

Learning and Reasoning about Interruption

Eric Horvitz
Microsoft Research
Redmond, Washington USA
+1 425-706-2127
horvitz@microsoft.com

Johnson Apacible
Microsoft Research
Redmond, Washington USA
+1 425-707-3167
johnsona@microsoft.com

ABSTRACT

We present methods for inferring the cost of interrupting users based on multiple streams of events including information generated by interactions with computing devices, visual and acoustical analyses, and data drawn from online calendars. Following a review of prior work on techniques for deliberating about the cost of interruption associated with notifications, we introduce methods for learning models from data that can be used to compute the expected cost of interruption for a user. We describe the *Interruption Workbench*, a set of event-capture and modeling tools. Finally, we review experiments that characterize the accuracy of the models for predicting interruption cost and discuss research directions.

Categories and Subject Descriptors

I.2.10, J.4 [Artificial Intelligence, Social and Behavioral Sciences]: Perceptual Analysis, Economics

General Terms

Human Factors, Economics, Experimentation, Theory

Keywords

Cognitive models, divided attention, interruption, notifications

1. INTRODUCTION

Interest has been growing over the last several years on methods for endowing computing systems with an understanding of users' focus of attention, workload, and interruptability. The work builds on psychological research on interruption and divided attention, extending from classic studies in the early Twentieth Century [13,15]. Efforts have focused in parallel on user studies, the formulation of models of attention and cognitive load, the construction of real-time sensing and reasoning platforms, and the development of applications such as notification managers, communication agents, and dialog systems [7]. In user studies, researchers have elucidated the effects of interrupting people in various ways in different situations [2,3,10,11], and have probed the workload and availability of people in office settings [6,8,9].

In this paper, we focus on efforts to build and use models of a user's attentional focus and workload within the Attentional User Interface (AUI) project [5,7]. The models of attention play a

central role in systems that perform ongoing cost-benefit analyses to mediate the flow of alerts and communications to users [5,6]. We have pursued models that can be employed to reason about a user's workload from observed events and, more specifically, to infer in an automatic manner the *cost of interruption* to users associated with different kinds of alerts and communications. Such models of interruption fuse together information from multiple sensory channels, including desktop events, calendar information, and ambient visual and acoustical analyses.

We shall present methods for building models that can be used to infer a user's state of interruptability from multiple event sources, and, that can provide a well-characterized expected cost of interruption. We first review work on the coupling of models of attention with event systems that provide streams of events, including desktop activity and sensory observations. We describe in particular techniques for computing the expected cost of interruption, given a probability distribution over attention and a utility assessment that encodes preferences about the costs of interruption in different situations. Then, we discuss the learning of models of attention and interruptability from data. We review the learning paradigm and tools, describe the learned models, and discuss experiments that probe the classification accuracy of the models. Then we present model ablation studies, including experiments with removing from consideration perceptual sensing and only including the discriminatory power of models that rely on events generated by interactions with a client computing system and information drawn from an online calendar.

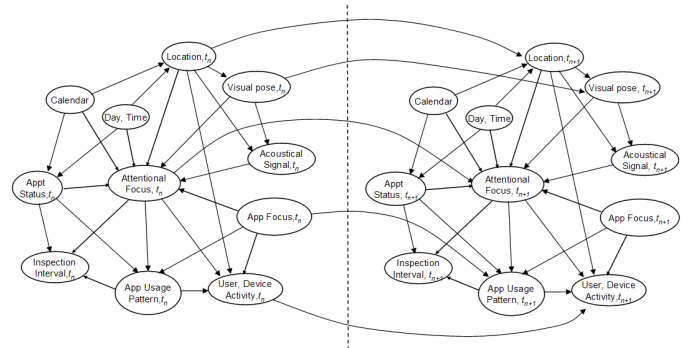


Figure 1. Dynamic Bayesian network for reasoning about a user's attentional focus, capturing key variables and intra- and inter-temporal probabilistic dependencies (from [5]).

2. INFERRING COST OF INTERRUPTION

Initial versions of an alerting mediation system, named the Notification Platform employed handcrafted dynamic Bayesian network models [5]. Two adjacent states of such a temporal Bayesian model are displayed in Figure 1. An important variable,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.
Copyright 2003 ACM 1-58113-621-8/03/0011...\$5.00.

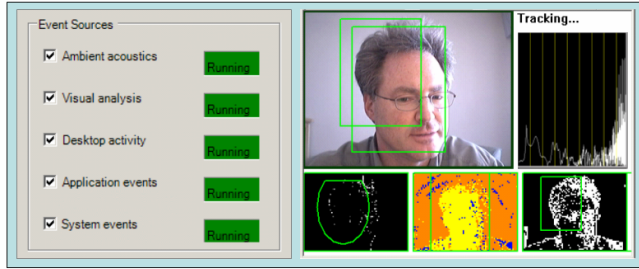


Figure 2. Control panel for Infocflow event system, showing event classes and graphical display of processing of acoustical and visual information.

represented over time in the Bayesian network, is *Attentional focus*. The states of this variable are structured into approximately fifteen mutually exclusive states of attention, representing a spectrum of user situations, capturing different amounts of cognitive workload and tolerance of interruptions. The states include such distinctions as *high-focus solo activity*, *medium-focus solo activity*, *low-focus solo activity*, *conversation in office*, *presentation*, *driving*, *private/personal time*, and *sleeping*.

2.1 Representing Preferences about Disruption

Beyond reasoning about the attentional states of users, an important goal of models of attention is to infer the cost of different types of interruption, conditioned on users being in particular states. To perform this inference, we consider the utility, $u(D_i, A_j)$, representing the cost of a user in attentional state A_j being disrupted by a task or communication event D_i . We have found that users are comfortable with assessing this cost as the *willingness to pay* to avoid a disruption in dollars for each outcome tuple. Willingness to pay to avoid outcomes has been used in decision analyses in several fields, including medical decision analysis. Given a set of dollar values that users assert that they are willing to pay to avoid different kinds of disruptions, and a probability distribution being inferred over the attentional state of a user, we compute the expected cost of interruption (ECI) by summing over the utilities, weighted by the likelihood of each state of attention, conditioned on the stream of incoming sensory information. That is, the ECI is

$$ECI = \sum_j p(A_j | E) u(D_i, A_j) \quad (1)$$

where $p(A_j|E)$ is the probability of the attentional state, conditioned on evidence stream E .

2.2 Event Systems for Sensing User State

Let us now explore additional details of a real-world implementation of a system that can compute the cost of interruption from a real-time stream of events, including interactions sensed by a computing device and perceptual information gathered from acoustical and visual sensors.

In the Notification Platform, we monitor the activity of a user interacting with their different client devices with event sensing and abstraction systems that sense computer events from the operating systems and applications executed on the clients. Given the availability of a microphone and camera, computing devices or appropriately equipped locations can also report visual pose

with a Bayesian head tracking system [14] and the presence of nearby conversation with an audio signal processing analysis. Finally, we examine information from users' online appointment information via an interface to the Microsoft Outlook application to see if a meeting is scheduled, and if so, we identify and consider as observational events several properties of the meeting.

Figure 2 displays a control panel from an event monitoring subsystem used in Notification Platform that we refer to as Infocflow. The event system provides an abstraction tool for coalescing patterns of low-level system events into higher-level events. We consider in the models of attention both low- and high-level events. For example, we capture such low-level states as the specific application currently in focus, whether the user is typing, clicking and/or moving the mouse, as well as a set of higher-level events such as the pattern of switching among applications (*e.g.*, single application focus versus switching among applications) and indications of task completion (*e.g.*, a message being sent, a file being closed, an application being closed, etc.). For the calendar events, we consider whether a meeting is in progress, the length of time until the meeting is over, and the location of the meeting as indicated by information from Outlook. For the acoustical and visual analysis, we note the presence of conversation or non-conversational sounds associated with user activity, and whether a user is present near a desktop system, and if so, if the user is gazing at or away from the computer. Figure 3 displays the status of a volatile event store used by Infocflow, named the *Event Whiteboard*. The Event Whiteboard is used to capture and share out the state of low-level and higher-level events gathered by Infocflow with other applications.

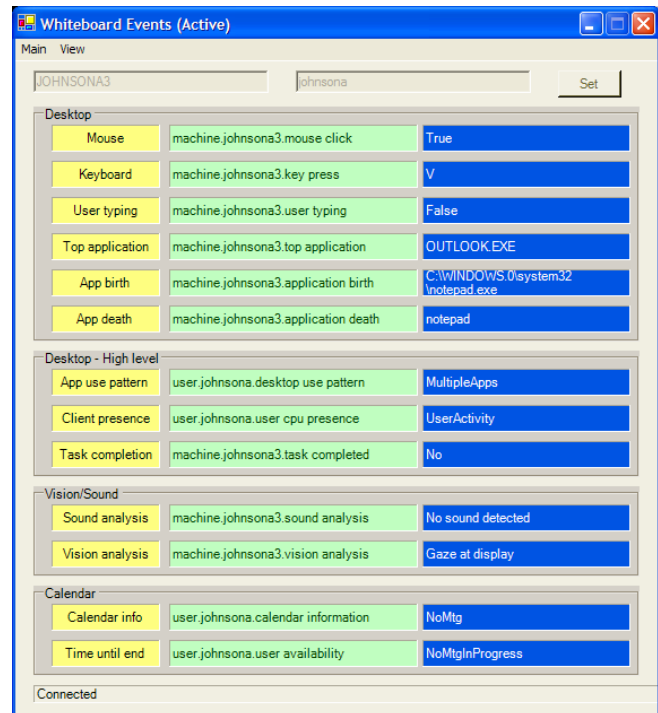


Figure 3. Event Whiteboard of Infocflow subsystem, displaying low-level and higher-level events considered by the probabilistic models of interruption.

As indicated in the figure, events include details about the birth and death of specific applications, the application currently in focus and being interacted with (Outlook), and events that capture usage patterns such as the desktop usage pattern, in this case, showing that a user is switching between different applications within a preset time horizon (15 seconds in this case).

2.3 Real-Time Analysis of Interruption

We shall now briefly examine the reasoning performed by the Notification Platform in mediating notifications. The Notification Platform considers the events described in Section 2.2, and employs a Bayesian network to infer a probability distribution over attentional states. Figure 4a displays the output of a model that considers eight states of attention, including *High-Focus Solo Activity*, *Low-Focus Solo Activity*, *Conversation in Office*, *Presentation or Meeting*, *Driving*, *Private Personal Time*, *Sleeping*, and *Now Available*. The curves in the figure depict that the initial high likelihood of *Conversation in Office* has, at the most recent time, become dominated by *High-Focus Solo*. Figure 4b shows inferences about the expected cost of interruption over time for different disruptions D_i . In this case, we compute, from the inferred probability distribution over the user attentional states, the expected dollars a user would be willing to pay to avoid different communication events.

The curves in Figure 4b represent, from top to bottom, the expected costs associated with six different interruptions, including a telephone call, a pager, a full visual alert with audio chime herald, a thumbnail display with audio chime, a full visual alert without the chime, and a thumbnail display without chime. Figure 5 displays samples of full visual alerts (for news, email, and instant messaging notifications) generated by Notification Platform, employing additional models that balance the cost of interruption and information value. Decision-analytic models are used to consider an inferred expected cost for different messaging actions and the expected value of different communications, as assessed in a separate analysis of the value of information. We shall not review the details of this cost-benefit analysis in this paper. Rather, we shall focus on learning and reasoning about the expected cost of interruption from data about users' behavior and context.

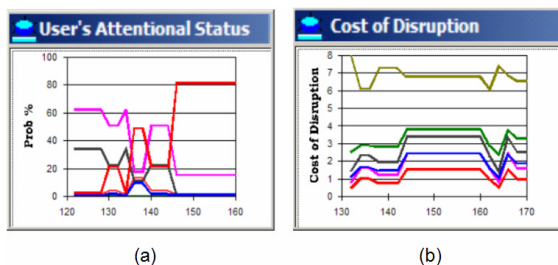


Figure 4. (a) Inference about the attentional state of the user over time. At the latest time, the dominant states, in order of likelihood are high-focus solo activity, conversation in office, and then low-focus solo. (b) Expected cost of interruption over time. Interruptions, sorted in descending order by expected cost, include (from the top) telephone call, pager, and full desktop herald with audio chime.



Figure 5. Sample visual alerts from Notification Platform, including news, financial alerts, email, instant messages, and output from services, such as a scheduling agent, background querying, and new document tracking. The service weighs the expected cost of interruption and the value of information in decisions about alerting and display modalities.

3. LEARNING PREDICTIVE MODELS FOR COST OF INTERRUPTION

Over the last several years, we have explored the construction of models of attention and interruption via manual knowledge acquisition, via machine learning from data, and combinations of direct assessment and learning. We shall focus here on the construction of personalized models via collection of streams of information, coupled with a process of tagging, machine learning, and evaluation. Our research in this realm has included the construction of models of attention with explicit temporal structure, including HMMs as well as more general dynamic Bayesian models, and models that represent temporal information intrinsically in the definition of variables.

We dwell in this paper on learning models that predict the state of interruptability of users in office settings. Such models are designed for making inferences in situations where a user has current or recent access to a computing system. Complementary work [6] on learning models of the cost of interruption based on a detailed consideration of distinctions about appointments drawn from an online calendar and time since a user has interacted with different applications or devices are reviewed in Section 6.

We shall consider the learning of models that predict the state of interruptability of users, where the detailed description of a user's attentional focus or workload remains implicit. That is, we bypass explicit, detailed representation about the state of a user, and pursue instead models that characterize the user's interruptability. We then map such inferences to a cost of interruption.

3.1 Interruption Workbench

We have developed a tool named the Interruption Workbench (IW) which provides facilities for event capture, annotation of segments of time, and building and testing statistical models of interruptability. A screenshot of the tagging tool being used during an annotation session is displayed in Figure 6.

3.1.1 Phases of analysis

The first phase of model building is *event and context capture*. During this phase, IW records streams of desktop, calendar, and available perceptual events. A video camera is employed to record

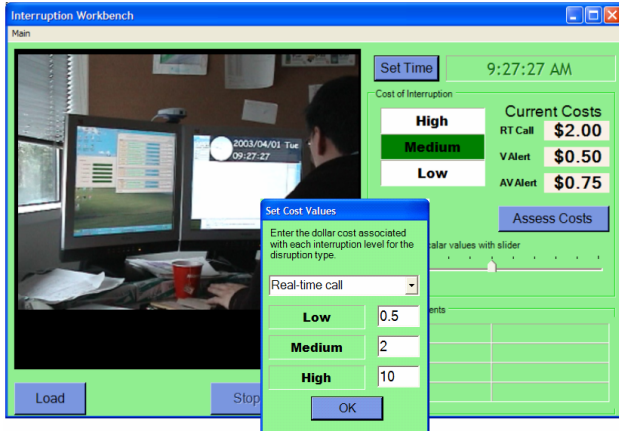


Figure 6. Screen from the Interruption Workbench. The tool captures and synchronizes sensed perceptual and client events with a video log of a user’s activities. Subjects can tag periods of time with cost of interruption for different types of interruptions. Foreground shows cost-assessment palette.

a subject’s activities and overall office context. The videotape with audio track is shot over the shoulder of subjects, revealing the content displayed on the user’s screen in addition to a portion of the user’s office environment.

The second phase of building models of interruptability is *tagging and assessment*. IW displays the video and synchronizes the video with events that were captured during the training session, enabling users to label segments of time by the state of their interruptability and to associate them with the constellation of sensed events. The labeling effort is minimized by allowing users to specify transitions among states of interruptability, rather than requiring users to label each small segment of time.

The tool provides a means for specifying the way that the variables representing the cost of interruptability are discretized and how cost is represented. Subjects can encode their assessments about their interruptability at different times in two ways, depending on their comfort. With the first method, subjects tag periods of time viewed on the video as *high*, *medium*, and *low* cost of interruption. As displayed in the foreground of Figure 6, users can separately map dollar values to each of the high-level states, for different kinds of interruptions (e.g., real-time telephone call, visual alert, audiovisual alert), reflecting the willingness to pay to avoid an interruption of each type during the states labeled as high, medium, and low cost of interruption. In a second approach to labeling time segments of a training session, subjects can define a scale and use a slider to tag periods of time with costs of interruption. These yield models that provide inferences about the probability distribution over real-valued values, representing the costs of interruption.

Finally, in the *generation and testing* phase, we construct and test a Bayesian network from the tagged library of cases generated in the first two steps. The task of tagging one or more sessions of office activity creates a database of two-second periods of time tagged with an interruptability label and containing a vector of logged event states. Given a set of tagged cases, the system can be instructed to build a classifier. We shall review the construction of Bayesian dependency models. These models

reveal the dependencies among key observations and variables. The Bayesian learning procedure employs graph structure search [1], and outputs a Bayesian network for the cost of interruption. The model can be used to make real-time predictions about the expected cost of interrupting users in different ways, given a live stream of sensed events obtained from the Event Whiteboard.

At run time, the probability distribution over the states of interruptability inferred by the model is used to compute the expected costs of interruption of different classes of interruption. For each disruption under consideration, we compute an expected cost of interruptability by invoking an expectation similar to the expected value calculation defined in Equation 1, substituting the likelihood of different states of interruptability, $p(I_i|E)$, for the explicit states of attention,

$$ECI = \sum_j p(I_j | E)u(D_i, I_j) \quad (2)$$

3.1.2 Beyond the Present Moment: Inferences about Future Cost of Interruption

In addition to reasoning about the current state of interruptability, we also generate several variables representing attentional forecasts about future state of interruptability. These include variables that capture inferences about the probability distributions over *times until* a low, medium, or high state of interruptability will be reached, and more specialized variables representing the times until states of interruptability will be achieved that *will persist for* different amounts of time. As an example, a variable in this family represents the time until a user will remain in a state of low cost of interruption for at least five minutes. Such predictions are important for deliberating about if, when, and how to mediate communications.

4. BUILDING AND TESTING MODELS

For learning models from data, IW provides tools for selecting a portion of training data for constructing a model and for testing the model on data held out for evaluating. Training data can be selected temporally as an initial segment of data or sampled randomly from the complete training set.

4.1 Sample Model

Figure 7 shows a Bayesian network model, output by the workbench. The model was built from a log of a subject’s (S1) activities tagged by cost of interruptability. For this model, the database of cases represents activities in the subject’s office, including the subject’s interactions with a personal computer, for five hours collected in one-hour sessions at different times. The database includes 9,384 two-second cases, representing 202 state transitions among interruptability levels. 85 percent of these cases were drawn randomly from the case library and used to build the model. The remaining cases were used for testing. To avoid overfitting, we first tuned a parameter, used in the Bayesian score for penalizing complexity, by splitting the training set into sub-training and holdout data sets to identify an optimal value of the parameter at a soft peak in the Bayesian score. This value was used to build the Bayesian network from the full training set.

We then test the abilities of the learned model to predict the outcomes in the test data set. We compute the classification accuracy for the learned and marginal models to characterize the power of these models. The classification accuracy is the likeli-

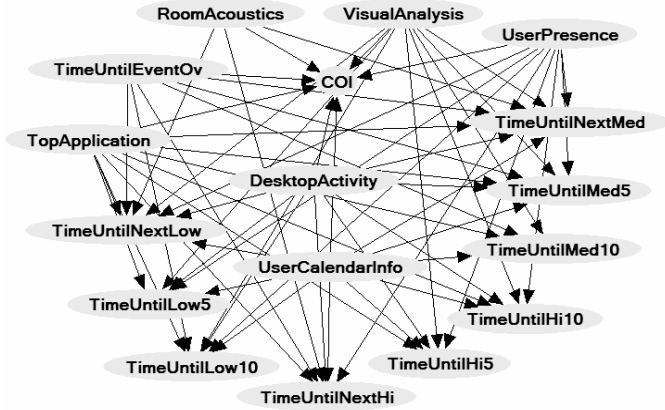


Figure 7. Bayesian network model learned from tagged data for training session. Model infers probability distribution over interruptability (COI), as well as predictions of time until reaching different future states of cost of interruption, and future states of interruption persisting for different times.

hood that the model will correctly identify the state of interruptability of the user. The variable in the Bayesian model representing the current state of interruptability (with states *Low*, *Medium*, and *High* cost) is labeled COI. Other variables include forecasts *Time Until Next Low*, *Time Until Next Medium*, *Time Until Next High*, and variants of these variables, representing the time until periods of low, medium, and high costs of interruption will be reached and persist for at least 5 and 10 minutes. In this case, the states of the forecasting variables are discretized into five time states, including *Less than one minute*, *1-5 minutes*, *5-10 minutes*, *10-15 minutes*, and *Greater than 15 minutes*.

Figure 8 displays a decision graph representing a compact encoding of the probability distribution underlying the COI variable of the Bayesian network. The bar graphs at the leaves of the tree represent probability distributions over *high*, *medium*, and *low* costs of interruptability (ordered, top to bottom, from high to low) for sets of observations represented by the paths leading to the leaves. The paths to the leaves identify important combinations of events for decomposing the probability distribution over COI and other variables into sub-distributions. The paths branch on key observations drawn from the user’s calendar and from the real-time activity event stream, including patterns of presence, application usage, and perceptual events. Similarly detailed trees, encoding variable state paths and associated probability distributions at the leaves, are produced for the attentional forecasting variables.

4.2 Evaluation of Learned Models

We have integrated tools into IW for testing the performance of sample models and for probing the discriminatory power of different features.

4.2.1 Accuracy of marginal and inferential models

Tables 1 summarizes the classification accuracies for COI, and for several of the attentional forecasting variables for the model for S1 displayed in Figure 7, and for a second model, built from data from another subject (S2). The data set for subject S2, also collected in 5 one-hour segments, is comprised of 8048 two-second cases, representing 926 transitions among states of cost of interruption. As in the procedure for S1, the model was

constructed by splitting the data 85/15 into training and test cases. The classification accuracies of the marginal models for each subject are listed under the variable names. The classification accuracies of marginal models are obtained by assuming the most likely states of the variable when testing predictions on the held-out data set. We found that the marginal models may provide good classification accuracy in some cases. In such situations, there may be little or only small lift in classification accuracy with the use of the inferential model. However, in most cases, the inferential model provides a discriminatory boost over the marginal model. For example, the inferential model, based on a consideration of all monitored events (column 1 of Table 1), yields a 0.73 accuracy for predicting COI for S1, while the marginal model provides a base classification accuracy of 0.53. For S2, the inferential model shows a 0.64 accuracy for predicting COI for S1, a boost over the marginal model’s classification accuracy of 0.37.

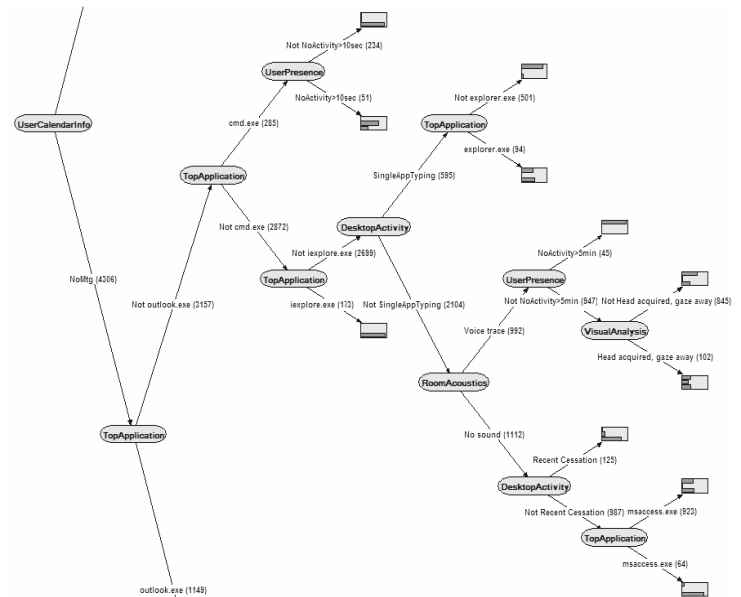


Figure 8. Portion of decision graph (for branch: *No Meeting, Not Outlook.exe as Top Applications*) encoding probability distribution for the COI variable of the Bayesian network in Figure 7. Bar graphs at leaves represents, from top to bottom, likelihood of high, medium, and low cost of interruption.

4.2.2 Probing information value with model ablation

Beyond testing the performance of models built from a consideration of all available observed events, we performed model-ablation studies, focusing on the sensitivity of classification accuracy to the loss of specific features and classes of features. We have been particularly interested in the sensitivity of the performance of the models with the removal of perceptual features from learning and inference. As many computers in use may not have acoustical and visual sensing capabilities, we have sought to better understand the base discriminatory power of events associated with rich patterns of desktop activity and from calendars. For the ablation studies, we constructed models in the same manner as the procedure described in Section 4.1 from the data from subjects S1 and S2. However, rather than using all observations, we omitted specific sets of features from the learning process.

Table 1. Classification accuracies for marginal and inferential models for current and future states of subjects' COI, showing contributions and synergies for system events, and acoustical (aud.), visual (vis.), and calendar (cal.) observations.

Attentional state	All events	Marg. model	Sys. only	Sys. + aud + vis.	Sys. + cal.
Current state of COI	S1: .73 S2: .64	.53 .37	.64 .60	.69 .63	.69 .63
Time until low	S1: .61 S2: .69	.35 .65	.51 .65	.58 .66	.58 .67
Time until med	S1: .68 S2: .82	.49 .76	.59 .76	.64 .81	.67 .76
Time until high	S1: .72 S2: .83	.66 .69	.67 .72	.70 .79	.71 .77

Table 2. Focus on classification accuracies for subsets of observations in the absence of the system event stream.

Attentional state	Cal. only	Aud. only	Vis. only	Vis. + aud.	Aud. + vis. + cal.
Current state of COI	S1: .53 S2: .44	.53 .54	.53 .46	.55 .54	.57 .58
Time until low	S1: .37 S2: .65	.35 .65	.35 .65	.38 .65	.42 .67
Time until med	S1: .58 S2: .76	.49 .76	.54 .76	.54 .76	.63 .78
Time until high	S1: .66 S2: .69	.66 .70	.66 .69	.65 .72	.67 .74

Table 1 shows the classification accuracies for models built from all events, system events only, system events and perceptual observations, and system events and calendar information. We found varying sensitivities of classification accuracy for different variables under consideration and different kinds of ablations. Such sensitivities also differed for the models for subjects S1 and for S2. Table 2 reviews the accuracy of models relying solely on calendar, acoustical, visual, and on acoustical, visual, and calendar events together. Although, lifts are not as high as provided by models that include system events, synergistic contributions among calendar, acoustical, and visual information sources is apparent. Overall, for the majority of cases where inferential models provided boosts over the marginal models, we found that classification accuracy of a base model, employing only system events, is dominated by models that also include either perceptual observations or calendar information. Adding a consideration of both classes of observations to the base model led to the most accurate predictions.

Beyond studies of classification accuracy, we can investigate the structural influences on predictive models of adding and removing different classes of observations. We can examine changes in variables and dependencies of the overall Bayesian network, or can explore the sensitivity of the structure of the decision graphs encoding the probability distributions for COI and COI forecasting variables. As an example, Figure 9 displays a refinement of a branch of the tree encoding the probability distribution for COI for S1 that was particularly sensitive to the introduction of audio and visual sensory information, given a base model of system events and calendar information.

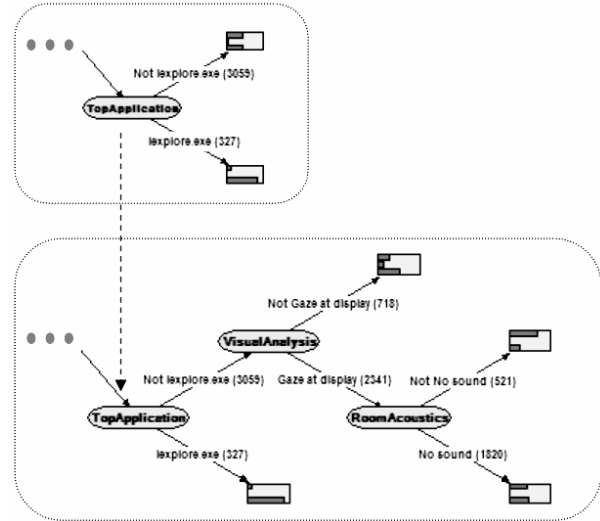


Figure 9. Sensitivity of a branch of decision graph for COI to gaining access to visual and acoustical sensory streams.

Introducing gaze and acoustical information enables the distribution at the former leaf in the tree (ending in the identification that *Internet Explorer is not in focus* at the moment) to be specialized into subdistributions that hinge on the status of visual and acoustical observations. Beyond such modular refinements, we noted situations where the removal of classes of features led to larger-scale changes in structure. For example, the loss of perceptual features may lead to shifts in the specific desktop activities being considered at key branches in the trees (e.g., removing vision and acoustics may lead to the introduction of *Outlook vs. Not Outlook in Focus* as a central consideration).

The decision graph in Figure 10, composed from data from S1 for system events and perceptual information, provides some insight about the use of visual pose information in predicting the time until the start of the next low cost of interruption period. The tree highlights how observations from a head tracking system about presence and pose can influence the probability distribution over the time until next low cost of interruption. Figure 11 shows the tree for the same prediction for the case where we have removed consideration of visual and acoustical information. The system is forced to rely on desktop activity and presence for making the forecast of time until low cost of interruption.

5. MODEL TRANSFER AND REUSE

In addition to developing and evaluating tools for generating personalized models that predict the cost of interruption, we are interested in the prospect for developing models and methods that can provide useful inferences to users with little or no effort. One goal of efforts on machine learning in this vein is the identification of important variables for determining the cost of interruption. Although we found that the detailed parameters and structure of predictive models shifts significantly from user to user, we found that specific sets of variables appear important across users. For example, information about the application that is currently active and in focus, and patterns of activity and quiescence are often important, early branches in decision graphs for COI. Machine learning can assist designers with formulating predicates, languages, and user controls that could enable people

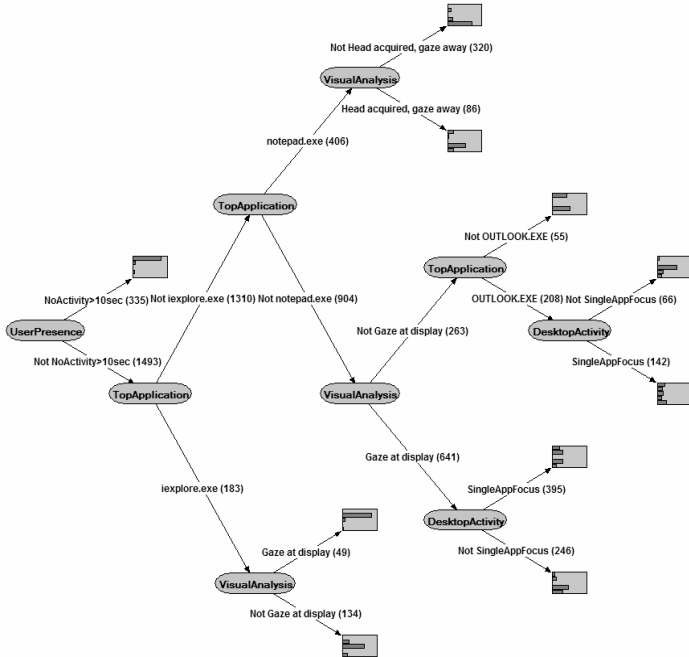


Figure 10. Decision graph for predicting the time until a user will next be in a low cost of interruption state, based on system events and perceptual sensing.

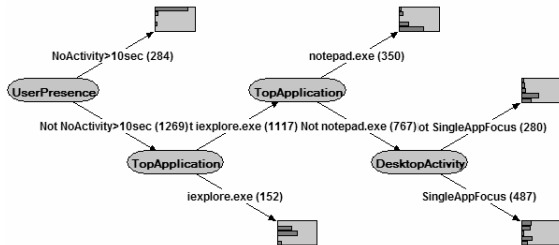


Figure 11. Decision graph for predicting time until a user will next be in a state of low cost of interruption, when perceptual events are removed from consideration.

to specify, in an efficient manner, policies that define the cost of interruption as a function of such variables.

In another approach to reducing the need for special training, we have been investigating prospects for applying models trained on one or more users to other users. We found that applying a personalized model from one user to predict the outcomes of another user may yield poor performance. As an example, Table 3 demonstrates the results of applying the predictive model developed for subject S1 to the test cases for subject S2 and vice versa. The performance is poor for the case of using all events and selected subsets of events, where inferential models can show poorer classification accuracy than the marginal models.

We found better performance with the construction of composite models from the training data of multiple individuals. Table 4 shows the performance of such a generalized model, constructed from the training data of subjects S1, S2, and a third subject, S3 who also provided 5 hours of data. Although the model does not perform as well as the users' own personalized models, we see significant boosts in classification accuracy, when compared with the simple substitution of models among users.

Table 3. Performance for cross-user classification of COI.

Attentional state	All events	Sys. only	Aud. + vis. + cal.	Aud. only	Vis. only	Cal. only
Current state of COI	S1→S2: .28 S2→S1: .32	.39 .35	.30 .25	.37 .34	.37 .38	.37 .31

Table 4. Performance of a composite model constructed from data from multiple subjects.

Attentional state	All events	Sys. only	Aud. + vis. + cal.	Aud. only	Vis. only	Cal. only
Current state of COI	All→S1: .55 All→S2: .66	.38 .60	.42 .54	.34 .54	.33 .34	.34 .31

As highlighted by the classification accuracies displayed in Table 4, multiple classes of observations work in a synergistic manner to boost the performance of the composite models. In contrast, combining multiple classes of features for the model substitution situation tends to diminish the classification accuracy, as captured by the results in Table 3.

6. RELATED RESEARCH

We have been exploring related challenges and opportunities via other ongoing projects. On the Coordinate project [6], we have explored the construction of models of the cost of interruption based on detailed appointment information. The Coordinate effort centers on the use of machine learning about presence and availability based on such events as the time of day, the user's current or recent presence on devices, a user's last sensed location, and details about meeting information drawn from a user's online calendar. The Coordinate server performs real-time machine learning and inference in response to ongoing, heartbeat queries or in response to special queries by a trusted colleague or communications agent. The system computes the cost of interruption now and in the future, considering multiple properties of a meeting drawn from the online calendar (number of attendees, organizer, location, duration, subject, etc.). In contrast to the work described here, Coordinate was initially developed to reason about forecasts about a user's location and availability when the user is out and about, rather than the detailed case of models of attention in an office setting, where we have access to the rich stream of data from a desktop system and perceptual sensing. Nonetheless, we have been working to unify the detailed office analysis with the detailed analysis of meetings and location provided by Coordinate. Beyond integration of the two kinds of models and modeling methodologies, there is opportunity for integrating the detailed meeting distinctions into the inferences about attention in an office setting. In contrast to the Coordinate work, the modeling efforts described in this paper employ only a few high-level distinctions about the likelihood and duration of a meeting in the office, and rely more centrally on system interactions and perceptual information.

Other relevant work includes the Seer effort on distinguishing among different office situations, such as identifying from perceptual information whether a user is on the telephone or in a conversation with other people in their office [12]. There is an

opportunity for integrating the inferences provided by systems like Seer into the Interruption Workbench tool set.

In other related work, a team at Carnegie Mellon has recently performed a Wizard of Oz study in an attempt to understand the value of different features in predicting the interruptability of several subjects [9]. In the study, the users' interruptability was probed with a "beeper study" methodology, intermittently seeking feedback from users about their interruptability. Features were coded by researchers and classifiers were constructed from the coded data in an attempt to identify the value of different observations. The researchers identified such features as whether the user is speaking, writing, sitting, or interacting with objects such as a keyboard or phone, the presence and activities of occupants, and whether the user's office door is open or closed. A list of top features was identified, including different kinds of talking, positions and configurations of people, and interaction with a keyboard. The work complements our ongoing research on modeling interruptions; features identified by the Wizard of Oz effort may provide valuable guidance as the tools are refined.

7. SUMMARY

We have described effort to build models that can predict the cost of interrupting users. We presented research on harnessing machine learning to generate statistical models for inferring the state of interruptability of users. We focused on a methodology and tools for logging and tagging a database of cases and discussed our attempts to probe the sensitivity of the classification accuracy and structure of the learned models in response to overlooking classes of observations. Finally, we touched on our investigation of the potential for building transferable models of interruptability and assessment tools, in pursuit of systems that could provide users with valuable attention-sensitive services without requiring costly training procedures. Beyond focusing solely on the opportunity for integrating statistical models in fielded systems, we discussed the potential value of leveraging insights from machine learning to identify important features that might be manipulated directly by users in crafting statements about their interruptability in different settings.

In ongoing research, we are continuing to investigate the discriminatory value of considering additional observational features about the environment and activities of users, drawn from computing devices and perceptual apparatus. Beyond exploring new system events, appointment information, and enriched acoustical and visual features, we are investigating the discriminatory power of other sensors such as accelerometers, proximity, and motion detectors [4]. We are also continuing to gather data from multiple subjects in pursuit of answers to several questions about the power and generality of the models we can build from the data. We are particularly interested in studying the accuracy of predictions of the cost of interruptability in settings where we build models with data from multiple users and use these models to predict the interruptability of new users. Characterizing the accuracy of such composite models promises to provide insights about the possibility for minimizing the need for custom-tailored training.

ACKNOWLEDGMENTS

We thank the subjects who have participated in our studies.

REFERENCES

- [1] Chickering, D.M., Heckerman, D. and Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 80-89.
- [2] Cutrell, E., Czerwinski, M. and Horvitz, E. (2001). Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance, *Proceedings of Interact 2001*, pp. 263-269.
- [3] Gillie, T. and Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity and complexity *Psychological Research*, 50, 243-250.
- [4] Hinckley, K., Pierce, J., Sinclair, M. and Horvitz, E. (2000). Sensing Techniques for Mobile Interaction, In *Proceedings of the ACM UIST 2000*, pp. 91-100.
- [5] Horvitz, E., Jacobs, A., and Hovel, D. (1999). Attention-Sensitive Alerting. In: *Proceedings of the Fifteenth Conference on Uncertainty and Artificial Intelligence*, pp. 305-313.
- [6] Horvitz, E. Koch, P., Kadie, C.M. Jacobs, A. (2002). Coordinate: Probabilistic Forecasting of Presence and Availability. *Proceedings of the Eighteenth Conference on Uncertainty and Artificial Intelligence*, pp. 224-233.
- [7] Horvitz, E., Kadie, C. M., Paek, T., D. Hovel. (2003). Models of Attention in Computing and Communications: From Principles to Applications, *Communications of the ACM* 46(3):52-59.
- [8] Hudson, J.M., Christensen, J., Kellogg, W.A., and Erikson, T. (2002). "I'd Be Overwhelmed, But It's Just One More Thing to Do:" Availability and Interruption in Research Management.' *Proceedings of CHI 2002*, pp. 97-104.
- [9] Hudson, S.E., J. Fogarty, C.G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J.C. Lee, J. Yang (2003). Predicting Human Interruptability with Sensors: A Wizard of Oz Feasibility Study, *Proceedings of CHI 2003*.
- [10] McCrickard, D.S., Catrambone, R., Chewar, C. M., and Stasko, J.T. (2003). Establishing Tradeoffs that Leverage Attention for Utility: Empirically Evaluating Information Display in Notification Systems. *International Journal of Human-Computer Studies*.
- [11] McFarlane, D. (1999). Coordinating the interruption of people in human-computer interaction. *Proceedings of Interact '99*, pp. 295-303.
- [12] Oliver, N., Horvitz, E., Garg, A. (2002). Layered Representations for Recognizing Office Activity, *Proceedings of ICMI 2002*, pp. 3-8.
- [13] Ovsiankina, M. (1928). Die wiederaufnahme unterbrochener handlungen. *Psychologische Forschung*, 11:302-379.
- [14] Toyama, E. and Horvitz, E. (2000). Bayesian Modality Fusion: Probabilistic Integration of Multiple Vision Algorithms for Head Tracking. *Proceedings of Fourth Asian Conference on Computer Vision*.
- [15] Zeigarnik, B. (1929). Das behalten erledigter und unerledigter handlungen. *Psychologische Forschung*, 9:1-85.

