Learning to Predict Information Needs: Context-Aware Display as a Cognitive Aid and an Assessment Tool

Bradley C. Love[†], Matt Jones[‡], Marc T. Tomlinson[†], Michael Howe[†]

[†]University of Texas at Austin
Austin, TX 78712 USA
{brad_love,tomlinson,howe}@mail.utexas.edu

[‡]University of Colorado
Boulder, CO 80309 USA
mcj@colorado.edu

ABSTRACT

We discuss the problem of assessing and aiding user performance in dynamic tasks that require rapid selection among multiple information sources. Motivated by research in human sequential learning, we develop a system that learns by observation to predict the information a user desires in different contexts. The model decides when the display should be updated, which is akin to the problem of scene segmentation, and then selects the situationally relevant information display. The model reduces the cognitive burden of selecting situation-relevant displays. We evaluate the system in a tank video game environment and find that the system boosts user performance. The fit of the model to user data provides a quantitative assessment of user behavior, which is useful in assessing individual differences and the progression from novice- to expert-level proficiency. We discuss the relative benefits of adopting a learning approach to predicting information preferences and possible avenues to reduce the negative consequences of automation.

Author Keywords

Context-Aware Computing, Adaptive Display, Experimentation, Learning, Event Segmentation

ACM Classification Keywords

Human Factors, Experimentation, Learning

INTRODUCTION

We increasingly find ourselves in information-rich environments. Often, many information sources are potentially useful for completing a task. For example, in coordinating disaster relief, sources of potentially useful information include video feeds, weather forecasts, inventories of relief supplies, GPS tracking of support vehicles, etc. Likewise, the many sensors, gauges, and navigation systems in a modern automobile are potentially useful to the driver.

CHI 2009, April 4 - 9, 2009, Boston, MA, USA.

One key challenge people face is identifying which source of information is desired at the current moment. Although the information available to a human operator can increase without obvious bound, our basic information processing capacities remain fixed. Each additional information source incurs a cost to the human operator by increasing the complexity of the selection process. As informational channels are added, at some point, the marginal costs (in terms of cognitive load) eclipse the marginal benefits. Indeed, one distinguishing aspect of human expertise may be the ability to rapidly assess which information is relevant and settle a plan of action [17].

In this report, we propose and evaluate a system that eases this selection process by highlighting the information channel desired by the user. The system, Responsive Adaptive Display Anticipates Requests (RADAR), learns to approximate the selection process of the human operator by observing the user's selection behavior. In cases where RADAR successfully approximates the human's selection process, the cognitive cost of information selection can be offloaded to RADAR.

RADAR is named after the character Walter "Radar" O'Reilly from the television series M*A*S*H. Radar O'Reilly had an uncanny ability to deliver information to his commander moments before the commander formulated his request, much like how RADAR learns to anticipate the information needs of the user to reduce cognitive load. In a series of well-controlled experiments, we evaluate RADAR's ability to increase situation awareness and thereby improve performance. We then evaluate whether RADAR's quantitative fits of individual performance provide a useful means for assessing expertise and individual differences.

RELATED WORK

The topic of plan recognition in AI is concerned with correctly attributing intentions, beliefs, and goals to the user. Plan recognition models tend to subscribe to the Belief-Desires-Intention framework [24]. This line of work relies on knowledge-based approaches for user modeling and encoding insights from domain-specific experts [10]. These approaches can involve identifying a user's subgoals through task-analysis [40]. Once a user's beliefs, intentions, and goals are understood, display can be adapted appropriately [10].

Instead of focusing on identifying the internal state of the

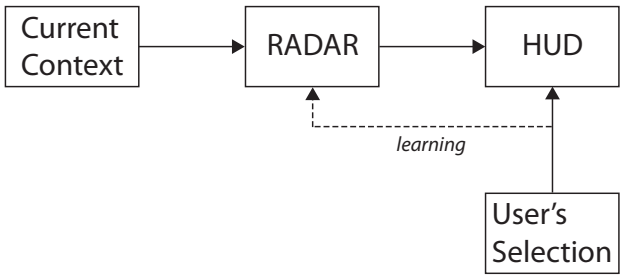

Figure 1. RADAR takes as input the current context (e.g., recent game history) and outputs its preferred display to the HUD. The user (e.g., the game player) can override RADAR's choice. Such corrections serve as learning signals to RADAR and increase the likelihood that RADAR will select the user's preferred display in similar situations in the future. Over time, RADAR approximates the information preferences of a specific user, allowing the user to offload the task of selecting the relevant information source (i.e., display) from numerous competing options.

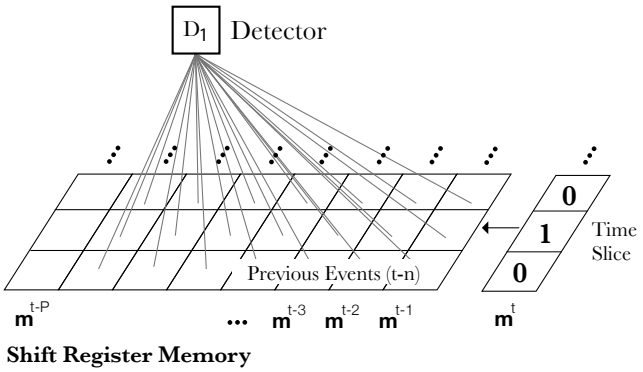

Figure 2. RADAR utilizes a buffer network to represent and learn from recent context (e.g., game history). Context is represented as a series of time slices. The tank game results are based on a context consisting of ten time slices of 250 ms each. The buffer functions as a shift register — the slice from the immediate time step enters one side of the buffer, all other time slices shift over one slot to accommodate the new entry, and the least recent time slice is removed from the buffer. Each time slice consists of a feature vector describing the current situation. Table 1 lists the features used for the tank game. Each possible display in the HUD has a detector that collects evidence to determine whether it is the situationally appropriate display. Association weights between features at various positions along the buffer and each detector are learned through error correction learning. For example, if a user prefers to have the fuel scope displayed when fuel is low, the weight from the fuel level feature's low value at various positions along the buffer to the fuel scope display detector will develop large, positive weights.

user, other approaches rely on input from domain experts to adapt display to emphasize the information to which the user *should* attend. For example human experts can label episodes and these episodes can serve as training instances for machine learning models that prioritize display elements [34]. Alternatively, input from human experts can be used to build expert systems or Bayesian models to prioritize display [13].

Our approach diverges from the aforementioned work. Rather than prescribe which information source a user should prioritize, we attempt to highlight the information a user would select if the user searched through all possible options. Our approach may be preferable in domains where it is unclear what is normative. Unlike work in plan recognition, we sidestep the problem of ascribing and ascertaining the user's internal mental state. Instead, RADAR learns to directly predict a user's desired display from contextual (i.e., situational) features. We do not deny that a user's explicit beliefs, desires, and intentions are important for determining information preferences. Rather, we suggest that some important aspects of cognition are grounded in lower-level mechanisms that are not effectively assessed through introspection and direct questioning. Furthermore, many higher-level beliefs may be embodied in terms of the display choices that people make in the environment. Thus, the correlates of some higher-level beliefs may be directly observable in users' actions. Our studies test these general notions by evaluating how successful a system can be in the absence of explicit representations of users' beliefs and intentions.

Our approach emphasizes learning as opposed to preprogrammed interfaces [22]. Adopting a learning approach to adaptive display has a number of positive consequences, including the ability to take into account individual differences across users [31]. Another positive consequence is that minimal input from subject matter experts is required to build a system. Like other context-aware applications that adopt a keyhole approach [2, 38], our approach infers a users preferences without interfering with or directly querying the user [15]. Interfaces that highlight recently selected menu items follow a similar logic [9], though the work we will propose is more open ended in terms of possible predictors and learnable relationships from predictors to display preferences.

Rather than anticipating a user's information needs like RADAR does, related work aims to predict when a user can be interrupted by a new task, such as a phone call [8, 14, 16]. However, work on the cost of user interruption may bear on RADAR's first decision stage (discussed below), which determines when to introduce new information. Additionally, models of user interruptibility provide information about the user's state that may be predictive of display preferences. Therefore, the outputs from these models, along with other measures of cognitive load, could serve as valuable inputs to RADAR.

OVERVIEW OF RADAR

RADAR is designed to operate in task environments in which the user must select which display among numerous displays to monitor. For example, we evaluate RADAR in an arcade game environment in which players select which of eight possible displays to show on a Head-Up Display (HUD). Figure 1 illustrates how RADAR operates in such task environments. RADAR takes as input the current context (e.g, recent game history) encoded as a feature vector and outputs to the HUD the display it thinks the user wishes to view. The user is free to override RADAR's choice. RADAR learns from the user's acceptance or rejection of its display choices and over time converges to selecting the displays the user desires. Alternatively, RADAR can observe and learn to mimic a user's display preferences offline. After online training, RADAR can be used to select displays. In the studies reported here, offline training was used.

RADAR employs a two-stage stochastic decision process at every time step. In the first stage, RADAR estimates the probability that a user will update the HUD given the current context. When the sampled probability from the first stage results in a display update, RADAR proceeds to the second stage (otherwise the current display remains unchanged). In the second stage, RADAR estimates the probability distribution for the next display choice given the current context, and samples this probability distribution to select the next display.

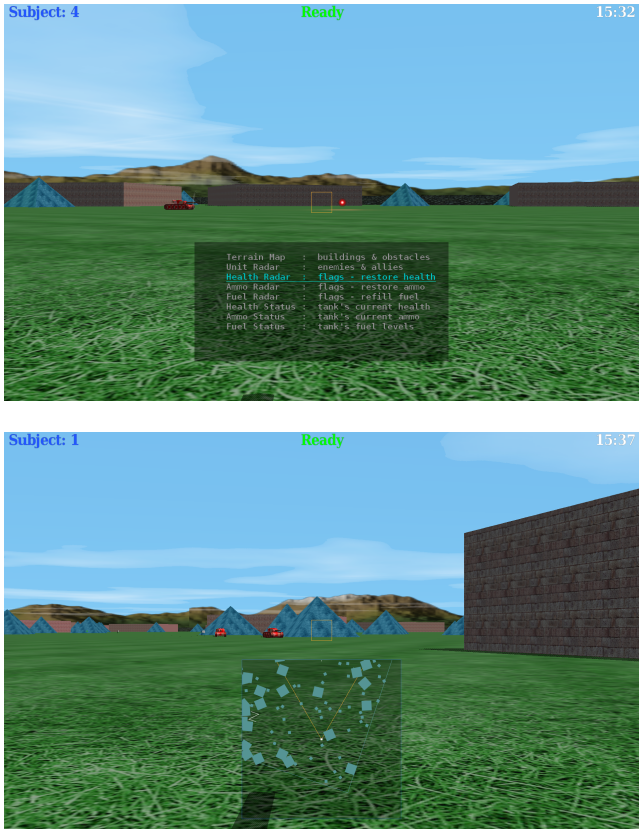

Figure 3. Screenshots from our modified version of the BZFlag tank game are shown. The top panel shows the selection menu listing the eight possible displays from which players can choose. These eight possible displays correspond to the first eight features listed in Table 1. Once a display is selected, the menu is replaced with the chosen display in the HUD, as shown in the bottom panel. Players can offload the task of selecting relevant displays to RADAR.

The motivation for the two-stage approach is both computational and psychological. Separating display prediction into two stages improves RADAR's ability to predict display transitions. The same display currently desired is highly likely to be desired in 250 ms. This constancy would dominate learning if both stages were combined. The second stage's focus on display transitions allows for improved estimation of these relatively rare, but critical, events.

Psychologically, the first stage corresponds to identifying key events in a continuous (unsegmented) environment, whereas the second stage corresponds to predicting event transitions. To make an analogy to speech perception, people segment the continuous speech stream into words (akin to RADAR's first stage) in the absence of reliable acoustical gaps between words [29]. Akin to RADAR's second stage, people anticipate which word (i.e., event) is likely to follow given the preceding words [23].

One view is that event segmentation serves an adaptive function by integrating information over the recent past to improve predictions about the near future (see [21], for a review). In support of this view, individuals who are better able to segment ongoing activity into events display enhanced memory [41]. People's judgments of event boundaries are reliable [33] and tend to show high agreement with others [26]. For example, two people watching a person make a peanut butter and jelly sandwich will tend to agree on the steps involved. These two people will also both segment off surprising or unexpected events, like the sandwich maker dropping the sandwich on the floor.

Behavioral measures reveal that cognitive load increases at event boundaries. Reading speed slows when event boundaries are crossed [28, 42]. Recognition for objects in picture stories, virtual reality, and movies becomes worse when an event boundary is crossed [27, 35]. In addition to these behavioral measures, neurophysiological measures track event boundaries. Events boundaries are associated with increased activity (as measured by fMRI) in bilateral posterior occipital, temporal, and parietal cortex, along with right lateral frontal cortex [33]. EEG measures corroborate these findings [32]. Furthermore, pupil dilation and increased frequency of saccades are associated with crossing event boundaries [36]. One hypothesis is that RADAR will benefit users by updating display at event boundaries because cognitive load, environmental change, and uncertainty are highest at such times. In Study 4, we assess whether display updates occur at event boundaries.

The probability distributions associated with both stages (event segmentation and event prediction) are estimated by simple buffer networks [6]. As shown in Figure 2, buffer networks represent time spatially as a series of slots, each containing the context (e.g., game situation) at a recent time slice, encoded as a feature vector. The buffer allows both ongoing events and events from the recent past to influence display prediction. Despite their simplicity, buffer networks have been shown to account for a surprising number of findings in human sequential learning [12]. At each time step, weights from the buffer are increased from activated features to the display option shown in the HUD, whereas weights to the other display options are decreased. Over time, this simple error correction learning process approximates a user's information preferences. RADAR's weights can be used to assess individual differences and user performance.

RADAR'S FORMAL DESCRIPTION

Player Model

Our model of the player's choice behavior assumes that the player's preferred channel at any time, t, is determined by the state of the game at that time, S^t, together with the recent history of the game, $(S^{t-l})_{1 \leq l < L}$. The recent history

is included, in addition to the current state, to allow for fixed delays in information need (e.g., the player wants to see channel Y, l timesteps after event X occurs). The parameter L determines the maximum delay, that is, the longest time that past information can remain relevant to the player's choice. Increasing this parameter initially improves system performance, though eventually performance declines as the ratio of data points to tunable weights becomes small. The choice of $L = 10$ (i.e., 2.5 s) for the applications described here attempts to balance these constraints.

For compactness, we write the sequence of current and recent game states as

$$\mathbf{S} = (\mathbf{S^{t-l}})_{\mathbf{0 \leq l < L}} \quad (1)$$

Because changing channels incurs a cost in terms of attention and motor resources, we do not assume that the player changes the HUD to his or her preferred channel whenever that preference changes. Instead, we assume a two-step stochastic process, in which at every timestep there is a probability that the player will change channels and, if the channel is changed, a probability distribution over the channel to be selected. The probability of switching channels is given by

$$p^t_{change}(C^t, \mathbf{S}) = \mathbf{P}[\mathbf{change(t+1)}|\mathbf{C^t}, \mathbf{S}] \quad (2)$$

where C^t is the current channel. If the player does change channels, the probability of selecting channel j is equal to

$$p^t_{choice}(j, \mathbf{S}) = \mathbf{P}[\mathbf{C^{t+1}} = \mathbf{j}|\mathbf{change(t+1)}, \mathbf{C^t}, \mathbf{S}] \quad (3)$$

Context Representation

The state of the game at any time, t, is represented by a vector of F feature values:

$$\mathbf{S^t} = (\mathbf{S^t_f})_{\mathbf{1 \leq f \leq F}}$$

These features used in the studies reported here are listed in Table 1. Continuous features are discretized, and all features are coded to take on values $0 \leq S_f < V_f$ (where V_f is the number of possible values of feature f).

Prediction

The display system operates by predicting two sets of probabilities, corresponding to the two steps in the model of the player's choice behavior: p_{change}, the probability that the player will change channels; and p_{choice}, the distribution over the new channel if the channel is changed. Both types of probabilities are predicted from the information in the game history, $\mathbf{S}$. The system learns a separate set of weights $\mathbf{w}$ for the two types of predictions, each indexed by the current channel (C^t), feature (f), value for that feature (v), and lag (l); the weights for p_{choice} are also additionally indexed by the value of the candidate new channel (j). The system's predictions are derived as a linear combination of these weights with the feature-value activations, $\mathbf{a^t}$, currently in the buffer:

$$p^t_{change}(C^t, \mathbf{S}) = \sum_{\mathbf{f,l,v}} \mathbf{w^{change}_{C^t,f,l,v}} \cdot \mathbf{a^t_{f,l,v}} \quad (4)$$

$$p^t_{choice}(C^t, j, \mathbf{S}) = \sum_{\mathbf{f,l,v}} \mathbf{w^{choice}_{C^t,j,f,l,v}} \cdot \mathbf{a^t_{f,l,v}} \quad (5)$$

Operation

At each timestep the system changes the channel with probability $p_{change}(C^t, \mathbf{S})$. When it does change the channel, it selects the channel j that maximizes $p_{choice}(C^t, j, \mathbf{S})$ subject to $j \neq C^t$.

Learning

The weights $\mathbf{w^{change}}$ and $\mathbf{w^{choice}}$ are computed from the player's manual choice behavior, by minimizing the following error terms:

$$E^{change} = \begin{cases} (p_{change})^2 & C^{t+1} = C^t \\ (1 - p_{change})^2 & C^{t+1} \neq C^t \end{cases} \quad (6)$$

$$E^{choice} = \left[1 - p_{choice}(C^{t+1})\right]^2 + \sum_{j \neq C^t, C^{t+1}} p_{choice}(j)^2 \quad (7)$$

The former is summed over all timesteps, and the latter is summed over all timesteps on which the player changed channels ($C^{t+1} \neq C^t$). In practice, the weights in RADAR's buffer networks are estimated directly and efficiently using optimized linear algebra routines rather than trial-by-trial error correction procedures. Both methods converge to the same solution, but trial-by-trial learning takes longer to do so.

Model Variants

In addition to the model formalized above, we have explored a variety other frameworks that instantiate RADAR's guiding principles, including Bayesian models and logistic regression. The results presented in this paper were based on the formalism presented above, but we have achieved similar results using other variants of the model.

Prescience

Others have hypothesized that information should be provided "just ahead" of the need [15]. We provide a testbed for such notions. RADAR is trained so as to predict players' display-selection behavior in advance of when that behavior would actually occur. This is accomplished by shifting the channel values relative to the feature values in the training set. The sequence of channel values selected by the player (i.e. on all timesteps in the model's training dataset) is moved earlier by τ steps, which effectively teaches the model to predict players' behavior τ steps into the future. Thus, when allowed to control the display, the model is able to immediately select the player's (predicted) preference τ steps into the future. The shift, τ, is currently set to 2 timesteps, i.e. 500 ms.

EVALUATING RADAR

Evaluating context-aware systems is challenging. Real-world studies are often impractical and difficult to properly control. Video game environments offer a number of substantial advantages for evaluation [3]. Our environment is a synthetic environment in that we aim to abstract functional relationships that we hope generalize to numerous actual operational environments [11]. Our synthetic environment is not intended to be realistic of an actual environment. Rather it is intended to allow us to test basic principles that generalize broadly. Unlike other studies involving context-aware systems, our studies provide rich, objective measures that can be quantitatively assessed, as opposed to relying on subjective self-report measures provided by subjects [4].

RADAR was evaluated in a video game task environment in which human players battled robot tanks. The task environment was adapted from the open source BZFlag 3D tank battle game (see www.bzflag.org). Modifications to BZFlag included expanding the state of a player's tank to include limited ammunition, fuel, and health. Players could pick up corresponding flags in the game to replenish these assets. Additionally, the display was modified to include a pop-up menu that allowed players to select one of eight possible displays to view on the HUD.

The eight possible displays for the HUD correspond to the first eight features listed in Table 1. Three of the displays provided the levels of the aforementioned assets. Three other displays were player-centered scopes that indicated the location of flags to replenish the corresponding asset. The remaining two displays consisted of a terrain map and a line-of-sight unit radar that provided the positions of enemy tanks and fire when not obscured by building structures. Figure 3 illustrates the menu for selecting which display to send to the HUD display as well as an example HUD.

RADAR's task was to anticipate the displays a player wished to have shown on the HUD, thus allowing the player to offload display selection to RADAR and devote full attention to game play. Successful game play requires maintaining situation awareness of the state of one's tank, the locations of flags to replenish assets, and the position of enemy tanks. Our prediction is that RADAR will improve players' situation awareness and performance by providing information at the appropriate time.

Below, we discuss results from a series of studies comparing player performance under RADAR to various controls. In each study, subjects were evaluated in game situations involving two enemy (robot) tanks. A game ended when the subject's tank was destroyed. When an enemy tank was destroyed, it was replaced by a new enemy tank at a random location. In between-subjects designs, subjects were randomly assigned to condition. In within-subjects designs, condition order was randomized across games. Players were recruited from the University of Texas's undergraduate population and participated in only one study. In all studies, experimenter and subjects were blind to condition. A typical game lasted around one minute.

Table 1. The features used to describe the current game context are listed. These features serve as inputs to RADAR. From these inputs, RADAR predicts which display the user wishes to view. The first eight features encode which channel was shown on the HUD (not the value of the displayed information).

Feature Type	Feature Name	
Display Shown (1-8)	Terrain Map	Unit Radar
	Ammo Status	Ammo Scope
	Health Status	Health Scope
	Fuel Status	Fuel Scope
Tank Condition (9-12)	Ammo Level	Health Level
	Fuel Level	Out of Fuel
Flag in View (13-16)	Any Flag	Ammo Flag
	Health Flag	Fuel Flag
Flag Picked Up (17-20)	Any Flag	Ammo Flag
	Health Flag	Fuel Flag
Dynamic/Battle (21-23)	Tank is moving	Tank hit
	Number of enemy tanks in view	

Overview of Studies

The studies presented here examine whether and how adaptive display aids performance and its utility in assessing user behavior. Study 1 served as an initial test of whether our adaptive display approach can improve users' task performance. Study 2 uses the same paradigm to assess whether RADAR promotes situation awareness. The remaining studies focus on issues revolving around individual differences and expertise. Study 3 evaluates the benefits of personalized models and whether RADAR's automation is preferable to purely manual operation. Study 4 compares the RADAR models for subjects at the novice and expert stage of development. This study also evaluates whether display updates occur at event boundaries. Study 5 evaluates RADAR's promise as an assessment tool by testing whether a user's pattern of display choices, as assessed by RADAR, can predict the user's task performance.

These studies are intended to guide RADAR's development and evaluate its promise for an array of real-world applications. Our synthetic task environment is demanding of both perceptual and cognitive resources and unfolds in real time. The task is engrossing and intensive, such that players continue to show improvement after one hundred hours of play. The environment is sufficiently complex for strong individual differences to be manifested. Like many real-world tasks, information relevancy in the game is situationally determined. To make a real-world parallel, a pilot may desire information about flap position at take-off and landing, but not during other portions of the flight. Likewise, a smart-phone user may welcome an unsolicited review of a nearby restaurant when (and only when) the user does not have dinner plans and it is dinner time. Our experimental environment embodies these aspects of real-world tasks.

We claim that these studies provide a general evaluation of RADAR. However, such a claim would be undermined by carefully tuning RADAR's features to yield the best results. For our task environment, we gathered features from volunteer players' verbal reports, as opposed to selecting features to improve RADAR's performance. As can be seen from an inspection of Table 1, the features are fairly rudimentary. Features include information about the basic state of the ve-

hicle (e.g., how much fuel is left) and events (e.g., whether the player has been hit by the enemy). Interestingly, in the studies that follow, we find that the best predictors of display preferences are which displays were previously viewed (e.g., the first eight features in Table 1). These features related to display use, along with other basic features tied to the tank's conditions (e.g., remaining fuel, see Table 1) could be determined in real-world applications without any additional sensors or signal processing. To test the robustness of our approach, only these features are incorporated in Study 3. Finally, Study 4 details a method for automatically determining which features are relevant to an individual from a large candidate set. In summary, we have constructed an environment and feature set that is intended to provide a strong evaluation of RADAR's potential for a number of real-world applications.

Study 1: Group Model Effects on Task Performance

Methods

Five undergraduate student volunteers in the laboratory served as the research subjects. These students each had over ten hours experience playing the tank game without RADAR operational (i.e., all displays were manually selected from the menu). Because this is the first evaluation of RADAR, the procedure was simplified to the greatest extent possible. RADAR's weights were estimated while users played without a functioning adaptive display (i.e., all display choices were determined by the subject), as opposed to incrementally training RADAR online. To further simplify evaluation, a single set of weights that predict the average display preferences of the group was calculated, as opposed to deriving a separate set of predictive weights for each subject. Thus, at test, each subject interacted and was evaluated with the same version of RADAR rather than a user-customized version. These evaluation choices make interpretation of the results clearer, but potentially reduced RADAR's benefits as individual differences in information preferences and drift within an individual's preferences over time are not captured by this procedure. The features that describe the game history for each time slice are listed in Table 1.

To provide a stringent test of the adaptive display system, subjects' ability to manually select displays (i.e., override RADAR) was disabled. Removing this ability forces subjects to completely rely on RADAR for information updates and simulates conditions in which operators do not have the option of scrolling through menus while on task. Performance with RADAR functioning was compared to a closely matched control condition. In the control condition, displays were shown for the same durations as the experimental condition (i.e., the base rates and mean durations of the eight displays were matched), but transitions between displays were determined at random rather than selected by RADAR. Thus, any benefit of RADAR over the control condition is attributable to RADAR's selecting the situationally appropriate displays for the HUD, as opposed to RADAR's merely learning which displays are most valuable in general. Each player completed fifty test games.

Results

The primary dependent measure was the mean number of enemy tanks destroyed per game. As predicted, subjects killed significantly more (4.54 vs. 3.29) enemy tanks in the experimental than in the control condition, $t(4) = 10.60, p < .001$. All five subjects showed an advantage with RADAR. These results indicate RADAR's effectiveness. Performance with RADAR could not have surpassed the control condition unless RADAR was successful in providing the situationally appropriate displays and doing so boosted subject performance. However, Study 1 does not establish that automation is beneficial over purely manual operation. This issue is addressed directly in Study 3.

Study 2: Maintaining Situation Awareness

Methods

The question of primary interest in Study 2 was whether RADAR helps subjects maintain situation awareness. If so, subjects using RADAR should be more aware of the state of their tank, and thus be more likely to replenish fuel and ammunition when necessary. Therefore we predicted that RADAR should reduce the rates of dying from causes that are somewhat avoidable, specifically, running out of fuel or ammunition.

Study 2 used the same methods and experimental conditions as Study 1. The weights for the RADAR model derived in Study 1 were also retained. Nine inexperienced players who had not participated in Study 1 served as subjects.

Results

The distribution of player deaths by condition is shown in Figure 4. As predicted, a greater proportion of games ended with fuel and ammunition depleted in the control condition than when RADAR was operating, $\chi^2(2, N = 713) = 12.58$, $p < .01$. These results suggest that players were less aware of the state of their vehicle in the control condition.

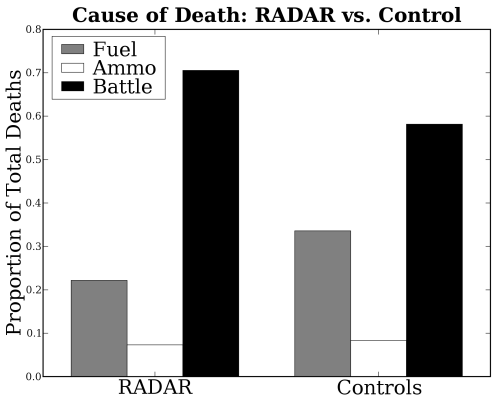

Figure 4. Study 2 demonstrates that players are more likely to lose situation awareness and die from somewhat avoidable causes, such as running out fuel, when RADAR is not operating.

Study 3: Individual Differences and Comparison to Manual Selection

Studies 1 and 2 establish RADAR's benefits over closely matched controls in terms of providing situationally relevant display. RADAR boosted overall performance and increased situation awareness relative to controls. However, Studies 1 and 2 do not establish whether RADAR is more effective than no automation of display choice. Indeed, automation could lower overall levels of performance relative to fully manual display selection. Study 3 assesses this possibility by using a manual control condition.

The second focus of Study 3 was the importance of individual differences in display preference. A separate RADAR model (i.e., set of weight parameters) was estimated for each player, and performance was compared between players using their own models and using other individuals' models. Additionally, to evaluate the robustness of our approach, a minimal feature set, consisting only of features 1-12 in Table 1, was used.

Methods

Five undergraduate student volunteers in the laboratory served as the research subjects. Each student had over ten hours experience playing the tank game without RADAR operational prior to test data collection. A user-specific RADAR model was fit to each subject using four hours of manual play data.

Each subject completed test games in three conditions (in a within-subjects design): Manual, Individual, and Other. In the Manual condition, no RADAR model was operable and subjects manually selected all displays (as in the training phase). In the Individual condition, each subject used the RADAR model derived from his or her own training data. In the Other condition, each player used the other players' models. In both experimental conditions, subjects were allowed to manually override RADAR's display choices.

To evaluate RADAR's promise in contexts where minimal input from subject matter experts is available, a minimal feature set was used to predict display preferences in all RADAR models. This minimal set consisted of the "Display Shown" and "Tank Condition" features shown in Table 1. Each player completed 12 test games in each of the three conditions. Game order was randomly determined for each subject with games from the various conditions interleaved.

Results

Mean kills per condition for the Manual, Individual, and Other conditions were 5.1, 6.2, 5.9, respectively. Subjects killed significantly more tanks in the Individual and Other conditions than in the Manual condition, $t(4) = 3.02, p < .05$ and $t(4) = 2.84, p < .05$, respectively. The advantage of these RADAR conditions over the Manual condition held for all five subjects.

These results indicate that individual RADAR models are more effective than purely manual operation. The strong performance in the Other Individual condition was attributable to relatively novice subjects benefiting from using the display models of more experienced subjects. This serendipitous result suggests that RADAR may prove effective as a training system in which novice subjects train under an expert's RADAR model.

Study 4: Novice vs. Expert Humans and Models

Study 4 assessed whether subjects' mental models shift as a function of experience on task. Data collected under manual play were assessed using RADAR to determine the features that novices and experts attend to when making display updates. Additionally, this data set was used to assess whether display changes are aligned with event boundaries and whether these boundaries become sharper as subjects become more expert.

Methods

Five paid undergraduate students provided twelve hours of data under manual play. Prior to the experiment, these subjects had no experience with the tank game.

Results

A Novice RADAR model was fit to each subject's first four hours of game play and an Expert RADAR model was derived from the final four hours. Rather than use all the features listed in Table 1, we determined the features that subjects actually entertained. This was done by evaluating subsets of all possible features using cross validation [19]. In cross validation, including features that are not "psychologically" real will decrease performance on the data held out to test for generalization.

Experts' second RADAR stage involved more features (4.1 vs. 2.8) in accord with findings from the expertise literature indicating that experts have richer feature vocabularies [37]. Interestingly, this difference (and every other comparison of novices and experts) strengthens (4.8 vs. 2.4) when one subject who did not improve (and therefore never truly became expert) is removed from the analysis. Novice and Expert models differed in the features typically included. Larger scale studies are necessary to assess the basis for these differences and to answer questions like whether expert models are organized along deeper principles [5].

RADAR contains two stages, the first of which we claim is akin to event segmentation. As previously reviewed, cognitive load and change in the environment are greatest at event boundaries (the very times one would want RADAR to update the display). If display changes in the tank game occur at event boundaries, then there should be relative stability in the environment following a display change. Furthermore, because event structure is learned, experts should exhibit sharper event boundaries than novices.

To evaluate these hypotheses, we measured feature change (see Table 1 for the features) across consecutive time slices (250 ms each) in the game ten seconds before and after each display change. As predicted, there was more feature change (.28 vs. .21) prior to a display update than after, $F(1, 4) = 72.14, p < .01$. Furthermore, this difference was larger (.08 vs. .06) for experts than for novices, $F(1, 4) = 11.71, p < .03$.

To assist in visualizing these data, Figure 5 illustrates an example expert subject that was run extensively on the task. Notice that feature change activity drops around the time of display update. Interestingly, the display update occurs after feature change activity begins to decrease. This lag might reflect the time required for subjects to complete decision and response processes in the course of making a manual display selection. RADAR's prescience (time shifting displays to be just-in-time) attempts to address this lag.

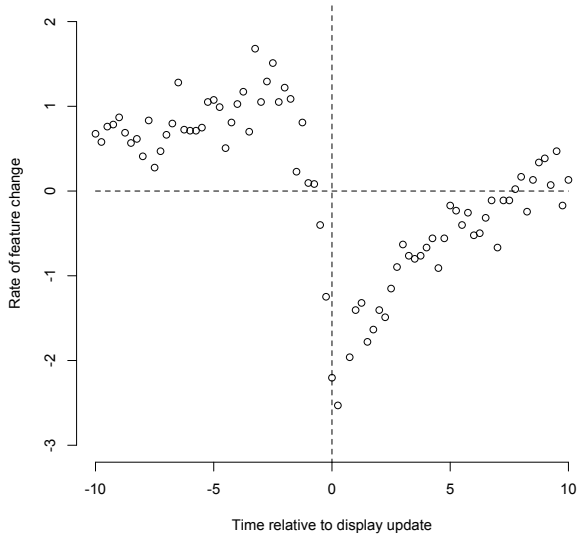

Figure 5. Feature change (a proxy for change in the environment) is plotted in z-scores. Time on the horizontal axis (in seconds) is relative to display updates (negative is prior to update, positive is post update). The plot indicates that feature change is greatest prior to a display change. These results support the notion that display updates are akin to event boundaries.

Study 5: External Markers of Proficiency and Expertise

One critical challenge in training is evaluating novices' knowledge structures. A variety of laborious and somewhat problematic techniques, such as think-aloud protocols and structural assessment, are often used to measure a person's knowledge [20]. These assessments are important because they can predict trainee comprehension, differentiate between novices and experts, and forecast future achievement [7]. Critically, as novices progress, their knowledge structures converge with those of experts [1]. One interesting question is whether RADAR can serve these functions without making recourse to subject matter experts or special evaluation procedures.

Experimental Method

Data were collected from forty-six novice subjects. In the first hour of game play, displays were shown randomly to familiarize subjects with the game and the displays. In the second and final hour of game play, subjects played under manual control. We fit each of the forty-six subjects' second hour of play with each of the models (five Novice, five Expert) from Study 4. For each subject, we determined which model predicted the subject's display selections best.

Experimental Results

Subjects' performance in the second hour under manual play could be predicted by which of the ten models best fit. The correlation between subject performance and that associated with the model that best fit was .26 ($p < .05$). Subjects that were best fit by one of the five expert models (18 of the 46 subjects) outperformed (2.5 vs. 1.8 kills per game) subjects best fit by one of the five novice models, $t(44) = 2.13, p < .05$. These results are very encouraging at this early stage of the project, especially given the sparsity of our data. RADAR offers the possibility of continuous evaluation and assessment of trainees without intervention. Somewhat surprisingly and consistent with our viewpoint, the results indicate that significant aspects of expertise are externalized in terms of information preferences revealed through display requests.

GENERAL DISCUSSION

Advances in information technology make large quantities of information available to human decision makers. In this deluge of information, finding and selecting the relevant piece of information imposes a burden on the user. This burden is particularly onerous in dynamic environments in which decisions must be made rapidly. RADAR is a domain-general system that learns to approximate the information search process of an individual user. By offloading this search process to RADAR, the user can concentrate on the primary task. Experimental results in a tank video game environment in which the player must maintain situation awareness demonstrate RADAR's promise. Players performed better with RADAR.

RADAR provides a powerful tool to quantitatively assess individual performance and the transition from novice to expert-level performance. Consistent with findings from the expertise literature, RADAR models derived from expert subjects involved more features than models derived from novice subjects. RADAR was also successful in evaluating novices' knowledge structures.

A variety of laborious and somewhat problematic techniques, such as think-aloud protocols and structural assessment, are often used to measure a person's knowledge [20]. These assessments are important because they can predict trainee comprehension, differentiate between novices and experts, and forecast future achievement [7]. Critically, as novices progress, their knowledge structures converge with those of experts [1]. We found that novices best fit by expert RADAR models performed best.

Finally, our results indicate that display updates in the tank game are akin to event boundaries. This finding suggests that are task environment is sufficiently rich to contain meaningful event structure. The fact that RADAR does a good job at identifying these boundaries is likely one of the reasons why its display updates boost user performance.

In the face of these successes, it is important to keep the limitations of the current system in mind. RADAR is not a cure all and is not intended to satisfy every user need. Alt-

hough RADAR can be viewed as an information agent, as it proactively retrieves context relevant information, RADAR does not perform other tasks commonly associated with information agents, such as information synthesis [18]. RADAR does not interpret information for the user, nor suggest how the user should act on information. Critically, RADAR's display predictions are not prescriptive. Rather, its choices reflect the user. For users, RADAR's function is to predict the displays that people desire. The displays people desire may in fact not be the best displays for the situation. RADAR might show limited benefits for users who chronically request inappropriate displays. Indeed, we find that RADAR models derived from novice users are in many ways inferior to expert-user models. This latter point highlights another function RADAR serves, namely serving as an assessment tool for scientists and practitioners.

Systems that automate tasks for humans often result in unexpected negative consequences [25]. One problem with automation is that automatic changes are often missed by human operators in systems with low observability [30]. We believe RADAR's design makes it less likely than most systems to suffer from these problems. Users can maintain basic control by overriding RADAR's display choices (see Figure 1). Mode errors are unlikely because all automatic updates involve a change of display, which the user should notice. Trust in the system should be high as RADAR learns to approximate a user's desired display preferences, rather than prescribe what the user should view. Finally, RADAR can be incrementally deployed with increasing rates of automation over time in order to maximize the benefits of automation and minimize the detriments [39].

One idea along these lines is make display update's in RADAR opt-in rather than opt-out. For instance, users could hit a key when they wish to advance to the display that RADAR's recommends. This opt-in operation walks the line between two basic modes of context-aware information retrieval: interactive (as in a web search engine where explicit queries are made) and proactive (as in the RADAR simulations reported here) [15]. Opt-in operation also eases the computational challenges in training RADAR models online so that RADAR and human operators can co-evolve. Our studies demonstrate that human users' behavior changes with RADAR operating, so it is critical for RADAR and human users to train simultaneously in order to converge to an optimal solution.

ACKNOWLEDGMENTS

This work was supported by AFOSR grant FA9550-07-1-0178 and NSF CAREER grant #0349101 to B.C.L. Special thanks to Kelvin Oie and Jeff Zacks for helpful comments.

REFERENCES

1. ACTON, W., JOHNSON, P., AND GOLDSMITH, T. Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology 86* (1994), 303–311.

2. ALBRECHT, D. W., ZUKERMAN, I., AND NICHOLSON, A. E. Bayesian models for keyhole plan recognition in an adventure game. *User Modeling and User-Adapted Interaction 8*, 1-2 (1998), 5–47.

3. BYLUND, M., AND ESPINOZA, F. Using quake iii arena to simulate sensors and actuators when evaluating and testing mobile services. *In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems* (2001), 241–242.

4. CHEVERST, K., DAVIES, N., MITCHELL, K., FRIDAY, A., AND EFSTRATIOU, C. Developing a context-aware electronic tourist guide: some issues and experiences. *In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems* (2000), 17–24.

5. CHI, M., FELTOVICH, P., AND GLASER, R. Categorization and representation of physics problems by experts and novices. *Cognitive Science 5* (1981), 121–152.

6. CLEEREMANS, A. *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT Press, Cambridge, MA, 1993.

7. DAY, E., ARTHUR, W., AND GETTMAN, D. Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology 86* (2001), 1022–1033.

8. FOGARTY, J., HUDSON, S. E., AND LAI, J. Predictability and accuracy in adaptive user interfaces. *In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems* (2004), 207–214.

9. GAJOS, K. Z., EVERITT, K., TAN, D. S., CZERWINSKI, M., AND WELD, D. S. Predictability and accuracy in adaptive user interfaces. *In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems* (2008), 1271–1274.

10. GOODMAN, B. A., AND LITMAN, D. J. On the interaction between plan recognition and intelligent interfaces. *UMUAI 2* (1992), 83–115.

11. GRAY, W. Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and laboratory tasks in basic and applied cognitive research. *Cognitive Science Quarterly 2* (2002), 205–227.

12. GURECKIS, T., AND LOVE, B. C. Transformational vs. statistical complexity in sequential learning. *Cognitive Science* (in press).

13. HORVITZ, E., AND BARRY, M. Display of information for time-critical decision making. In *Proc. of Conf. on Uncertainty in AI* (1995), pp. 296–305.

14. HORVITZ, E., KOCH, P., AND APACIBLE, J. Busybody: Creating and fielding personalized models of the cost of interruption. *In Proceedings of the ACM Conference on Computer Supported Cooperative Work* (2004), 507–510.

15. JONES, G. J. F., AND BROWN, P. J. *Context-Aware Retrieval for Ubiquitous Computing Environments*. Springer, 2004.

16. KAPOOR, A., AND HORVITZ, E. Experience sampling for building predictive user models: a comparative study. *In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems* (2008), 657–666.

17. KLEIN, G. A. Recognition-primed decisions. In *Advances in man-machine research*, W. Rouse, Ed. JAI Press, Greenwich, 1989, pp. 47–92.

18. KLUSCH, M. Information agent technology for the internet: a survey. *Data Knowl. Eng. 36*, 3 (2001), 337–372.

19. KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial lntelligence* (1995), 1137–1143.

20. KRAIGER, K., AND FORD, J.K. ANDSALAS, E. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology 78* (1993), 311–328.

21. KURBY, C. A., AND ZACKS, J. M. Segmentation in the perception and memory of events. *Trends in Cognitive Sciences 12* (2008), 72–79.

22. MÄNTYJÄRVI, J., AND SEPPÄNEN, T. Adapting applications in mobile terminals using fuzzy context information. In *Mobile HCI* (2002), pp. 95–107.

23. MCRAE, K., SPIVEY-KNOWLTON, M. J., AND TANENHAUS, M. K. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language 38* (1998), 283–312.

24. MCTEAR, M. F. User modeling for adaptive computer systems: a survey of recent developments. *Artificial Intelligence Review 7* (1993), 157–184.

25. MILLER, C., FUNK, H., GOLDMAN, R., MEISNER, J., AND WU, P. Implications of adaptive vs. adaptable UIs on decision making: Why automated adaptiveness is not always the right answer. In *Proc. of the 1st Inter. Conf. on Augmented Cognition* (2005).

26. NEWTON, D. Foundations of attribution: The perception of ongoing behavior. In *New Directions in Attribution Research*, J. Harvey, Ed. Erlbaum, Mahwah, NJ, 1976.

27. RADVANSKY, G. A., AND COPELAND, D. E. Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition 34* (2006), 1150–1156.

28. RINCK, M., AND WEBER, U. Who when where: An experimental test of the event-indexing model. *Memory & Cognition 31* (2003), 1284–1292.

29. SAFFRAN, J. R. Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science 12* (2003), 110–114.

30. SARTER, N., MUMAW, R., AND WICKENS, C. Pilots monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors 49* (2007), 347–357.

31. SCHNEIDER-HUFSCHMIDT, M., KÜHME, T., AND MALINOWSKI, U. *Adaptive User Interfaces: Principles and Practice*. North-Holland, 1993.

32. SHARP, R. Electrophysiological correlates of event segmentation: how does the human mind process ongoing activity. *Proceedings of the Annual Meeting of the Cognitive Neuroscience Society 235* (2007).

33. SPEER, N., SWALLOW, K., AND ZACKS, K. Activation of human motion processing areas during event perception. *Cognitive, Affective & Behavioral Neuroscience 3* (2003), 335–345.

34. ST. JOHN, M., SMALLMAN, H. S., AND MANES, D. I. Assisted focus: Heuristic automation for guiding users' attention toward critical information. In *Proc. of the 1st Inter. Conf. on Augmented Cognition* (2005).

35. SWALLOW, K. Perceptual events may be the episodes in episodic memory. *Abstr. Psychon. 12* (2007), 25.

36. SWALLOW, K., AND ZACKS, J. Hierarchical grouping of events revealed by eye movements. *Abstr. Psychon. 9* (2004), 81.

37. TANAKA, J. W., AND TAYLOR, M. Object categories and expertise: Is the basic level in the eye of the beholder. *Cognitive Psychology 23*, 2 (1991), 457–482.

38. TEEVAN, J., DUMAIS, S. T., AND HORVITZ, E. Personalizing search via automated analysis of interests and activities. *Proceedings of SIGIR* (2005), 449–456.

39. WICKENS, C. D., AND DIXON, S. R. The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomic Science 8* (2007), 201–212.

40. YI, W., AND BALLARD, D. Behavior recognition in human object interactions with a task model. In *AVSS* (2006), pp. 64–64.

41. ZACKS, J., SPEER, N., VETTEL, J., AND JACOBY, L. Event understanding and memory in healthy aging and dementia of the alzheimer type. *Psychology & Aging 21* (2008), 466–482.

42. ZWAAN, R., RADVANSKY, G., HILLIARD, A., AND CURIEL, J. Constructing multidimensional situation models during reading. *Scientific Studies of Reading 2* (1998), 199–220.