ACT-R models of a delayed match-to sample task

Sarah Cebulski (sarahcebulski@cmail.carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive
Ottawa, On., Canada

Sterling Somers (sterling@sterlingsomers.com)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive
Ottawa, On., Canada

Abstract

The current paper presents two ACT-R models of a delayed match-to sample task, and performs equivalence testing against human performance data to evaluate them. Success of an episodic model which avoids interference from previously encountered visual stimuli, and implements a serial search and rehearsal strategy lends insight into how individuals may encode, maintain and retrieve visual information.

Keywords: ACT-R, visual memory, rehearsal

Introduction

ACT-R (Anderson & Lebiere, 1998) is a cognitive architecture that includes a theory of how higher-level processes interact with a visual system. ACT-R's visual module identifies objects in the visual environment and through the use of buffers passes this information to the declarative memory module in the form of chunks. A chunk is a vector representation of individual properties, and in the case of visual information, is often represented with vector locations of the presented stimuli. Once visual information is represented in declarative memory, it can be retrieved according to task demands. In the past there has been little in the way of research which connects low-level visual processes with high-level cognition. Fortunately, this trend has been reversing over the last several decades and a wealth of research in the ACT-R community examines exactly how low-level processing constrains and influences visual encoding. These constraints include, among others: the time required for visual attentional shifts, the noise accompanying conjunction searches and the feature scale directing object recognition (Anderson, Matessa, & Lebiere, 1997).

Despite strides towards understanding encoding constraints, most computational models of high-level visual processing continue to take visual representations for granted. Many of these models assume representations are deposited into declarative memory once they have been successfully encoded without accounting for intermediate processes between encoding and chunk formation. Often, for example, models do not account for rehearsal strategies that actively maintain complex visual stimuli in memory in order to prevent their decay. Extant models that do include visual rehearsal processes (e.g., Winkelholz & Schlick, 2006) do not do so as a primary research focus, and it is thus difficult to disentangle observed effects owing to rehearsal from those owing to other lines of inquiry. It is thus our aim to examine, as a primary focus, the rehearsal mechanism involved in actively maintaining complex visual stimuli in memory for a brief period of time. Specifically, we are interested in determining whether an ACT-R model implementing a serial rehearsal strategy can account for human performance differences observed across two versions of a delayed match-to sample task.

Versions of the delayed match-to sample task exist throughout the literature (Della Sala, Gray, Baddeley, Allamano, & Wilson, 1999; Warrington & James, 1967). In its most basic form, the task requires participants to encode a matrix grid pattern, rehearse it across a delay period, and compare it to a test grid. This task was selected for a number of reasons. First, its simplicity reduces many of the major confounds introduced by individual differences in strategy use, such as the tendency to recode presented visual information verbally. This notion is supported by the finding that articulatory suppression does not impair performance on similar tasks (Salway & Logie, 1995; Vandierendonck, Kemps, Fastame, & Szmalec, 2004). Second, the randomized nature of the grid pattern ensures that the structure does not become more familiar with time, so there is no expectation that implicit learning occurs resulting in faster and more efficient linking of environmental features to object-locations (Winkelholz & Schlick, 2006). Third, the instituted delay period between encoding and retrieval is longer than the time visual information is purported to survive in sensory memory (Phillips, 1974). This necessitates some form of active maintenance or rehearsal strategy. Finally, it is possible to create different versions of the selected task that vary only in complexity, such that a high-workload version contains more visual data to be encoded and rehearsed than a low-workload version.

The present paper describes two ACT-R models of visual rehearsal. As a starting point, both models assume similar low-level processes, with absolute screen position used to encode visual stimuli in a serial fashion (i.e. objects are encoded as single chunks, without any Gestalt-type grouping). If model performance employing this serial encoding and rehearsal strategy does not fit the experimental data, it would suggest differences in encoding strategies (i.e., perceptual grouping of visual information) should be investigated in future work. The two models diverge in their implementation insofar as whether they represent each trial as an episode. While one model allows

visual information encountered on previous trials (i.e., previous episodes) to interfere with the encoding, maintenance and retrieval of the current trial (i.e. current episode), the second model tags the current episode, encoding a slot/value pair maintained in the *imaginal* buffer (updated each trial) into the memory chunk. Rehearsal and recall uses the slot/value pair from the *imaginal* buffer in all rehearsal and retrieval, preventing interference from previous episodes (tagged with a different episode value). Performance data is generated for each model as it performs a low- and high-workload version of a delayed match-to sample task, and is compared to human performance data from a behavioural experiment using the same tasks. It is predicted that the high-workload version of the delayed match-to sample task will be accompanied by increased rehearsal demands that will account for increased response times and decreased accuracy measures in the high- relative to low-workload versions of the task.

The first part of the paper describes the task itself, as well as the design and results of the behavioural experiment mentioned. The second part of the paper describes the two ACT-R models, presenting major differences between them. The final section of the paper fits the model parameters for threshold, latency and noise to the human performance data, and discusses implications of the findings.

Behavioural Experiment

The behavioural experiment was conducted in order to generate empirical data for comparison to computational models, as well as to determine whether the performance on a low-workload version of a delayed match-to sample task was better than performance on a high-workload version of the task.

Materials

The tasks used throughout this paper are two versions of computerized delayed match-to sample tasks that vary in complexity. Fig. 1 shows the matrix structures that were used, which included a 5x5 grid with 4 shaded cells for the low-workload condition, and a 7x7 grid with 7 shaded cells for the high-workload condition. The location of shaded cells was randomized with the constraint that no two adjacent cells be filled.

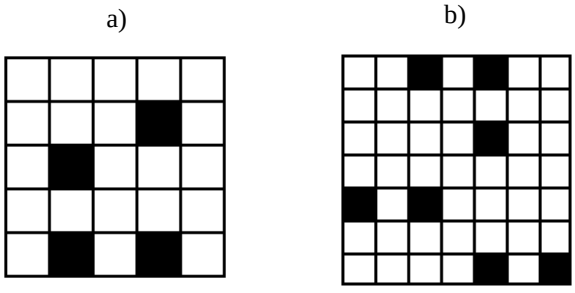

Figure 1: Sample grids: a) low-workload, b) high-workload condition.

Design and Procedure

A single factor (workload: low vs. high) repeated measures design was used. Each participant completed two blocks, one for each workload version of the task. Blocks were counterbalanced across participants, and each block consisted of 30 trials. Each trial consisted of three phases: encoding, maintenance and retrieval. The encoding phase began with the presentation of a study matrix for 2000 ms. After study matrix presentation, the participant maintained the visual information in memory for 2500 ms, during which time a masked screen was presented. This time period is longer than the time visual information is purported to survive in sensory memory (Phillips, 1974), but shorter than the time required for it to enter long-term memory. This necessitates some form of active maintenance or rehearsal strategy and renders simple retrieval from long-term memory unlikely (Cowan, 2008). After the delay screen, a test matrix was presented which either matched the study matrix exactly, or did not. Non-matching test matrices were consistent in that the shaded cell had been relocated to an unfilled non-adjacent cell. Participants were required to retrieve their representation of the study matrix, and using a response pad, respond "yes" if the study matrix matched the test matrix, and "no" if it did not. Participants had 3000 ms to respond, and were told to try to respond within this time. If no response was made within 3000 ms, a timeout screen was presented, and the trial was labelled as a "miss". Missed trials, which accounted for 2.10% of the entire data, were not included in analysis. Participants included 11 individuals (5 male, 6 female; mean age 20.0 years) recruited through Carleton University's SONA system.

Results

A repeated measures ANOVA revealed a main effect of workload ($p < 0.001$) such that percent accuracy was higher and reaction time was lower in the low-workload condition relative to the high-workload condition (Table 1).

Table 1: Contingency table of means and 95% confidence intervals for reaction time (RT) and accuracies (ACC) at low- and high-workload conditions.

	Low RT	Low ACC	High RT	High ACC
Lower Bound	1.105	0.865	1.437	0.712
Mean	1.266	0.93	1.584	0.75
Upper Bound	1.33	0.960	1.681	0.802

Models

The two models created to investigate the human performance data presented were written in the Python

variant of ACT-R (Stewart & West, 2005). The first, which we call the *interference* model, is the most naïve model. The second, which we call the *episodic* model, was built in response to early analysis of the interference model. We consider both models to be early in development. As will be discussed, we expect further behavioural measures to help guide which model offers a better explanation of the data. This section will outline both models, highlighting the key differences between the two.

One of the key aims of the research presented here is to gain insight into whether a simple, serial encoding of visual stimuli is used by individuals in our delayed match-to sample task. Previous research regarding visual encoding of stimuli (Anderson et al., 2004; Ehret, 2002) suggests that visual encoding of items on a computer screen can be accomplished using the computer screen itself as a reference frame and encoding (*x,y)* screen coordinates based on this frame of reference. Work by Winkelhoz & Schlick (2006) suggests that a more complex visual encoding is used. Though they present their own vision module with its own set of sub-symbolic parameters, we find their model to be more complex than necessary as dictated by the needs of our experimental design. We instead adopt a simpler approach which attempts to model the encoding, maintenance and retrieval phases of the experimental task.

Encoding Phase

Both models use Python ACT-R's SOS (Simple Operating System) vision module (West & Emond, 2002) to perceive the environment. The SOS vision system makes use of a chunk-based representation environment. In our task, the chunks representing the environment consist of a slot for *isa*, two slots which represent the absolute coordinates (*x, y*) of the filled cells, a slot for *location*, and a slot to represent *salience*. The salience of all filled cells was set to 1.0. The SOS vision system is intended as a first-pass vision system where reaction time and attention simulation does not need to be as accurate as vision systems such as EMMA (Eye Movements and Movement of Attention) (Salvucci, 2001). SOS uses the salience factor to probabilistically choose which visual chunk to push into the visual buffer. Because estimates of scanning time are not used, all vision requests take 85 ms. The SOS vision system assumes that over a number of trials, the scanning differences are averaged out.

The interference model is the more naïve of our two models in that once a filled cell is detected, the cell's *x,y* coordinate is simply stored in declarative memory. The episodic model, however, encodes the *x,y* position together with the contents of the *imaginal* buffer, which also contains a representation of the trial (i.e., a slot/value pair tagging the filled cells as belonging to the current trial). In the episodic model, the *imaginal* buffer keeps track of the trial with a slot *trial* and a value which increments at the end of each trial. Retrieval and rehearsal of cells includes the slot/value pair representing the current episode maintained in the *imaginal* buffer.

Finally, while the grid is still visible, both models rehearse the grid by re-scanning it. To ensure that the model scans the entire grid before re-scanning, a visual *finst* was added to the SOS vision module. For simplification, the finst size is set to 7 to account for both conditions.

Maintenance and Recall

During the maintenance phase, both models rehearse chunks from memory. The maintenance phase is essentially a production loop which continually conducts declarative memory request for any recalled block. To avoid rehearsing the same cell continually during this period (resulting from a high activation of the first retrieved filled cell), a declarative memory *finst* is used. The model will rehearse from memory as many times as it can until it sees a new grid, which is the indicator that the recall portion of the trial has begun.

The recall phase uses a first failure strategy to reject the test matrix. Like in the encoding phase, a visual finst drives search for new filled cells. When a cell's *x,y* coordinates are matched in the visual buffer, a declarative memory request for those coordinates are made. If the declarative memory retrieval is successful, the recall phase loops to the next filled cell. If the recall is unsuccessful, the model assumes that this is an indication that the model has not seen that cell configuration before and a negative response is issued. A positive response is only issued when a finst-enabled vision module request fails (indicating the model has looked up all the filled cells) and no declarative memory failure occurs.

The interference model is so named because early analysis indicated that some false-positives resulted from the model rehearsing and recalling filled cells it had experienced in a previous trial. Although no statistical analyses were conducted in terms of false-positives between the model and the behavioural data, we decided to implement a second model, the episodic model, and compare results of this model to performance on the interference model as well as empirical data.

Validation of the models

One of our underlying assumptions for the current versions of our models is that individuals use the same strategy for low- and high- workload versions of the visual task. Given this assumption, a model should account for performance data across both workload versions of the task, since a good cognitive theory explains many different kinds of empirical findings (Simon & Wallach, 1999; Stewart & West, 2010). In order to assess model validity, we therefore tested for equivalence between the model and human performance across parameter space. The models were also tested for equivalence across the four performance variables measured: reaction time and accuracy measures in the low-workload condition, and reaction time and accuracy measures in the high-workload condition. The model was required to pass equivalence testing at each of the four performance variables in order to be considered to be predictive of the human data.

Equivalence testing

Traditionally, the success of ACT-R models is evaluated based on the magnitude of the Root Mean Squared Difference between the model and real-world data (Stewart & West, 2010). This approach, however, is problematic since it does not properly weight sampling error, and fails to consider that the true value of the mean can lie anywhere within the sample confidence interval with equal probability (Tryon, 2001). A better approach, as suggested by Stewart and West (2010), is to identify a set of models that could be correct, and use equivalence testing to indicate that there is insufficient evidence to distinguish between them.

In line with this, inferential confidence intervals were determined for the four performance variables for the human data, as well as for each model at each set of considered parameters as suggested by Tyron (2001). Equivalence testing was then performed, whereby maximum likely differences (MLD) were calculated reflecting the maximum difference between the model data (at a given parameter set) and the human data. The values were calculated according to Equation 1, where R_l to R_u are the 95% inferential confidence intervals for the real-world (i.e., empirical) data, and M_l to M_u are the model 95% confidence intervals.

$$MLD = \max(M_u - R_l, R_u - M_l) \quad (1)$$

When the MLD is less than a threshold value, which is the maximum difference deemed unimportant on substantive grounds (Tyron, 2001), then the 95% CI test for statistical equivalence is also satisfied. The minimum threshold value for fitting computational data to human data is suggested to be the size of the confidence interval of the real-world data (Stewart & West, 2010). However, because there is an important difference between how human participants and our models perform the experimental tasks, we suggest there are grounds for increasing this threshold. The reason for this is that while it is likely human participants guess on a proportion of their responses, especially since they were encouraged to respond within the 3000 ms timeout period, no guessing occurs in our ACT-R models. Unfortunately, modifying the models to accommodate for guessing is not a simple task, and is beyond the aims of the current paper. Individual guesses are almost certainly not random, but rather, based on complex probabilistic mechanisms related to the level of uncertainty, the activation level of shaded cells within memory, and the ratio of previous responses. In fact, including a simple guessing strategy (e.g., guessing "yes" half of the time on a subset of trials) may decrease the fit of the model to the experimental data since it may not reflect the actual mechanisms individuals employ when guessing. Guessing increases the noise for both the reaction times and accuracies of human data relative to computational data, and should therefore increase the acceptable level of error in the model. Rather than tackle this issue by including guessing strategies in our model, we suggest increasing the threshold and examining the resulting set of parameters where the model is equivalent on all measures to the empirical data (i.e. the MLD is less than the threshold for all of the measures considered). Based on a recent study (Kemps & Andrade, 2012) that employed similar visual stimuli and found individuals were 'sure' of their responses approximately 80% of the time, we opted to increase the threshold by a factor of 0.2.

Results

Equivalence testing revealed that the interference model is not equivalent to the empirical data for any of the parameters searched. The episodic model, on the other hand, is equivalent to the empirical data at a range of parameters between thresholds of 0.45-0.6, latencies of 0.25 and 0.315 at a noise of 0.5 (Figure 2).

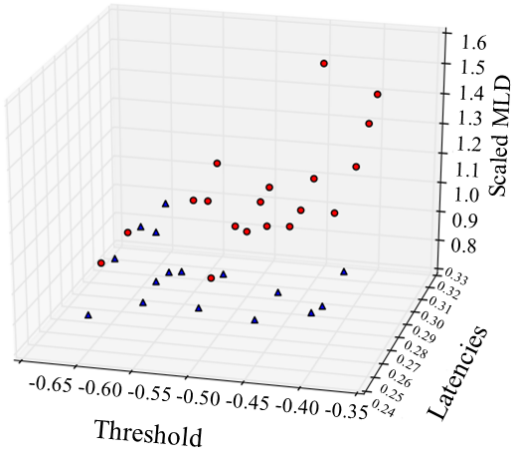

Figure 2: Statistical equivalence of episodic model data to empirical data. Blue triangles represent models that fall below the threshold (success) and the red circles represent models that are above the threshold.

An example of statistical equivalence between reaction time data generated by the episodic model for the low-workload condition and human performance data on the same condition is presented in Figure 3. From this figure, it is apparent that the maximum likely difference (MLD) between the empirical data (Y_2) and the model data (Y_1) is less than the threshold data (Delta) that represents our acceptable level of error.

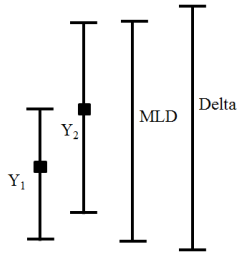

Figure 3: Statistical equivalence of episodic model data to empirical data

Discussion

The interference model did not pass equivalence testing on all four empirical measures in the parameter space searched, indicating that it is a poor fit to the behavioural data. This was in part expected upon an initial investigation of errors, as mentioned, which revealed a bias towards false-positives. Of course, it is possible that the observed poorness of fit does not necessarily point to an inaccurate model, but rather problematic empirical data. However, because the second model experienced more success over a relatively broad parameter space, it is believed that the failure of the interference model to fit the empirical data is due to a failure of the model itself, rather than a problem with the empirical data. This failure could owe to an inability of the model to account for the visual rehearsal mechanisms actually used by individuals, or to the interference of visual information from previous trials—interference that was not actually encountered by individuals performing the behavioural experiment.

The episodic model, which reduced this interference, was met with more success than the interference model. The fact that reducing visual interference resulted in a model that passed equivalence testing across a relatively broad parameter space, and across all four performance variables bolsters the suspicion that interference was behind the poorness of fit in the original model, and that visual information within a given trial does not suffer significantly from interference with visual stimuli seen in previous trials. The broad coverage of the episodic model is an indication that this model, especially as a starting point, is a potential candidate for modelling how human participants actually search for and rehearse visual information.

In order to expand on this research, and to refine the current model, more behavioural data is necessary to confirm and expand on the trends currently seen. Eye-tracking data, in particular, will help to guide the next steps that relate to encoding strategies (e.g., perceptual grouping of shaded cells) that might increase the fit of our episodic model. It will also be important to modify the delayed match-to sample task such that guessed responses are more difficult to make. Performing a broader search of parameter space will also shed light onto how broad the coverage of the current model extends, and can guide modelling. The use of the maximum likely differences method is particularly useful in this regard, since the parameter space that passes equivalence can be visualized iteratively, and areas in space that do not fit well to the performance data can be isolated, and reasons for this poorness of fit considered.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. Psychological Review, 111(4), 1036–60

Anderson, J. R., & Lebiere, C. (1998). The Atomic Components of Thought. Mahwah, NJ: Lawrence Erlbaum Associates Ltd.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention. Human–Computer Interaction, 12(4), 439–462

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? Prog Brain Res., 169, 323-338

Della Sala, S., Gray, C., Baddeley, a, Allamano, N., & Wilson, L. (1999). Pattern span: a tool for unwelding visuo-spatial memory. Neuropsychologia, 37(10), 1189–99

Ehret, B. (2002). Learning where to look: Location learning in graphical user interfaces. Proceedings of the SIGCHI Conference on Human …, (4), 211–218

Kemps, E., & Andrade, J. (2012). Dynamic visual noise reduces confidence in short-term memory for visual information. Cogn Processes., 13, 183-188

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. Perception & Psychophysics, 16(2), 283–290

Salway, A.F.S., & Logie, R.H. (1995). Visuospatial working memory, movement control and executive demands. British Journal of Psychology. 86, 253-269

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. Cognitive Systems Research, 1(4), 201–220

Simon, H., & Wallach, D. (1999). Cognitive modeling in perspective. Kognitionswissenschaft, 8, 1–4

Stewart, T. C., & West, R. L. (2005). Python ACT-R: A New Implementation and a New Syntax. In 12th Annual ACT-R Workshop

Stewart, T.C., & West, R. (2010). Testing for Equivalence: A Methodology for Computational Cognitive Modelling. Journal of Artificial General Intelligence, 2(2), 69–87

Tryon, W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis. Psychological Methods, 6(4), 371–386

Vandierendonck, A., Kemps, E., Fastame, M.C., & Szmalec, A. (2004). Working memory components of the Corsi block task. British Journal of Psychology. 95(1), 57-79

Warrington, E. K., & James, M. (1967). Disorders of visual perception in patients with localised cerebral lesions. Neuropsychologia, 5(3), 253–266

West, R. L., & Emond, B. (2002). SOS: A Simple Operating System for modeling HCI with ACT-R. In Seventh Annual ACT-R Workship. Pittsburg, PA.

Winkelholz, C., & Schlick, C. M. (2006). Modeling human spatial memory within a symbolic architecture of cognition. In T. Barkowsky, M. Knauff, G. Ligozat, & D. R. Montello (Eds.), Lecture notes in artificail intelligence: Proceedings of Spatial Cognition (pp. 229–248). Berlin, Germany: Springer-Verlag