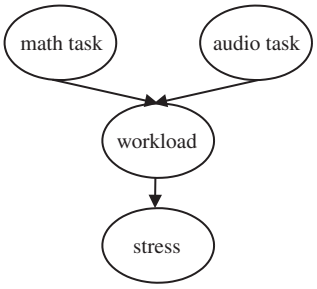

Fig. 18. A Bayesian network model for inferring the ground-truth stress from the task workload.

Varying the pace of the two tasks causes a change in workload, which, in turn, causes a change in subject's stress. After parameterizing the model with training data, the subject stress can then be inferred from the task levels. Since each person relates stress and workload differently, the model can be individualized to each subject.

During the experiments, the user is required to perform two different tasks: a math task that requires the addition or subtraction of two two-digit integers, and an audio task in which single letters of the alphabet are presented. For the math task, the user has to decide whether the answer presented on the screen is correct or not; for the audio task, the user has to either indicate whether the current letter precedes or follows the prior letter in the alphabet, or, determines whether the current letter (t) is equal to or different from the letter that was two back ($t-2$). Two types of task trials are arranged: single-task trial, where user performs only math or audio task; and dual-task trial, where user performs both math and audio tasks simultaneously. Each experiment session consists of approximately eight 10-min blocks. For the single-task trial, each block consists of eight intervals of 72 s (seconds) and the tasks are presented at the speed of 1, 2, 4, or 6 s. While for the dual-task trial, each block consists of 16 intervals of 36 s and the tasks are presented at the speed of 2, 4, or 6 s in each block.

6.3.2. Stress modeling

The structure of the dynamic ID for user stress is presented in Fig. 19. We set the cost of performing user assistance to be extremely high. Under this setting, the model degenerates to a stress monitor.

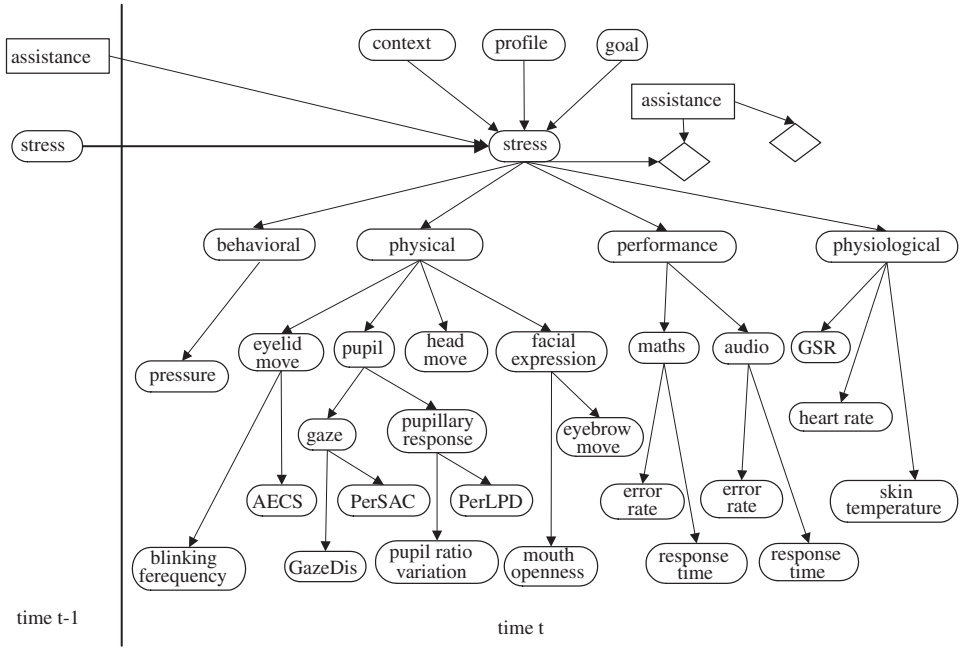

Fig. 19. Dynamic ID model for recognizing human stress. Due to space limit, we skip the sensory nodes and utility nodes that are associated with the evidence (leaf) nodes.

6.3.3. Results

Multiple subjects of different ages, genders and races have been tested in the system. Each subject attends one experiment session (eight blocks, around 80 min). For each subject, the data from five blocks are used for learning with the EM learning algorithm (Lauritzen, 1995), while the data from other three blocks are used for testing. In the following, we only report the results from six subjects, where three of them (A, B, C) perform the single-task trial, and the other three (D, E, F) perform the dual-task trial.

6.3.3.1. Stress vs. individual features. To demonstrate the relationship between each individual feature and stress, we carried out three types of statistical analysis—cumulative analysis, correlation analysis, and ANOVA test, where cumulative and correlation analysis are performed on the dual-task trial, and the ANOVA test is performed on the single-task trial. These analysis will help parameterize the DBN model.

In the cumulative analysis, we try to analyse the sensitivity and robustness of each feature to stress in average. The stress is divided into three levels and the feature values are grouped by each stress level. Then, the mean and standard deviation of individual feature in each group are computed. If a feature is sensitive to stress, the mean values vary with different stress levels; and if a feature is robust to stress changes, the standard deviations are small. Fig. 20 illustrates the results of 11 features for three subjects in the dual-task trial. It demonstrates that most features are sensitive and robust to the stress. As stress increases, a participant blinks less frequently, closes the eyes faster, dilates the pupils more often, focuses the eye gaze more on the screen, moves the head and opens the mouth less frequently, and clicks the mouse button harder. In the meantime, the heart rate increases, and GSR decreases. However, for different subjects, the same feature may have different levels of sensitivity to stress. For example, blinking frequency is a very good feature for subject D and F, while its mean values for subject E vary little as stress level changes. Thus, it is necessary to train the DID model for each subject.

In the correlation analysis, we study how feature values change as the stress changes over time. Fig. 21 illustrates the correlation curve between stress and three features in the dual-task trial for subject E. For charts (a) and (b), as stress level increases, the subject's pupil dilation is larger and his gaze focuses more on the central region of the computer screen where the math tasks are displayed. By contrast, chart (c) represents a negatively correlated relationship between AECS and stress: as stress level increases, the subject's average eye closure speed becomes slower. These observations are consistent with those in the cumulative analysis.

However, in real-time, individual features are not always reliable. Let us take the coefficients between stress and PerLPD as an example (Fig. 21(a)). Although it is approximately positively correlated to stress, negative coefficients occur in a time period; in addition, at some time steps, the coefficients fall below .3, thereby indicating that the correlation between stress and PerLPD is somehow weak. However, by combining the individual features with the DBN model, the inferred stress level has a very high correlation with the ground-truth stress, as will be shown later.

ANOVA is another statistical analysis to quantitatively determine the sensitivity of each feature to user stress. Table 1 displays the ANOVA test results for subjects A, B and C in the single-task trial. The data in each cell indicates the p-value. If the p-value is less than 0.05, it is believed the test result is statistically significant, which means the feature is sensitive to stress. The table shows most features are sensitive to stress. Similar to cumulative analysis, it shows that, for different subjects, the same feature may have different degrees of sensitivity to stress. For example, AECS is sensitive to stress for subjects A and B, while it is insensitive for subject C. Also, some features, e.g. AudioError, are almost not sensitive for all the subjects.

6.3.3.2. Stress inference with single-modality features vs. multiple-modality features. Our system has extracted four-modality features for stress recognition. One question is whether all these features are necessary or not. The experimental results demonstrate that single-modality features are not sufficiently precise for stress recognition, while multiple-modality features integrated by the DID model are very helpful for accurate stress inference.

Fig. 22 shows the results for the single-task trial. The x-coordinate indicates the ground-truth stress level from 1 to 4, and the y-coordinate indicates the means (the median points of the bars) and standard deviations (the heights of the bars) of the inferred stress levels. Ideally, the mean values should be very close to 1, 2, 3, and 4; the standard deviation should be quite small. However, as shown in Fig. 22(a)–(c), the results are not good enough when only using single-modality features. The mean values are not very close to 1, 2, 3, 4, and the standard deviation is not small. The inference result from the performance-modality feature is the worst. One possible explanation is that user may make more efforts to maintain his performance when stress level increases. Thus, the performance does not correlate well with stress. The inference result from the physiological features is better. However, sometimes the physiological signals may not be accurate enough since the user may move his hand irregularly while using the emotional mouse. Compared to other single-modality features, the visual features bring better inference result, which is contributed by our robust computer vision techniques in the system.

In summary, single modality alone cannot infer the stress reliably because it cannot provide enough features to infer the stress. Therefore, we tried to combine all these four modalities together to infer the stress. Fig. 22(d) displays the inferred stress results with multiple-modality features

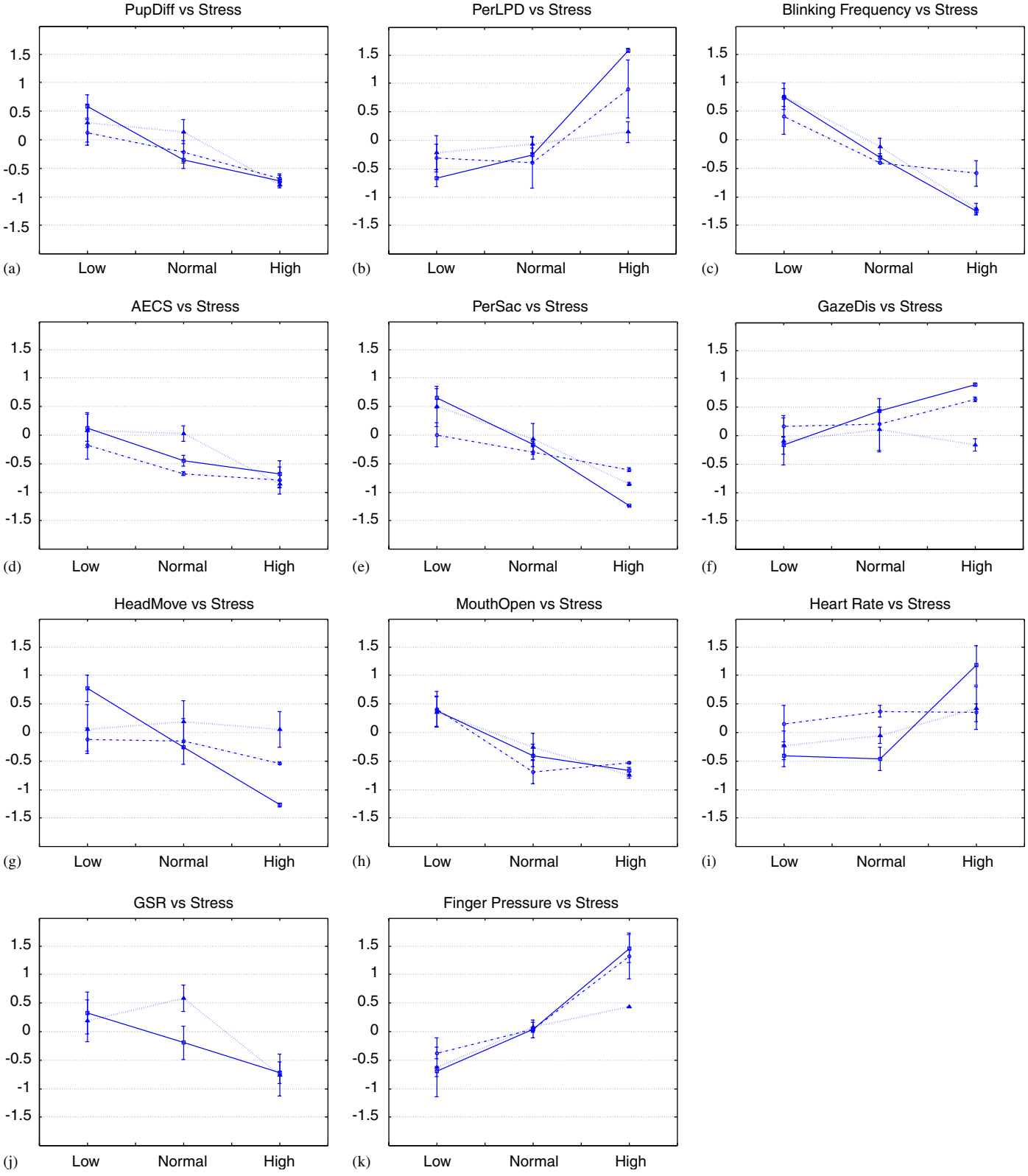

Fig. 20. The relationship between individual features and stress in the dual-task trial: (a) PupDiff; (b) PerLPD; (c) BF; (d) AECS; (e) PerSac; (f) GazeDis; (g) HeadMove; (h) MouthOpen; (i) heart rate; (j) GSR; (k) finger pressure. Vertical bar denotes standard deviation, and the median point denotes mean value. Each chart plots individual data for three subjects (subject D—solid line, subject E—dashed line, and subject F—dotted line) for the dual-task trial. All the feature values are normalized to zero mean and unit standard deviation. GSR feature for subject E was not available.

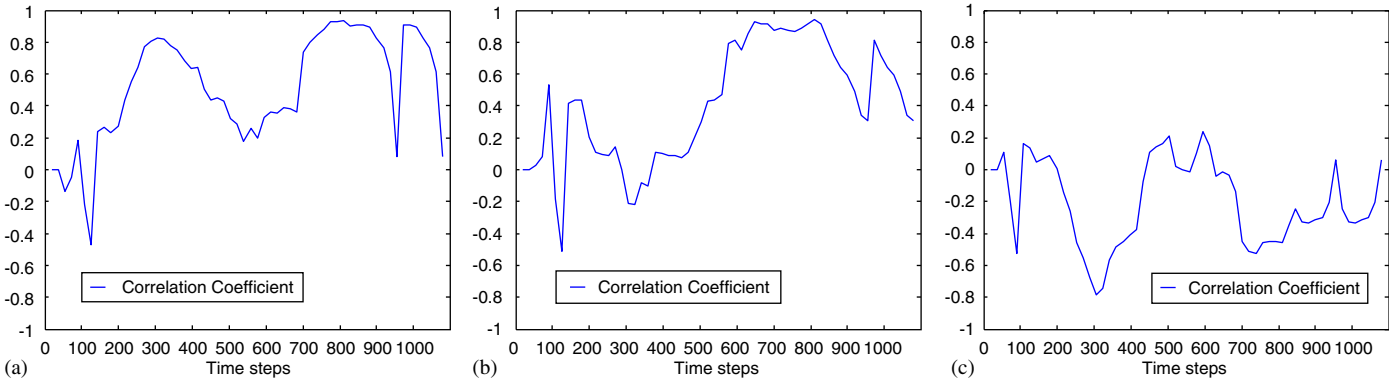

Fig. 21. The correlation curve (running average correlation coefficients) between individual visual features and stress for subject E in the dual-task trial: (a) PerLPD vs. stress; (b) GazeDis vs. stress; (c) AECS vs. stress. In the x-coordinate, the time unit is second.

Table 1
Summary of sensitivity results with ANOVA test in the single-task trial

Features	Subject A	Subject B	Subject C
HeartRate	.0001	.0026	.0040
GSR	.0019	.2289[a]	.0324
FingerPressure	.0000	.0013	.0009
AECS	.0062	.0001	.0557[a]
BF	.0000	.0000	.0070
GazeDis	.0000	.0036	.0723[a]
PerSac	.0000	.0002	.0318
PerLPD	.0000	.0204	.0747[a]
PupDiff	.0002	.0001	.0510[a]
MouthOpen	.0036	.0163	.1440[a]
HeadMove	.0000	.0094	.1841[a]
MathError	.0004	.0009	.0377
MathRes	.0000	.0000	.0141
AudioError	.0846[a]	.0918[a]	.3285[a]
AudioRes	.1684[a]	.1979[a]	.0000

The data denote the p-values.
[a]The values denote that the feature is not sensitive to stress changes for the corresponding subject.

using the DID model. The mean values of the stress indexes are close to the desired values. Quantitatively speaking, the average mean values for the three subjects are 1.18, 2, 2.98, and 3.93, which are much more closer to the desired values compared to charts (a)–(c) in Fig. 22. The standard deviations are also smaller (.08, .08, .1, .06), which means the inferred results from multi modality evidence are more consistent, accurate, and robust than those from single-modality evidence.

6.3.3.3. Stress recognition. Since single-modality features are not capable of recognizing user affect very well, our system integrates the multi modality features and performs efficient stress inference with the active sensing strategy proposed in Section 4.2. The experiments prove that our system can successfully monitor human stress in an efficient and timely manner. We show the inference results for the dual-task trial as an example.

The results for subject D are shown in the top two charts in Fig. 23. For each chart, the solid curve denotes the ground-truth stress and the dashed curve denotes the inferred stress levels. The top-left chart shows inferred stress based on six types of evidence. The top-right shows inferred stress based on 10 types of evidence. In both cases, the evidence selected is dynamically determined by the active sensing strategy. We see that inferred stress curves roughly trace the stress curves. In addition, we see that at most time steps the stress inferred from 10 pieces of evidence is more correlated to the actual stress than that inferred from six pieces of evidence.

6.3.3.4. Active sensing. The remaining charts in Fig. 23 demonstrate that the active sensing approach outperforms a passive (random selection) approach. The middle two charts show the inference performances for the random selection strategy, where the left selects six features and the right selects 10 features. To quantitatively show how much the active sensing approach outperforms, we conducted statistic correlation analysis. In both charts, the solid (dashed) curve denotes the correlation coefficients along time steps between ground-truth and inferred stress with the active (passive) sensing approach. In general, the solid curve lies above the dashed curve. This implies that the inferred stress is more correlated to ground-truth in the active sensing case. Consequently, the active sensing approach is effective in improving the inference performance in efficiently estimating stress.

6.4. Fatigue recognition

In addition to stress recognition, the system has also been tested in fatigue recognition on human subjects. The study includes a total of eight subjects. Two test bouts are performed for each subject. The first test is done when they first arrive in the lab at 9 p.m. and are fully alert. The second test is performed early around 7 p.m. in the following day, after the subjects have been deprived of sleep for a total of 25 h. During the study, the subjects are

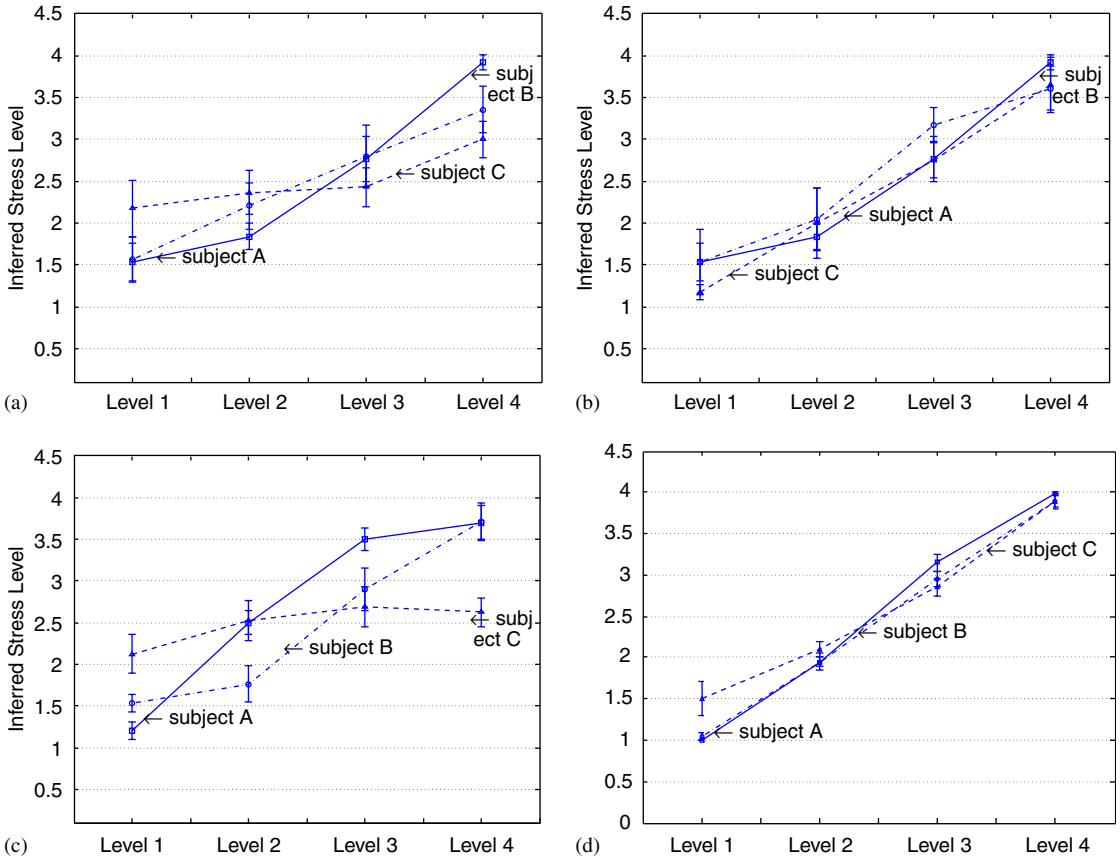

Fig. 22. Inferred stress level with DID for three subjects in the single-task trial: (a) with physiological and behavioral features only; (b) with visual features only; (c) with performance features only; (d) with multiple-modality features. The x-coordinate denotes the ground-truth stress level and the y-coordinate denotes the mean inferred stress levels and standard deviations. Inference results with multi modality features are better than the results with single-modality features.

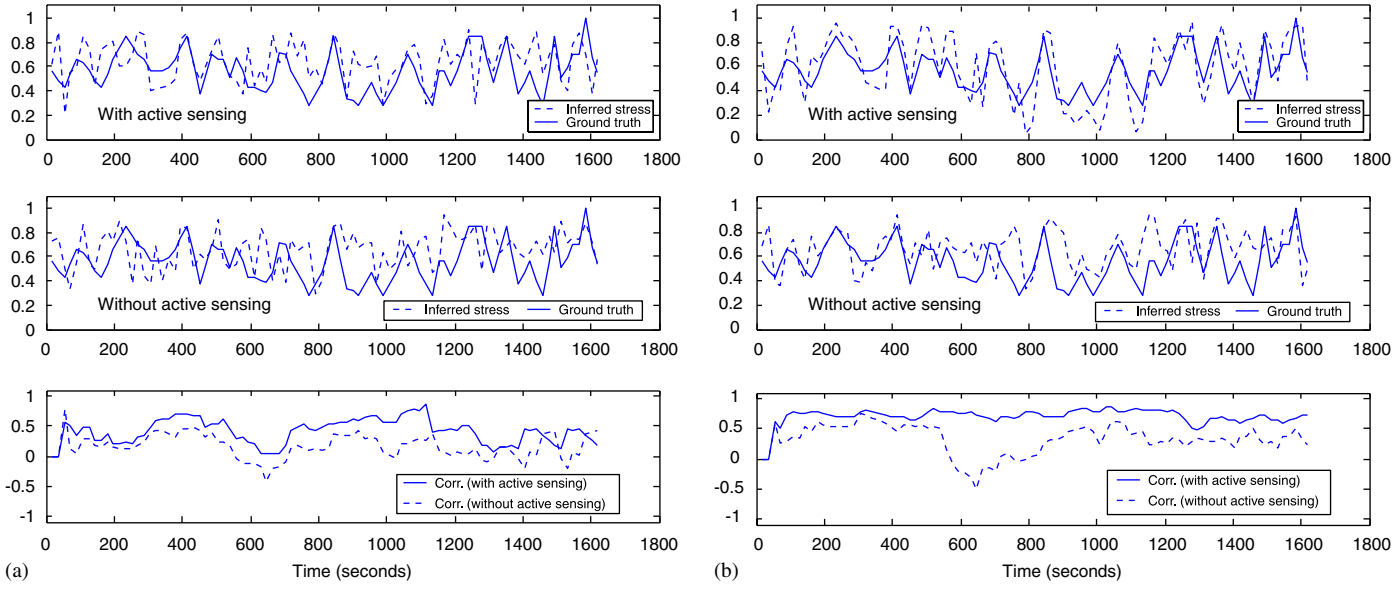

Fig. 23. Actual stress vs. inferred stress level with and without active sensing: (a) select six pieces of evidence in each time step; (b) select 10 pieces of evidence in each time step.

asked to perform a test of variables of attention (TOVA) test. The TOVA test consists of a 20-min psychomotot test, which requires the subject to sustain attention and respond to a randomly appearing light on a computer screen by pressing a button. The TOVA test keeps record of several measures related to the subject's behavior: the person's

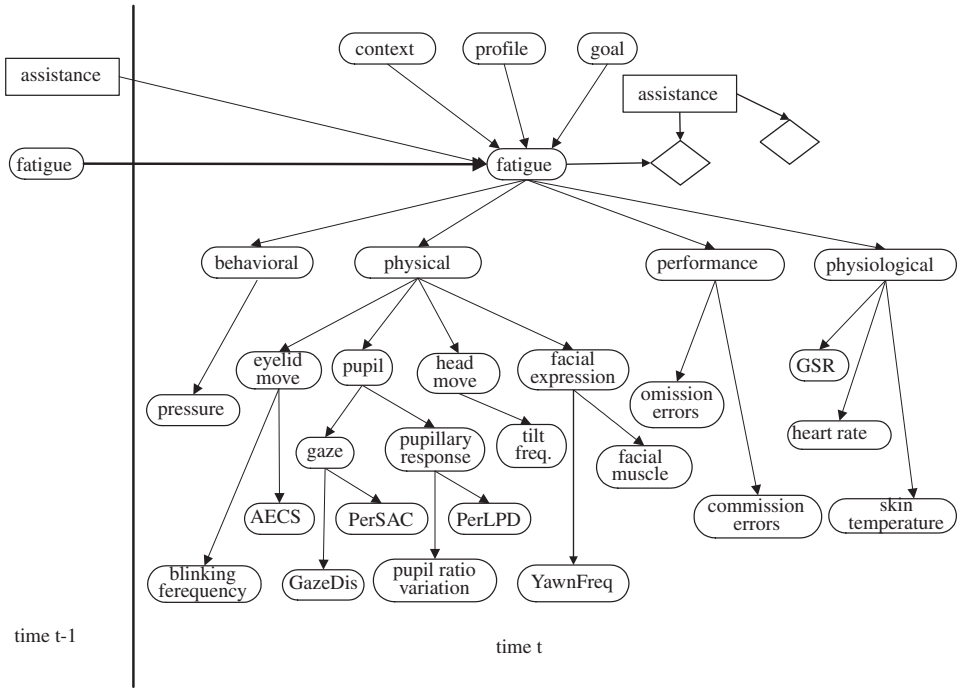

Fig. 24. Dynamic ID model for recognizing human fatigue. Due to space limit, we skip the sensory nodes and utility nodes that are associated with the evidence (leaf) nodes.

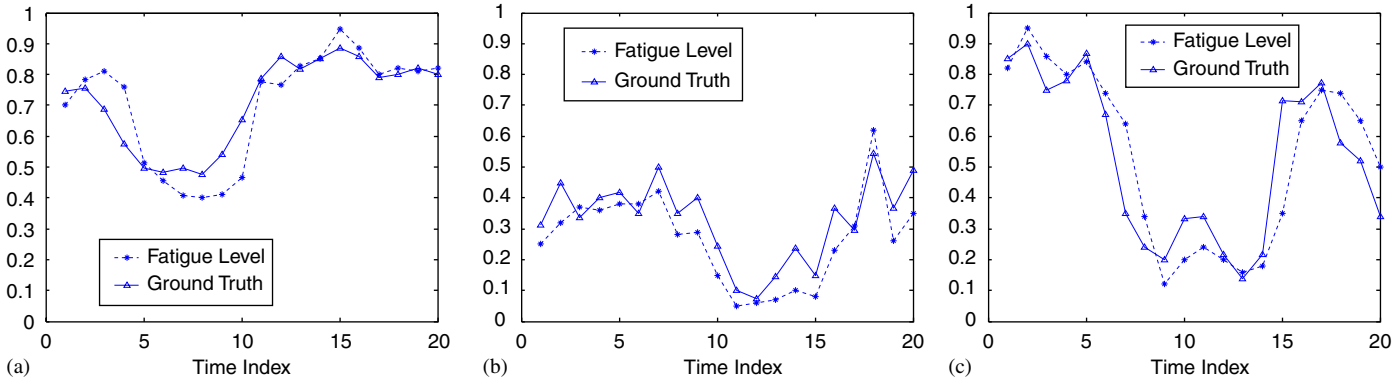

Fig. 25. Inferred fatigue vs. actual fatigue for: (a) subject 1; (b) subject 2; (c) subject 3.

response time when clicking the button, whether the person correctly responds to the symbols, and if not, what kind of mistakes the person makes. The response time is selected as a metric to quantify a person's performance so as to reflect the ground-truth fatigue (Dinges et al., 1998).

6.4.1. Fatigue modeling

The structure of the dynamic ID for user fatigue is presented in Fig. 24. We set the cost of performing user assistance to be extremely high. Under this setting, the model degenerates to a fatigue monitor.

6.4.2. Results

We have collected data from eight subjects of different ages, genders and races. For each subject, part of the data is used for learning with the EM algorithm (Lauritzen, 1995), while part of the data is used to validate the model. Fig. 25 plots the estimated fatigue level vs. the actual fatigue over time for three subjects. In each chart, it is clear that the two curves' fluctuation match well, qualitatively proving their correlation and co-variation. The correlation coefficients of the two curves for the three subjects are 0.89, 0.91 and 0.88, respectively, therefore quantitatively proving the good performance of the fatigue monitoring system.

7. Conclusions and future work

This paper presents a dynamic decision framework based on IDA for simultaneously modeling affective state recognition and user assistance. The technical issues involve user affect recognition, active sensing and user assistance determination. The proposed framework can be

used to unify the three tasks: (1) user affect recognition by probabilistic inference from the dynamically generated evidence of different modalities, (2) active sensor selection by selectively combining the most effective sensor measurements for efficient and timely user state recognition, and (3) automated user assistance by balancing the benefits of improving user productivity in affective states and the costs of performing possible user assistance. To validate the proposed framework, we design a simulation system to emulate user behavior in order to validate the model correctness and effectiveness. The simulation results show that the framework successfully realizes the two central functions in intelligent user assistance systems. Furthermore, a non-invasive real-world human affect monitoring system is built to demonstrate the user affect recognition capability on real human subjects. Such a system non-intrusively collects the four-modality evidence including physical appearance, physiological, behavioral, and performance measures. To our knowledge, this integration from four-modality evidence, together with the probabilistic approach, is unique in user affect research.

Several directions deserve further investigations in the future. First, although the proposed decision-theoretic work has been validated in a simulation system, the current real-time system does not fully integrate the assistance function yet. We would like to work on it in the future. Second, it would be interesting to integrate more contextual information and observable evidence in affect recognition. The contextual information such as user's age, physical fitness, and user's skill level can further improve the recognition accuracy and robustness. In addition, the evidence from other modalities may also embody user affect. For example, user acoustic features have proven useful as reported in linguistic research work (Ball and Breeze, 2000). The contextual information and additional evidence can be integrated into the proposed framework and further improve the accuracy of user affect recognition. Third, our study reveals that a simple model cannot be generalized well to each individual since different persons may have different symptoms even under the same affect. Instead, under the same framework, an individual model should be constructed for each person. One possible future research is to improve the model learning algorithms so that the framework can be better individualized. Finally, it would be interesting and also necessary to integrate a computational affect model with a cognitive user model in order to identify the causes for certain user states and to provide appropriate assistance. Our current work mainly focuses on recognizing human affect from the external symptoms and providing user assistance accordingly. Like other user assistance systems, the user assistance is only superficial since the current system does not really understand the causes of the user affect. For example, the affective state could be stressed; the cognitive steps that lead to user's stress may be numerous. The probabilistic user model cannot figure out why the user is stressed. Understanding the cause for user's affect is crucial to offering correct augmentation and appropriate assistance. In this regard, we are investigating the integration of a cognitive model based on ACT-R (Anderson and Lebiere, 1998) into the loop so that it is feasible to analyse the sub symbolic parameters in cognitive side and explain the identified user affect. The ongoing research in this lab is pursuing in this direction.

Acknowledgments

The research described in the paper is supported in part by a grant from the Defense Advanced Research Projects Agency (DARPA) under the Grant number N00014-03-01-1003 and in part by a grant from the Air Force Office of Scientific Research under Grant number F49620-03-1-0160.

References

Anderson, J.R., Lebiere, C., 1998. The Atomic Components of Thought. Lawrence Erlbaum Associates, Mahwah, NJ.

Ark, W.S., Dryer, D.C., Lu, D.J., 1999. The emotion mouse. The Eighth International Conference on Human–Computer Interaction: Ergonomics and User Interfaces, vol. I, pp. 818–823.

Ball, G., Breeze, J., 2000. Emotion and personality in a conversational agent. In: Cassel, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.), Embodies Conversational Agents. MIT Press, Cambridge, MA, pp. 189–219.

Bauer, M., Gymtrasiewicz, P.J., Vassileva, J., 2001. User Modeling. Springer, New York.

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin 91, 276–292.

Berthold, A., Jameson, A., 1999. Interpreting symptoms of cognitive load in speech input. In: Proceedings of the Seventh International Conference on User Modeling, pp. 235–244.

Boverie, S., Leqellec, J.M., Hirl, A., 1998. Intelligent systems for video monitoring of vehicle cockpit. In: International Congress and Exposition ITS. Advanced Controls and Vehicle Navigation Systems, pp. 1–5.

Breazeal, C., 1999. Robot in society: friend or appliance? Workshop on Emotion-based Agent Architectures, pp. 18–26.

Buntine, W., 1994. Operations for learning with graphical models. Journal of AI Research 159–225.

Carskadon, M.A., Dement, W.C., 1982. The multiple sleep latency test: what does it measure? Sleep 5, 67–72.

Cohen, I., Grag, A., Huang, T.S., 2000. Emotion recognition using multilevel-HMM. NIPS Workshop in Affective Computing.

Conati, C., 2002. Probabilistic assessment of user's emotions in educational games. Journal of Applied Artificial Intelligence, Special Issue on Merging Cognition and Affect in HCI 16, 555–575.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human–computer interaction. IEEE Signal Processing Magazine 18, 32–80.

Diez, F.J., 1993. Parameter adjustment in Bayes networks: the generalized noisy or-gate. In: Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI93), pp. 99–105.

Dinges, D.F., Mallis, M.M., Maislin, G.M., Powell, J.W., 1998. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. NHTSA Report No. DOT HS 808 762, April.

Ekman, P., Friesen, W., 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press.

Elliott, C., Rickel, J., Lester, J.C., 1999. Lifelike pedagogical agents and affective computing: an exploratory synthesis. Artificial Intelligence Today 195–211.

Empson, J., 1986. Human Brainwaves: The Psychological Significance of the Electroencephalogram. The Macmillan Press Ltd, New York.

Gardell, B., 1982. Worker participation and autonomy: a multilevel approach to democracy at the workplace. International Journal of Health Services 4, 527–558.

Grace, R., 2001. Drowsy driver monitor and warning system. International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, August.

Gu, H., Ji, Q., Zhu, Z., 2002. Active facial tracking for fatigue detection. IEEE Workshop on Applications of Computer Vision.

Hartley, L., Horberry, T., Mabbott, N., Krueger, G.P., 2000. Review of Fatigue Detection and Prediction Technologies. National Road Transport Commission, ISBN 0 642 54469 7.

Hass, M.W., Hettinger, L., 2001. Adaptive User Interface. Lawrence Erlbaum Associates, Mahwah, NJ.

Healy, J., Picard, R., 2000. Smartcar: detecting driver stress. The 15th International Conference on Pattern Recognition, pp. 4218–4221.

Heishman, R., Duric, Z., Wechsler, H., 2004. Using eye region biometrics to reveal affective and cognitive states. CVPR Workshop on Face Processing in Video.

Heitmann, A., Guttkuhn, R., Trutschel, U., Moore-Ede, M., 2001. Technologies for the monitoring and prevention of driver fatigue. In: Proceedings of the First International Driving Symposium on Human Factors in Driving, Assessment, Training and Vehicle Design, pp. 81–86.

Horvitz, E., 1999. Uncertainty, action, and interaction: in pursuit of mixed-initiative computing. IEEE Intelligent Systems September Issue, 17–20.

Horvitz, E., Kadie, C.M., Paek, T., Hovel, D., 2003. Models of attention in computing and communications: from principles to applications. Communications of the ACM 46, 52–59.

Howard, R., 1967. Value of information lotteries. IEEE Transactions of Systems Science and Cybernetics 3 (1), 54–60.

Howard, R., Matheson, J., 1981. Influence diagrams. Readings on the Principles and Applications of Decision Analysis 2, 721–762.

Hudlicka, E., McNeese, M.D., 2002. Assessment of user affective and belief states for interface adaptation: application to an air force pilot task. User Modeling and User Adapted Interaction 12, 1–47.

Ji, Q., 2002. 3D face pose estimation and tracking from a monocular camera. Image and Vision Computing 20, 499–511.

Ji, Q., Zhu, Z., Lan, P., 2004. Real time non-intrusive monitoring and prediction of driver fatigue. IEEE Transactions on Vehicle Technology 53 (4), 1052–1068.

Jones, F., Bright, J., 2001. Stress: Myth, Theory and Research. Prentice-Hall, Englewood Cliffs, NJ.

Jordan, M.I. (Ed.), 1999. Learning in Graphical Models. MIT Press, Cambridge, MA.

Kaapor, A., Mota, S., Picard, R., 2001. Toward a learning companion that recognizes affect. AAAI Fall Symposium: Emotional and Intelligent, North Falmouth, MA, November 2–4.

Kaliouby, R.E., Robinson, P., 2004. Real-time inference of complex mental states from facial expressions and head gestures. IEEE Workshop on Real-time Vision for Human–Computer Interaction in Conjunction with IEEE CVPR, Washington, DC, July.

Kapoor, A., Picard, R., Ivanov, Y., 2004. Probabilistic combination of multiple modalities to detect interest. IEEE International Conference on Pattern Recognition 3, 969–972.

Karasek, R.A., 1979. Job demands, job decision latitude and mental strain: implications for job design. Administrative Science Quarterly 24, 285–308.

Kataoka, H., Yoshida, H., Saijo, A., Yasuda, M., Osumi, M., 1998. Development of a skin temperature measuring system for non-contact stress evaluation. In: Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2, pp. 940–943.

Kevin, B., Ann, E., 2003. Bayesian Artificial Intelligence. Chapman & Hall/CRC, London, Boca Raton, FL.

Lauritzen, S.T., 1995. The EM algorithm for graphical association models with missing data. Computational Statistics and Data Analysis 19, 191–201.

Lemmer, J.F., Gossink, D., 2004. Recursive noisy-or: a rule for estimating complex probabilistic causal interactions. IEEE Transactions on Systems, Man, and Cybernetics 34 (6), 2252–2261.

Li, X., Ji, Q., 2004. Active affective state detection and user assistance with dynamic Bayesian networks. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 35 (1), 93–105.

Machleit, K., Enoglu, S., 2000. Describing and measuring emotional response to shopping experience. Journal of Business Research 49, 101–111.

Maes, P., Schneiderman, B., 1997. Direct manipulation vs. interface agents: a debate. Interactions IV (6).

Massaro, D.W., 2000. Multimodal emotion perception: analogous to speech processes. ISCA Workshop on Speech and Emotion.

Mindtools, 2004. ⟨http://www.mindtools.com/stress/understandstress/stressperformance.htm⟩.

Moriyama, T., Saito, H., Ozawa, S., 1997. Evaluation of the relation between emotional concepts and emotional parameters on speech. IEEE International Conference on Acoustic, Speech, and Signal Processing 2, 1431–1434.

Murray, R.C., VanLehn, K., Mostow, J., 2004. Looking ahead to select tutorial actions: a decision-theoretic approach. International Journal of Artificial Intelligence in Education 14, 235–278.

Ortony, A., Clore, G.L., Collins, A., 1988. The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, England.

Pantic, M., Patras, I., Rothkrantz, L. M., 2002. Facial mimics recognition from face profile image sequences. Data and Knowledge Systems Group, Delft University of Technology, Netherlands.

Partala, T., Surakka, V., 2003. Pupil size variation as an indication of affective processing. International Journal of Human–Computer Studies 59, 185–198.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inferences. Morgan Kaufmann Publishers, Los Altos, CA.

Petrushin, V.A., 1999. Emotion in speech: recognition and application to call centers. Conference on Artificial Neural Networks in Engineering (ANNIE'99), St. Louis, November 7–10.

Petrushin, V.A., 2000. Emotion recognition in speech signal: experimental study, development, and application. The Sixth International Conference on Spoken Language Processing.

Picard, R., 1997. Affective Computing. Cambridge University Press, Cambridge, England.

Picard, R., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: analysis of affective physiological state. IEEE Transactions on Patter Analysis and Machine Intelligence 23 (10), 1175–1191.

Qi, Y., Picard, R.W., 2002. Context-sensitive Bayesian classifiers and application to mouse pressure pattern classification. The Proceedings of International Conference on Pattern Recognition, Qué., City, Canada, August.

Qi, Y., Reynolds, C., Picard, R.W., 2001. The Bayes point machine for computer–user frustration detection via pressure mouse. The Proceedings of the 2001 Workshop on Perceptive User Interfaces, vol. 15, pp. 1–5.

Rani, P., Sarkar, N., Smith, C.A., 2003. Anxiety detection for implicit human–robot collaboration. IEEE International Conference on Systems, Man and Cybernetics 4896–4903.

Rimini-Doering, M., Manstetten, D., Altmueller, T., Ladstaetter, U., Mahler, M., 2001. Monitoring driver drowsiness and stress in a driving simulator. First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, pp. 58–63.

Rosekind, M.R., Co, E.L., Gregory, K.B., Miller, D.L., 2000. Crew factors in flight operations XIII: a survey of fatigue factors in

corporate/executive aviation operations. National Aeronautics and Space Administration NASA/TM-2000-209610.

Saito, H., Ishiwaka, T., Sakata, M., Okabayashi, S., 1994. Applications of driver's line of sight to automobiles—what can driver's eye tell. Proceedings of 1994 Vehicle Navigation and Information Systems Conference, Yokohama, Japan, pp. 21–26.

Scherer, K.R., 1993. Studying the emotion-antecedent appraisal process: an expert system approach. Cognition and Emotion 7, 325–355.

Searle, B.J., Bright, J.E., Bochner, S., 1999. Testing the three-factor model of occupational stress: the impacts of demands, control and social support on a mail sorting task. Work and Stress 13, 268–279.

Shachter, R., 1986. Evaluating influence diagrams. Operations Research 34 (6), 871–882.

Sherry, P., 2000. Fatigue countermeasures in the railroad industry—past and current developments. Counseling Psychology Program, Intermodal Transportation Institute, University of Denver.

Ueno, H., Kaneda, M., Tsukino, M., 1994. Development of drowsiness detection system. Proceedings of 1994 Vehicle Navigation and Information Systems Conference, Yokohama, Japan, August 1994, pp. 15–20.

Yeasin, M., Bullot, B., Sharma, R., 2004. From facial expression to level of interest: a spatial-temporal approach. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 922–927.

Zhang, N.L., Poole, D., 1996. Exploiting causal independence in Bayesian network inference. Journal of Artificial Intelligence Research 5, 301–328.

Zhang, Y., Ji, Q., 2005a. Active and dynamic information fusion for facial expression understanding from image sequence. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27 (5), 699–714.

Zhang, Y., Ji, Q., 2005b. Sensor selection for active information fusion. The 20th National Conference on Artificial Intelligence (AAAI-05).

Zhu, Z., Ji, Q., 2004. 3D face pose tracking from an uncalibrated monocular camera. IEEE International Conference on Pattern Recognition.

Zhu, Z., Ji, Q., 2005. Eye gaze tracking under natural head movements. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR05), San Diego, CA, June 2005.

Zhu, Z., Ji, Q., Fujimura, K., Lee, K., 2002. Combining Kalman filtering and mean shift for real time eye tracking under active IR illumination. IEEE International Conference on Pattern Recognition, Que., Canada.