"An Unscented Hound forWorking Memory" and the Cognitive Adaptation of User Interfaces

Bruno Massoni Sguerra, Pierre Jouvelot

"An Unscented Hound for Working Memory" and the Cognitive Adaptation of User Interfaces

Bruno Massoni Sguerra
MINES ParisTech, PSL University, France
bruno.sguerra@mines-paristech.fr

Pierre Jouvelot
MINES ParisTech, PSL University, France
pierre.jouvelot@mines-paristech.fr

ABSTRACT

We introduce An Unscented Hound for Working Memory (AUHWM), a new framework for the real-time tracking of human Working Memory (WM) that can be used to adapt computer interfaces to users' available cognitive resources. WM is the part of human cognition responsible for the short term storing and handling of information; it can, in stressful situations, under information overload, or when suffering from dementia-like diseases, become severely limited, possibly leading to poor decision making. Our preliminary results suggest that AUHWM can provide a precise and timely assessment of WM capacity, so that the cognitive load a specific task imposes on users can be adapted, e.g., at the User Interface (UI) level.

AUHWM is based on a low-level stochastic discrete model of human WM dynamics, implemented as a Gradient-Boosting-derived deterministic algorithm that simulates users' oblivion. AUHWM also performs Unscented Kalman filtering to track users' WM-specific parameters in real time, thus providing a dynamic assessment of their cognitive resources. Our approach has been tested and validated using data collected from Match^{2}s, a visual memory game played by 18 users in another study. Going beyond real-time WM tracking, AUHWM is intended to also be used for WM prediction, paving the way to the adaptation of tasks and their UIs in real time as a function of users' cognitive abilities; we detail an example of such an adapted system, and provide experimental evidence this approach has the potential to lead to enhanced WM-adapted UIs in the future.

CCS CONCEPTS

• **Human-centered computing** → **User models**; • **Computing methodologies** → **Modeling and simulation**; • **Applied computing** → *Consumer health*;

KEYWORDS

UI adaptation, Working Memory, Gradient Boosting, Cognitive adaptation, Unscented Kalman filter

ACM Reference format:
Bruno Massoni Sguerra and Pierre Jouvelot. 2019. "An Unscented Hound for Working Memory" and the Cognitive Adaptation of User Interfaces. In *Proceedings of UMAP'19, June 09-12, 2019, Larnaca, Cyprus, ACM,* 8 pages.

UMAP'19, ACM, June 09-12, 2019, Larnaca, Cyprus

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/3099023.3099052

https://doi.org/10.1145/3099023.3099052

1 INTRODUCTION

Working memory (WM) is the part of human cognition responsible for the storing and processing of short-term information [2]. It is essential for daily activities, ranging from having a conversation to following instructions to problem solving. WM is known to be extremely restricted, being limited by the amount of information that can be stored [8] as well as the period of time during which such information is available [17]. Its capacity is responsible for the biggest differences in cognitive abilities [10] and is usually associated with fluid intelligence [7]. Deficit in WM capacity can be linked to life-altering problems in learning, language understanding and many others activities. One of the main drivers of WM capacity weakening is dementia, usually linked to the onset of Alzheimer-like diseases in the elderly.

In this article, we present An Unscented Hound for Working Memory (AUHWM, pronounced "om"), a new cognitive framework capable of estimating in real time a person's WM capacity through interaction with a computer system. WM capacity is often described by the fixed number of information items, or "chunks", that can be stored at the same time in memory [4]. In the case of complex information storage, this limit fluctuates around four [4]; while performing simpler tasks, the capacity can increase up to seven [8].

AUHWM is able to dynamically track a user's WM cognitive capabilities over both short- and long-term time intervals. One key component of AUHWM is a simple, general and yet well-validated low-level cognitive model of human Working Memory (see Section 3) that abstracts, via a single integer parameter, a user's memory capacity. A second key feature behind AUHWM is Unscented Kalman filtering (see Section 5.1), a control theory tool used here to track this parameter as the user interacts with an AUHWM-equipped system. Building upon such well-understood mechanisms allows cases that could not be previously handled to be now tackled. For instance, AUHWM-based applications can be used with different populations without going through the burdensome process of collecting personalized training data that plagues typical Machine Learning-based systems. Also, by having a clear understanding of what AUHWM modelling parameters stand for, high-level explanations of the system choices can be provided.

We believe that the AUHWM framework for cognitive tracking can be of extreme importance when developing User Interfaces (UI) that are sensitive to the user's reasoning and memorizing abilities. Knowing how taxing in cognitive capacity a task is would allow its UI in stressful situations to be simplified, providing the user with only the necessary information to the task at hand. One can

think of aircraft interfaces in crisis situations as a clear application domain of our approach.

AUHWM could also represent a significant contribution in the area of assistive technologies as it has the potential to be of great benefit to individuals suffering from memory deficits. By adjusting the UI to the user's cognitive capacities, it could render computer interfaces more accessible to the elderly population suffering from dementia-linked diseases. AUHWM can provide as well data specific to the user's evolving cognitive capacities. Beyond its clear relevance in the design of simpler user interfaces for computer-assisted daily-life activities, such information can be used by caregivers as signals suggesting a possibly setting-in of neurodegenerative diseases. It can also be used to track the gradual temporal decline of cognitive abilities as the disease progresses.

To summarize, our paper includes four major contributions:

- AUHWM, a new framework for the dynamic, real-time modelling, tracking and ultimately prediction of human WM performance, which uses a quanta-based stochastic model of memory and Unscented Kalman filtering;
- a Gradient Boosting-based, deterministic, approximate implementation of this stochastic memory model, for performance-efficiency purposes;
- an experimental evaluation of AUHWM ability to track WM capacity, using pre-existing data extracted from the visual memory game Match^{2}s, played by 18 players;
- a new AUHWM-based framework for automatic UI task adaptation, using previously tracked WM parameters as estimates for future performance, and its evaluation using the same data set.

The structure of the paper is as follows. We introduce in Section 2 the previous works related to our research. Section 3 describes the Markov Decision Process-based model used to probabilistically abstract the WM maintenance dynamics. We show in Section 4 how this stochastic system can be well approximated by a much more efficient Gradient Boosting-based algorithm. AUHWM, introduced in Section 5, builds upon this deterministic version to track users' WM parameters via Unscented Kalman filtering. We provide experimental evidence of the validity of AUHWM in Section 6, using actual user interactions with Match^{2}s. Section 7 describes how AUHWM can be used to predict future WM performances, which thus paves the way to the temporal adaptation of UIs. We discuss possible future work in Section 8, before concluding in Section 9.

2 RELATED WORK

Cognitive Load Theory [6] posits that a person has a finite amount of cognitive resources, and that different tasks apply different cognitive loads, resulting in more or less available resources for the consecutive tasks and therefore different performance. Cognitive Load Theory provides the basis for a number of different studies related to the measurements and compensation of working memory limitations. Most of these works deal with problem-solving activities such as learning or decision making.

Long Short-Term Memory networks are used in [12] to learn different patterns of sequential behavioural data in order to classify dynamically user's behaviour into either (1) under cognitive load or (2) not. The approach is based on data collected from users playing a memory game as well as data generated using a theoretical memory model whose parameters were set so that the generated data closely resemble the collected one.

In [5], authors learned a Hidden Markov Model (HMM) from data where the hidden states correspond to different levels of cognitive load. Here the user's reaction time, accuracy and error signal are used to infer the hidden states. Once the user's cognitive load is known, the proposed model is used to adapt the collaboration between humans and software agents.

Closer to our goal of adapting systems to user's cognition, [19] describes the design of a data-driven Socially-Assistive Robot system for personalized robot-assisted training. In this work, interactive reinforcement learning is used to adapt the robot's behavior to users' performance. Data corresponding to users' electroencephalography (EEG) signals, performance and engagement is collected and analized in order to find clusters. These clusters of users are then used to create simulation models and learn user-specific policies through reinforcement learning.

Most of these systems are data-based. The neural network learned in [12] might work poorly when presented to a different population, as would the HMM in [5] and the user models of [12]. Deep learning and machine-learning tools work effectively when the data used for learning is comprehensive enough to create a representation of the use case, as they try to fit a function capable of performing meaningful association. This can be problematic when the system is used with different populations that are not represented in the data, say people with cognitive deficits, for instance, resulting in a lack of flexibility and adaptability. Our approach uses an online filtering technique able to adapt smoothly to any user's specific capabilities, even in the presence of scarce data.

Moreover, trustworthiness is widely recognized as crucial for the acceptance of "intelligent" systems in diverse domains [16]. Being able to explain a system's choice of action is crucial for building trust, in particular when dealing with assistive technologies, where the user has to trust the system's decision for it to be effective. This is, for now, critically lacking when dealing with black-box classifiers such as neural networks. Our approach, which uses clearly separable, well-understood and interpretable components, helps here.

3 WORKING MEMORY MODEL

The WM model at the core of AUHWM was proposed by J.W. Suchow in [18][17]. The evolution dynamics of the information stored in the WM is considered as a Moran process [9], a stochastic formalism often used to describe the dynamics of finite populations in biology. At each instant where the state of the population may change, an individual, chosen at random, dies and another is chosen for reproduction, ensuring a constant yet varying population.

Suchow models the evolutionary dynamics of information in WM as the evolution of a finite population of "memory quanta". A number of quanta is allotted to a each information item in the WM: the more quanta assigned to an information there is, the better encoded it is and therefore the easier it is to be retrieved. Although the authors in [17] are non-committal about what these quanta represent (they could take a number of forms, such as clusters of neurones in the prefrontal cortex, cycles in time-based refreshing

processes or other elements), they make it clear that this is a limited commodity whose availability affects performance. Logically, the total number of quanta is positively correlated to the cognitive capacity of an individual. The more available quanta, the better the quality and stability of memory.

Following the rules of Moran processes, at each time step, a random quantum assigned to an information "dies" while the WM maintenance mechanism selects another quantum to "reproduce". The quantum chosen for reproduction can be related to the same information as the dead one, thus ensuring the persistence in memory of this information. If, however, the quantum selected for reproduction isn't one allotted to the degraded information, but to a different one, the latter is then reinforced in detriment of the former. This dynamics results in a competition for quanta, i.e., for fixation in memory. This model also uses a stability threshold L: any information associated to less than L quanta is considered forgotten and cannot be restored via reproduction.

3.1 MDP Formulation

Suchow's WM dynamics is modeled as a Markov decision process (MDP). A MDP is used to model decision making in partially stochastic environments, where an agent (or decision maker) selects actions to optimize a cost (or reward function) [21]. Formally, a MDP is defined by a state space S, a set of actions A, a probabilistic transition function $\tau : S \times A \rightarrow P(S)$ to move to the next state s' given the present state s and a selected action a and finally, a reward (or cost) function $\rho : S \times A \rightarrow R$ that yields the immediate consequence the agent taking an action in a given state gets – in some extended MDP models, the reward also depends on the next state the agent finds itself in. The goal when using a MDP is to find the optimal policy $\Pi^* : S \rightarrow A$ that maps a given state s to the optimal action a the agent should take in order to maximize (or minimize) its accumulated reward (or cost).

In Suchow's model, the WM maintenance mechanism acts as the MDP agent. The state space corresponds to all the possible allocation of Q quanta to k information bins: $s = [n_1, \ldots, n_k] \in S$, where $\sum_{i=1}^{k} n_i = Q$. Each action a_i from A represents the selection of a quantum from a specific memory bin b_i for reproduction. Following Moran's principle, at each system iteration, one randomly selected quantum from a bin, say b_i, decays (i.e., dies), with probability $P = n_i/Q$, while the maintenance mechanism chooses a specific action, say a_j, to have one of the quanta of bin b_j reproduced; so, if the system is at state $s = [n_1, n_2, \ldots, n_k]$ and the agent selects action a_1, the probability of the agent landing in state $s' = [n_1 + 1, n_2 - 1, \ldots, n_k]$, given by $\tau(s, a_1)$, is $P(s'|s, a_1) = n_2/Q$, which is the probability that one quantum from the second bin was selected to decay.

Regarding the reward function ρ, the behaviour of the maintenance mechanism handling the information stored in the WM might vary along the user's goal; information items can be remembered or forgotten intentionally. Thus ρ is clearly task-dependent. Here, we posit that the reward corresponding to the specific use-case task of this study, i.e., playing the Match^{2}s visual memory game (see Figure 2 and Section 6 for more details) favours storing as many information items as possible for the longest period of time; ρ thus positively correlates with the number of bins with L quanta or more.

3.2 Optimal Policy

The authors in [18] suggest that the optimal policy of the previously defined MDP can be approximated by a simple strategy known as Luce's choice axiom. This axiom states that when faced with a choice, the decision maker will mostly base his/her decision on the perceived values of the various options at the time of choice, in a "greedy" fashion. Therefore the probability $P(a)$ of selecting action a from a set of alternatives A is given by

$$P(a) = \frac{v(a)^\sigma}{\sum_{x \in A} v(x)^\sigma},$$

where $v(x)$ stands for the strength of the signal generated by action x, and σ is the sensibility of the decision maker[1]. By varying the value of σ for a fixed definition of v, Suchow shows that one obtains different macroscopic behaviors for the WM maintenance mechanism, adapted to different tasks, and draws attention to five specific values of sensibility: $0, 1, -1, +\infty$ and $-\infty$.

Choosing $\sigma = 0$ leads to an unconditional policy, i.e., action choice is independent of the current state and insensitive to the perceived signals. If $\sigma = 1$, the policy will give preference to actions that have the highest perceived value, while the opposite occurs when $\sigma = -1$. Finally, when $\sigma = +\infty$, the maintenance mechanism will always choose the action that has the strongest perceived signal, while when $\sigma = -\infty$, the weakest one will be selected.

4 OBLIVION SIMULATION

Keeping track of users' WM capacity to model information recall and oblivion relies on the simulation of the MDP defined above; this requires the setting of six parameters, given in Table 1, and the definition of the strength function v. One also needs to specify an initial state $s_0 = [n_1, \ldots, n_k]$ representing the default distribution of quanta between information items.

Table 1: MDP simulation parameters for Suchow's WM model.

Q	Number of quanta in WM
k	Number of information items in WM
L	Stability threshold [number of quanta]
δ_t	Time step between actions [ms]
T	Total simulation time [ms]
σ	Sensibility of the decision maker

Following [17], we set L and δ_t[2] to 7 and 10 respectively. We present below two simulation strategies: a straightforward stochastic implementation and a deterministic, approximated-yet-efficient version.

[1] Care must be taken to avoid divisions by zero; we don't address these details here.
[2] If need be, these two parameters could be set to different values in the user interfaces that rely on AUHWM for personalization, possibly yielding a better model of the user's WM.

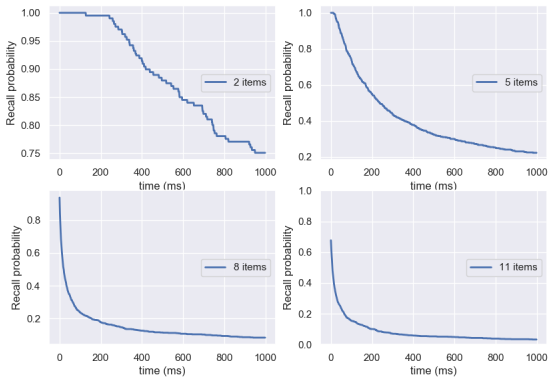

Figure 1: Recall probability $r(t)$ for different numbers, k, of items ($Q = 60$, $L = 7$, $\delta_t = 10$ ms, $T = t$)

4.1 Stochastic Simulation

As said before, WM management is task-dependent. The setting of the initial state s_0 and the definition of the signal generated by the possible actions v and the sensibility parameter σ, which characterise the Luce choice axiom underlying the MDP policy used in AUHWM, depend thus on the task. In this paper, we are interested in simulating the WM of a user performing the specific use-case task Match^{2}s. In Match^{2}s, players score higher if they are able to retain the maximum number of information items for the longest period of time possible. One can then assume that, on average, players will try to remember as much information as they possibly can, without giving particular preference to a particular stimulus. Accordingly, s_0 is set by distributing the Q quanta in the k bins homogeneously; if Q is too small to fill each bin with at least L quanta, the maximum number of bins are filled with L quanta, and the remaining ones are distributed randomly across bins. Also, we define $v(a_i) = n_i$, i.e., the strength of information fixation in bin b_i, while setting $\sigma = -1$. The probability of choosing action a_i that reinforces bin b_i will thus increase proportionally to $1/n_i$, i.e., when the number of quanta in b_i is low. This ensures that the maintenance mechanism will focus on the least stable information in order to try to keep it in memory as long as possible, which is the natural way to obtain a good performance in Match^{2}s. Using AUHWM with other types of tasks will require finding the best setting accordingly.

Once the initial state and the simulation parameters are set, one can perform various stochastic simulations of memory degradation, using the optimal policy specified above. They yield recall curves $r(t)$, with time t varying from 0 to T, by increments of δ_t. At each time t, the number of bins with more than L quanta divided by k represents the recall probability $r(t)$ of a given information item in the WM at time t. Given the stochastic nature of the model, a large number of simulations is necessary to average the recall curve. Figure 1 presents the average recall curves of 100 simulations for different values of k.

4.2 Deterministic Simulation

Unfortunately, running stochastic simulations is very time consuming; this is not acceptable for our goal of tracking, in real time, the user's cognitive capacity. Thus, we propose to implement an approximation of our adaptation of Suchow's model using a gradient-boosting (GB) approach for regression. To do so, we first sampled from uniform distributions over the key simulation parameter limits: $Q \sim \mathcal{U}(0, 120)$, $k \sim \mathcal{U}(1, 8)$ and $T \sim \mathcal{U}(0, 2500)$. The limits for k and T come from the set of possible configurations for the Match^{2}s parameters, while the minimum and maximum values for Q where identified as pertinent in a previous study where validation data was collected (see [15] for more details). The sampled parameters were then used to generate simulation data, from which we trained our GB for regression using the GradientBoostingRegressor class from `sklearn` [11]. The gradient-boosting modelling finds the relationship between Q, k and T, thus providing an approximate recall probability $f(Q, k, T) = r_{Q,k}(T)$. With this approach, we are able to retrieve user- and task-dependent recall performance with good accuracy (0.93 ± 0.02 on average in 10-fold cross-validation) without having to go through a large number of expensive stochastic simulations.

5 AN UNSCENTED HOUND FOR WORKING MEMORY

In this section, we go into the details of how AUHWM is implemented and how it's able to track users' cognitive capacity. Subsection 5.1 reviews the concepts behind Unscented Kalman Filtering, while Subsection 5.2 describes our tracking framework.

5.1 Unscented Kalman Filter

An Unscented Kalman Filter (UKF) is a standard estimation tool mostly used in nonlinear dynamic systems or in probabilistic parameter estimation [20]. Much as the traditional Kalman Filter (KF), which however only works for systems with linear dynamics [14], a UKF also provides estimations of a system's current state by propagating the previous estimation through a dynamic system model, getting evidence from sensors, and updating the system's state belief with the new data [14].

An UKF works by applying the Unscented Transformation (UT), which is a method for calculating statistics of random variables undergoing nonlinear transformations. The UT relies upon carefully selected "sigma points", i.e., chosen sample points from an initial Gaussian random variable (GRV), that wholly capture its mean and covariance, and having them undergo a nonlinear transformation. The resulted points are used to reconstruct a new GRV. For Gaussian inputs, UT is accurate to the third order (in Taylor series expansion).

A UKF nonlinear system model relies on a transition function F and an observation function G, both assumed to be known. The sampled sigma points from the prior state (x_{t-1}) go through the transition model as $F(x_{t-1}) + u_t$, where u_t is an added process noise; the resulted points are used to approximate a predicted state GRV ($\bar{x}_t$). They then go through the observation model $G(\bar{x}_t) + v_t$, with measurement noise v_t, generating $\bar{y}_t$, a predicted observation GRV. Using the real observation value y_t from the sensor, the predicted state is updated as $x_t = \bar{x}_t + K(y_t - \bar{y}_t)$, where K is the Kalman

gain. We recommend [20] for information about the Kalman gain and a more detailed discussion about UKF.

When used for parameter tracking instead of state estimation, UKF requires some slight modeling modifications. The estimated state x_t becomes the parameter w_t to be tracked, modeled as a GRV. The modeled observation becomes $G(w_t, z_t) + v_t$, viewed as an observation of w_t, linked to an application-specific input z_t.

5.2 AUHWM Unscented Kalman Filter

We have observed in Section 3 that strong links exist between the number Q of quanta a person has available and WM performance. AUHWM is thus designed to track a single paramer $w_t = Q_t$, which corresponds to the persons' cognitive capacity at time t and which drives the evolution of the recall curve, which we will be observing.

Fluctuations of a person's available cognitive capacity are bound to happen during the day, given factors such as motivation, attention or fatigue [3]. A more constant and long-term degradation might also happen with the onset of neurodegenerative diseases. Our UKF transition function F for the parameter Q is thus set such that $Q_t = Q_{t-1} + u_t$, meaning that those fluctuations on the available quanta are driven by a process noise, u_t (but see Section 8 for possible extensions).

Using the previous notations of Section 5.1, the input z_t for the observation of the state Q_t corresponds to the application-dependent tuples (k_t, T_t). The observation function $G(w_t, z_t)$ driving the nonlinear evolution of Q_t is the recall probability $r_{Q_t,k_t}(T_t)$, computed by our GB model as $f(Q_t, k_t, T_t)$, which deterministically enforces the relation $r_{Q_t,k_t}(T_t) = f(Q_t, k_t, T_t)$; measurement noise v_t is supposed null. This UKF for the estimation of Q over time was implemented using the `pykalman` library for Python [13].

6 AUHWM EXPERIMENTAL VALIDATION

We present preliminary results obtained when experimentally assessing how well AUHWM actually tracks the cognitive capacities of human users.

6.1 Data

We tested AUHWM with data collected in [15]. The data set corresponds to 18 participants (7 females), ranging in age between 18 and 40 (26.10 ± 5.37) playing the visual memory game Match^{2}s. In a series of consecutive turns, the game consists of (1) displaying some squares (maximum 8) of different colors during 500 ms to the player, (2) hiding then the colors using yellow boxes with a "?" during a variable time t_{wait}, and (3) having the player answer a popup box asking to click on the hidden square of a determined color. Figure 2 depicts a typical turn where seven colors are presented at once (note that, for this turn, since seven colors are presented, one of the eight squares remains gray); the player is then asked to click on which box s/he thinks contained the determined color before the hiding phase, thus testing his/her memory capabilities.

Every participant played the game for 125 turns; the first 5 were used to familiarize them with the game's mechanics. For the next 120 turns, at every batch of 20 turns, the number of presented colors (that corresponds to our k) as well as the hiding time t_{wait} (our T) changed according to the player's performance (see [15] for a detailed description of the feedback mechanism used then). The

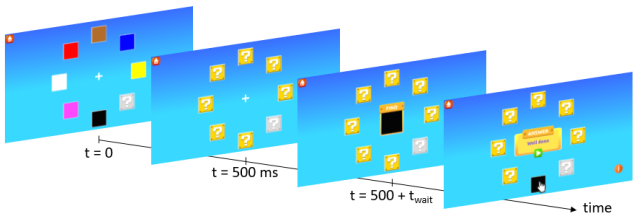

Figure 2: Example of a Match^{2}s game turn

player's actual recall probability r for each batch is then computed as $n/20$, where n is the number of successful answers for the queried colors. Overall, this resulted in a dataset of six (120/20) data points $((k, T), r)$ for each of the 18 players, which we are using to provide estimates of the WM capacity, in numbers Q of quanta, of each player.

6.2 Results

We applied the UKF of AUHWM to the data of the 18 players, starting with an initially quite loose state GRV estimate $Q_0 \sim \mathcal{N}(100, 32)$. One example of a tracked Q_t, defined by its mean and standard deviation for each "time" t (t being a batch number), for one player is shown in Figure 3. One can see that AUHWM is able to zoom in on this user's cognitive capacity, here around a somewhat steady 70 quanta at the time this game was played.

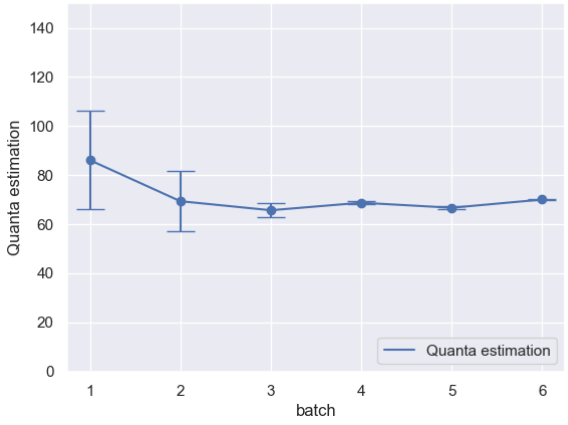

Figure 3: Example of the means and standard deviation bars of UKF-predicted states Q_t, as tracked by AUHWM, for one Match^{2}s player.

Once equipped with the estimated numbers of quanta Q_t provided by AUHWM, we can use our GB model to estimate the recall probabilities each user should have had for each batch, based on Suchow's memory model and the actual (k_t, T_t) values, remembering that the recall probability is given by $f(Q_t, k_t, T_t)$. Figure 4 depicts (dashed line in red) the evolution of the recall probability the same player displayed when presented with 6 different combinations of k_t and T_t as well as the recall probability obtained using the number

of quanta estimated by AUHWM, together with the corresponding values for k_t and T_t, as input to the GB model $f(Q_t, k_t, T_t)$.

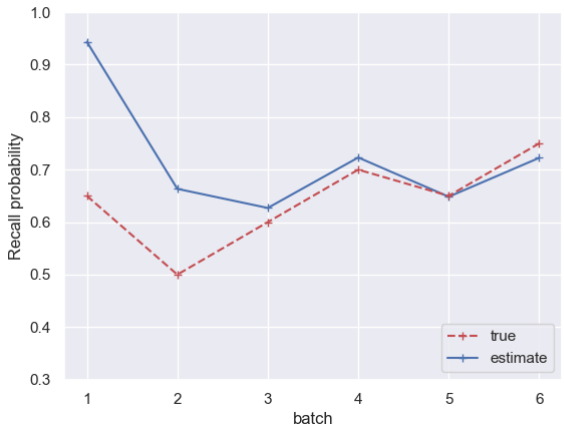

Figure 4: Actual vs. estimated recall curves generated by the GB model, using the quanta estimates in Figure 3.

One can see that after the first two batches, AUHWM is correctly assessing the number of quanta that corresponds to the player performance, that is, is tracking reliably the player's cognitive capacity. Figure 5 depicts the evolution of the Root-mean-square error (RMSE) of the estimated recall probabilities with respect to the actual 18 players' performance per batch. The last three estimations present a mean RMSE error of approximately 10%, therefore showing that after the initial batches, AUHWM is indeed finding accurate quanta numbers for each player in each batch, thus providing additional support for its validity.

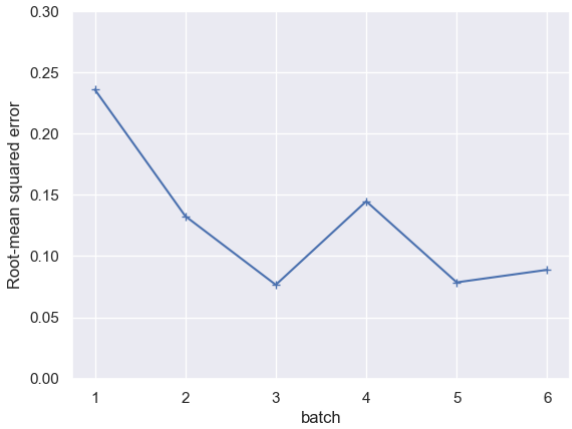

Figure 5: Evolution of the RMSE for the recall probability of all the 18 Match^2 players, per batch.

7 UI ADAPTATION

We have provided experimental evidence that suggest that AUHWM enables the real-time tracking of a person's cognitive capacity when observing his/her performance on a task. There are many possible applications of such a system. A direct, health-motivated one is the assessment of a user's cognitive decay over time. Another one could be the quantitative determination of the *load* a given task induces on someone's abilities; this could, for instance, be measured by comparing the estimated numbers of available quanta when performing a given task to the values found when performing a supplementary task before it, similarly to the experiment in [12]. However, we discuss below AUHWM intended application, namely the personnalization of UIs.

7.1 Oblivion Adaptation

As mentioned in the introduction, AUHWM's estimations are mostly intended to be used in adapting task-specific UIs to users' cognitive limitations in real time. We introduce in Figure 6 a possible AUHWM-based framework for real-time UI adaptation.

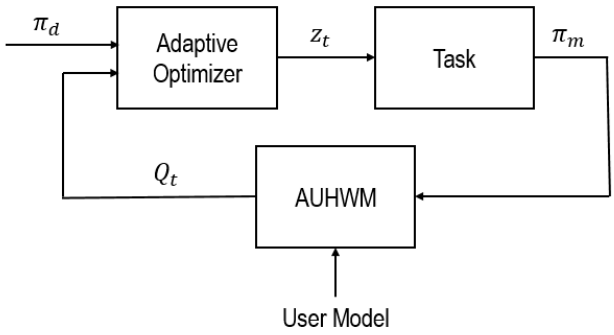

Figure 6: AUHWM-based UI adaptation of a task to users' cognitive capacity Q_t at time t.

In such a framework, a task-specific parameter π_d must be specified by the task manager; this parameter corresponds to the desired performance one wants the user to have when performing the task. For instance, in the context of Match^{2}s, this parameter corresponds to the probability of recall; if set to 1, one wishes the user to get a perfect score in Match^{2}s; if set to 0.5, the user global performance would show 50% successful answers, on average. At a given time step t, the Adaptive Optimizer (see Figure 6) is responsible for finding the task parameters z_t that will ensure that, on average[3], the user will perform with performance π_d, given the initial quanta estimation Q_{t-1}. Once again, in the context of Match^{2}s, z_t would correspond to k_t and T_t, that is, the number of information items presented as well as the duration of time during which the player has to hold this information in his/her WM. These optimized parameters are therefore the ones that ensure the constraint $r_{Q_{t-1},k_t}(T_t) = \pi_d$. Once given z_t, the task is adapted accordingly and presents its possibly updated UI to the user. The

[3]Finding a proper z_t could be done, for instance, by searching the task-parameter space for the combination of values that would result in the desired performance (or its closest approximation).

user's measured performance π_m is then used by AUHWM to estimate the next state Q_t, corresponding to the updated assessment of the user's cognitive capacity. As the user interacts with the interface over time, as depicted in Figure 3, AUHWM generates ever more precise and real-time-updated WM estimations, therefore resulting in a better adaption of the task execution.

Moving beyond the context of Match^{2}s-like games, z_t could correspond to the period of time before the UI refreshes a previously presented information, ensuring the user will be able to perform a task without forgetting more than $(1 - \pi_d)$ of the information content. In the context of decision-making processes, the UI could make sure that the user is solving a problem while considering all the essential information. For assistive technologies, z_t could stand for the number of presented information items; this would enable patients suffering from Alzheimer's disease to interact with the adapted UI autonomously, without the help of family or caregivers, restoring some of their lost autonomy.

7.2 UI Adaptation Experimental Evaluation

We tested the adaptation framework introduced in Figure 6 using the data from the 18 Match^{2}s users, once again. In order to do so, instead of having the Adaptive Optimizer find the optimal task parameters z_t that would result in the performance π_d, we assume that T_t and k_t, i.e., the hiding time and number of squares presented at batch t, are already the optimal parameters. Therefore the user measured performance π_m must be equal to π_d. This corresponds to having a perfect Adaptive Optimizer that even when presented with faulty quanta estimations, finds the optimized task parameters.

Therefore, if AUHWM correctly predicts the user's current cognitive capacity, the previously estimated quanta number output by AUHWM, Q_{t-1}, together with the corresponding task parameters T_t and k_t, when run through our GB model, should result in π_m. In practice, this means running the same test as the one of Section 6, but while "looking ahead", that is, using the state Q_{t-1} to predict the recall probability $r_{Q_{t-1},k_t}(T_t)$ of the next turn. Once again, AUHWM is initialized with a state $Q_0 \sim \mathcal{N}(100, 32)$, and run to obtain the players' quanta estimates according to the measured performance π_m.

Using the obtained quanta-number estimates as well as the corresponding task parameters as inputs of the GB model, one obtains, for one of the 18 players, the recall probability curve presented in Figure 7. Remember that in this configuration, we are "looking ahead", and thus the recall curve is given by $f(Q_{t-1}, k_t, T_t)$. Of course, since the initial guess for Q_0 is very generic, Figure 7 shows that the system takes at least two interactions to zoom towards the user's capacity; also, the performance predictions are not as accurate as the ones in Figure 4, which provide a posteriori estimations.

To assess more globally the performance of this adaptation framework, we computed the RMSE of the actual vs. estimated recall probabilities for all 18 players, by batch. The resulting curve is presented in Figure 8. This time, the mean of the last three errors is slightly more than 33%, which is considerably worse than the result obtained when assessing the quanta locally (that is, the result presented in Figure 5). The reason for this performance deterioration

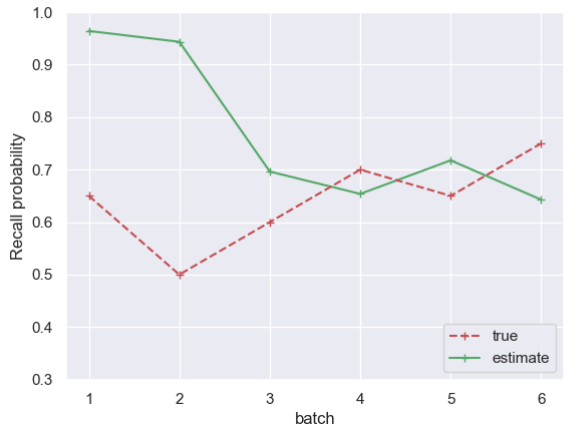

Figure 7: Example of predicted (using last quanta estimate) vs. actual recall probabilities for one Match^{2}s player.

is that there are a number of factors that result in performance variation other than cognitive capacity. For instance, from one batch to the next, some players lost motivation, due to the task being repetitive and therefore became less attentive (remember that motivation and attention are key factors capable of modulating WM). Moreover, some players started developing better strategies, or became used to Match^{2}s, resulting in better scores and therefore apparent higher cognitive capacities. The used model of memory dynamics doesn't take all these factors into consideration (but see Section 8), which consequently leads to fluctuations on the estimated quanta numbers from batch to batch.

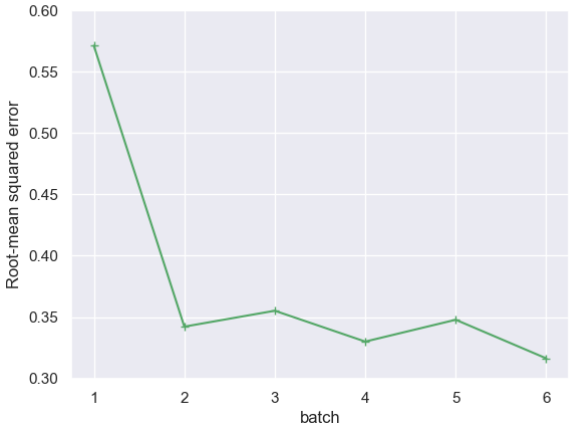

Figure 8: Evolution of the recall probability RMSE for the 18 Match^{2}s players, per batch, when "looking ahead" (prediction based on previous batch quanta number).

8 FUTURE WORK

At the fundamental level, it would be interesting to look at ways to improve the UKF transition function for Q_t. For instance, a quantitative estimate of a user's attention to the task at hand would clearly impact positively the assessment of the short-term evolution of her cognitive retention capabilities. A first approach to such an estimation process could be via the use of dedicated sensors, as brain computer interfaces (BCI) such as the EEG-based headband Muse[4], which can be used to assess attention estimates in real time. Even though this doesn't constitute a workable solution in the long term for obvious usability reasons, such a study could nonetheless provide ways to refine the AUHWM memory capacity tracking process, and spur further research into finding more pragmatic ways to assess users' attention. In practice, this would result in having attention estimates entering the AUHWM and Adaptive Optimizer modules in Figure 6. AUHWM would then be able to assess changes in concentration in real time by the use of the BCI device. Therefore, instead of having the quanta estimates changing from batch to batch as seen in Section 7, the quanta estimations could be modulated by changes in concentration levels, increasing or decreasing the predicted performance accordingly. The Adaptive Optimizer could also take into account local changes in attention to adjust the task parameters and better adapt the interface (or could, for instance, issue sounds to demand a certain attention level).

Looking at practical applications, future work should focus on applying AUHWM to more meaningful UI-adaptation use cases than Match[2]s. For instance, a framework such as AUHWM could be adapted to perform scaffolding in intelligent tutoring systems [1]. By adding a theoretical transition model of learning and taking into account users' accuracy and reaction times into its observation model, AUHWM could be used to track a second parameter corresponding to the user's mastery of the knowledge being tutored. Therefore, complex applications such as Photoshop or CAD software, which are daunting for novice users, could be personalized by increasing the complexity and amount of information presented according to the user's competence and/or cognitive capacity.

9 CONCLUSION

We introduced AUHWM, a new paradigm for tracking human WM that we posit could be a key component for the adaptation of UIs to users' cognitive limitations. AUHWM employs a model of WM dynamics implemented as a GB model, and is able to assess in real time users' memory recall abilities. The time adaptation of WM is based on UKF-tracked estimates of users' memory capacity, measured in numbers of memory quanta. AUHWM has been experimentally proven successful when tracking the cognitive capacity of 18 players in an existing visual memory game, thus providing a strong degree of assurance about the model pertinence.

Short-term fluctuations of users' motivation, attention and fatigue, not taken into account yet, usually result in significant changes on cognitive abilities. However, preliminary experiments suggest that current AUHWM-tracked quanta estimations already provide some crude prediction capability for the assessment of users' future performance, up to about 33%. Given the UKF's capability of improving its estimations through sensor fusion, AUHWM could clearly be enhanced through the use of BCI devices. They would help modulate the quanta estimations according to users' concentration and other key factors that drive WM capacity, therefore improving further the automatic adaptation of UIs.

We believe that AUHWM can thus be of great value when developing UIs that are sensitive to users' cognitive abilities. More specifically, in the context of assistive technologies, adapting interfaces to patients' capabilities would be of great benefit to individuals suffering from memory deficits. Moreover, a memory tracking framework such AUHWM can assess if a person is tired, and therefore not in her best capacity for optimal decision making, or can perform the adaptation of an interface in order for the user to be able to fully grasp and consider what is being presented.

[4]https://choosemuse.com

ACKNOWLEDGMENTS

We thank Patryk Kiepas at CRI (MINES ParisTech, PSL University) for useful discussions and help.

REFERENCES

[1] John R. Anderson, C. Franklin Boyle, and Brian J. Reiser. 1985. Intelligent Tutoring Systems. *Science* 228, 4698 (1985), 456–462.
[2] Alan Baddeley. 1992. Working Memory. *Science* 255, 5044 (1992), 556–559.
[3] Annette Brose, Florian Schmiedek, Martin Lövdén, and Ulman Lindenberger. 2012. Daily variability in working memory is coupled with negative affect: the role of attention and motivation. *Emotion* 12 3 (2012), 605–17.
[4] N. Cowan. 2001. The magical number 4 in short term memory. A reconsideration of storage capacity. *Behavioral and Brain Sciences* 24, 4 (2001), 87–186.
[5] Xiaocong Fan, Po-Chun Chen, and John Yen. 2010. Learning HMM-based cognitive load models for supporting human-agent teamwork. *Cognitive Systems Research* 11, 1 (2010), 108–119.
[6] Jacek Gwizdka. 2010. Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology* 61, 11 (2010), 2167–2187.
[7] Michael J. Kane and Randall W. Engle. 2002. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review* (2002).
[8] George A Miller. 1956. The Magical Number 7, Plus or Minus 2 - Some Limits on Our Capacity for Processing Information. *Psychol. Rev.* 63, 2 (1956), 81–97.
[9] P. A P Moran. 1958. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* 54, 1 (1958), 60–71.
[10] Klaus Oberauer and Hsuan-yu Lin. 2016. An Interference Model of Visual Working Memory. *Psychological Review* 124, 1 (2016), 1–39.
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[12] Felix Putze, Mazen Salous, and Tanja Schultz. 2018. Detecting Memory-Based Interaction Obstacles with a Recurrent Neural Model of User Behavior. In *23rd International Conference on Intelligent User Interfaces*. ACM, 205–209.
[13] pykalman. 2012. pykalman 0.9.2 documentation. [Online] https://pykalman.github.io/. (2012).
[14] Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.
[15] B. Massoni Sguerra, A. Benamara, S. Benveniste, and P. Jouvelot. 2018. Adapting Human-Computer Interfaces to Working Memory Limitations Using MATCHS. In *2018 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*. 1309–1314.
[16] Michael Siebers and Ute Schmid. 2018. Please delete that! Why should I? *KI-Künstliche Intelligenz* (2018), 1–10.
[17] Jordan W Suchow, Benjamin Allen, Martin A Nowak, and George A Alvarez. 2013. Evolutionary dynamics of visual memory. *J.of Vision* 13, 9 (2013), 20–20.
[18] Jordan W Suchow and Thomas L Griffiths. 2016. Deciding to remember: memory maintenance as a Markov decision process. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.
[19] Konstantinos Tsiakas, Maher Abujelala, and Fillia Makedon. 2018. Task Engagement as Personalization Feedback for Socially-Assistive Robots and Cognitive Training. *Technologies* 6, 2 (2018), 49.
[20] Eric A Wan and Rudolph Van Der Merwe. 2000. The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000.* Ieee, 153–158.
[21] Elena Zanini. 2014. Markov Decision Processes. [Online] https://www.lancaster.ac.uk/pg/zaninie/MDP.pdf. (2014).