Games with Hypotheses: Color, Text and Texture

Nathan Moroney[1] and Sabine Süsstrunk[2]
Hewlett-Packard Laboratories[1], Palo Alto, USA and Ecole Polytechnique Fédérale (EPFL)[2], Lausanne, Switzerland

Abstract

An online memory matching game is used to explore the collective performance for stimulus sets varying in color, text and texture. The game is a consistent five wide by four high array of 70 by 70 pixels squares for a total of eleven unique test pairs. Users click squares to find matching pairs and once all of the pairs have been found the time to complete and number of clicks to complete are saved on a server. For the initial draft of this paper eleven images sets were tested. In three cases the test pairs were solid colors and corresponded to a basic color set and two non-basic color sets. In four cases the test pairs were the text corresponding to color terms. The text cases included black 12 pixel Arial for the basic color terms and two test cases for the non-basic color terms used previously. The text cases also included a set with Stroop colored basic color terms or text colored roughly the opposite to the corresponding color term. Finally, four texture image sets taken from the Outex texture database[11] were used for testing. Two of the texture sets were for higher key, mostly white textures sets of wallpaper and flour. Two of the texture sets were for more colorful images of textiles and floors. The analysis is presented for a web-based experimental game based on the completion time and number of clicks to completion. The use of single simple game unifies each of these perception and memory tasks.

Introduction

The use of the web for conducting color experiments has been previously described. This paper extends this work in several ways. First a game is used as the primary interface to web volunteers and the responses to be analyzed are the collective performance. Second the game is simple enough that it can be used to consider multiple, interrelated domains such as solid colored patches, texture and written color terms.

There is a range of related current work[1-7] on the use of games for either research or engineering purposes. Applications of labeling games for machine learning have been shown by multiple authors. The use of action videos games has been proposed to enhance the contrast sensitivity function of participants. Large-scale analysis of chess playing has been performed to explore rapid human decision making. The game of concentration has also previously been used to assess the visuospatial memory of children[8] and monkeys[9].

Our contribution is on a web-scale version of the game of concentration, focused on perceptual stimuli sets with the potential to generalize to other specific hypotheses within a consistent and familiar framework. The game is also simple enough that it can be modeled in its ideal form, for example for an ideal participant with perfect memory or a completely random player. The fundamental hypothesis is that collecting web-scale user data for the game of concentration based on widely different stimuli sets can yield statistically significant differences in performance across stimuli. This in turns implies the possibility of a univariate scaling of differences between sets of stimuli using a game interface. Therefore the focus of this paper is on use of a web-scale game not for labeling, therapeutic benefits or analysis but for testing of specific hypotheses.

Experiment

The experimental game used was the concentration game, also known as pairs, shinkei-suijaku, hūsker dū or the memory game. The game consists of a five wide by four high array of square patches. The squares were all 70 by 70 pixels with gaps between each of the patches. Upon loading the web page the experiment would be initialized with one of eleven sets of experimental patches. Initially the user sees the instructions and all of the patches as a gray patch with a single pixel black border. Below the patches is a button that must be pressed to start the game. Once the button is pressed the button becomes a timer displaying the minutes and seconds taken. Figure 1.A and 1.B shows a screen shot of the initial view of the experiment and the game underway. The objective of the participant is to find matching pairs on the reverse side of the patches. The participant can examine the patches two at a time and once a pair matches they stay visible. Selected pairs remain visible as long as no new sample has been selected. The goal is to complete the matching process as quickly as possible.

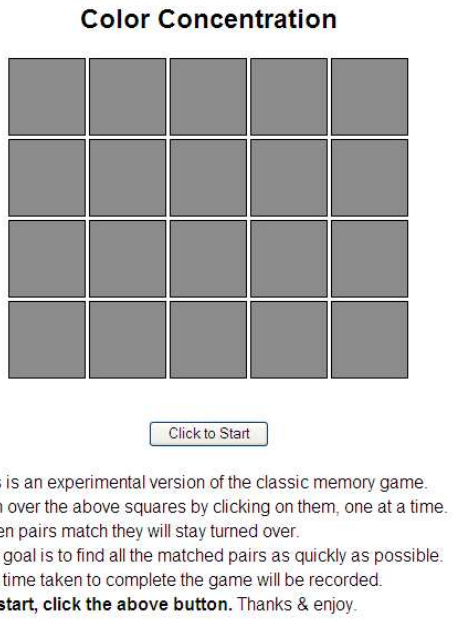


Figure 1.A. *Screen shot of the central portion of the experimental color concentration game at the beginning of play.*

Figure 1.B. *Screen shot of the central portion of the experimental color concentration game after 38 seconds of play and four matched pairs have been found.*

The eleven experimental sets used for the game are now described in additional detail. First is a set of solid colored and text patches, shown on the left and center columns of table 1. This table shows the samples from one to ten one per row and each column as an experiment. The second column shows the solid basic color patches. This set of ten colors consisted of the ten color centroids for the validated Munroe and Ellis data[10] for a sub-set of the color terms hypothesized to be more universally shared across languages and arising in a partially fixed sequence during language evolution. The third and fourth columns are for two sets of ten non-basic color pairs. The colors were also based on the Munroe and Ellis data and were selected to be both approximately in order of frequency of use and also to have a minimum ΔE^*_{ab} from neighbors of more than 20.

The next three columns show the patches for columns two through four in corresponding text form. The color terms were show as black text on a white background and were rendered as 12 pixel high, anti-aliased Arial text. Finally a Stroop set of basic color terms was used. In this case the color terms were the basic terms but the color of the text was not black but was its color antonym. In all cases the patches were surrounded by a one pixel black border.

The remaining four experimental test sets are a sub-set of texture images from the Outex texture database.[11] These images were all downloaded as 100 dpi, Inca illuminated BMP files and were then scaled down by 30% before being cropped to 70 by 70 pixels. The specific texture image used are shown in the right columns of Table 1. Column nine shows the texture patches taken from the wallpaper texture series or a set of high key or nearly white wall coverings. Column ten shows the texture patches taken from the floor series or a muted color set of floors samples. Column eleven shows the texture patches taken from the flour series or a set of surfaces covered with a layer of shades of nearly white flour. Column twelve shows the texture patches taken from the canvas series or a set of textile samples. These four texture sets were used to provide a range of colors and textures ranging from more or less similar.

Results

The first step in analyzing the results is to consider the participants performance in the concentration game across all the game conditions. Over 1000+ trials of the game are therefore considered in aggregate and the distribution of total number of clicks and total number of seconds to complete the game is considered. This data is shown plotted in Figure 2. The median number of clicks across all games is 50 and the median number of seconds to complete is 53.5 seconds, or roughly one click every 1.07 seconds. The fastest and slowest completion times differ roughly by a factor of 3. To minimize privacy concerns, no demographic data was collected and no cookies were used. Therefore for these results it is not possible to determine the number of unique subjects or cluster participants.

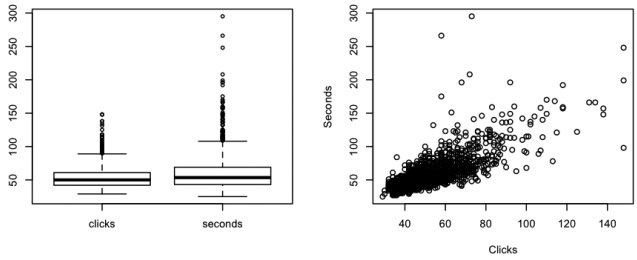


Figure 2. *Box plots for number of clicks and number of seconds for all games played, shown on left. Plot of clicks versus seconds for all games played, shown on the right.*

The above results can also be shown according to their sorted values. That is across all games and participants the data can be sorted from least to most. This provides a rough visualization of the distribution and as can be seen in Figure 3 strongly suggests a skewed distribution. The Shapiro-Wilks tests with an alpha level of 0.05 results in a rejection of the hypotheses that total number of clicks and total time elapsed are normally distributed. This is skewed distribution is also evident is the histograms of the data, shown in Figure 4.

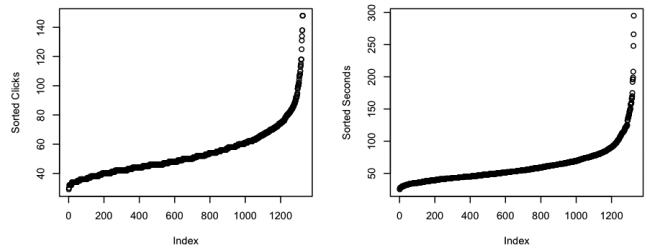
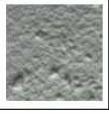


Figure 3. *The game results sorted by number of clicks, on the left, and the number of seconds, on the right.*

Table 1. Listing of all stimuli used in the web-based game of concentration. The first row is the name or description of the experiment. The second row is the experimental identification number is in order that the stimuli set were added to the game database and are not in order numerically. The first column is the stimulus number per experiment. The second through fourth columns are for solid colored basic and non-basic colors. The fifth through eighth columns are for text basic and non-basic colors and includes a Stroop colored basic color text. The ninth through twelfth columns are for the wallpaper, floor, flour and canvas Outex textures. All text was rendered as 12 pixel high Arial using Gimp. Solid colors were computed using the results of lab-validated web-scale color naming database. The Stroop inverted colors were computed using Yc_ic_{ii}. The texture images were selected from the Outex texture database.[11]

Name	Solid Basic	Solid Non-Basic One	Solid Non-Basic Two	Text Basic	Text Non-Basic One	Text Non-Basic Two	Text Basic Stroop	Wallpaper Textures	Floor Textures	Flour Textures	Canvas Textures
No.	1	2	5	3	6	7	4	8	9	10	11
1				black	lime	teal	orange				
2				blue	sky	lilac	white				
3				brown	magenta	mustard	black				
4				green	olive	fuchsia	red				
5				orange	aqua	peach	green				
6				pink	lavender	salmon	blue				
7				purple	maroon	beige	yellow				
8				red	tan	seafoam	pink				
9				white	rust	burgundy	purple				
10				yellow	cream	cobalt	brown				

The results shown in Figure 4 are limited to being positive values and skew to the left. These characteristics are also common to log-normal distributions. In fact using the MASS package in R results in an excellent fit to the log-normal distribution. This is shown for the total number of seconds elapsed data in Figure 5. The log-normal distribution is associated with events and phenomena exhibiting multiplicative errors.[12]

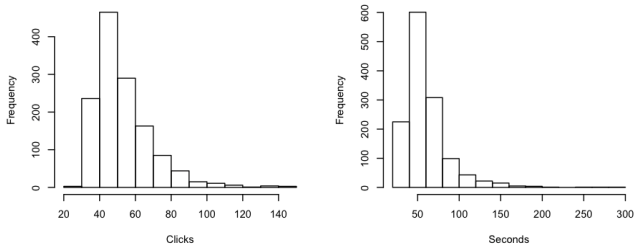


Figure 4. *The histogram of the total number of clicks, on the left, and the histogram of the total seconds elapsed, on the right.*

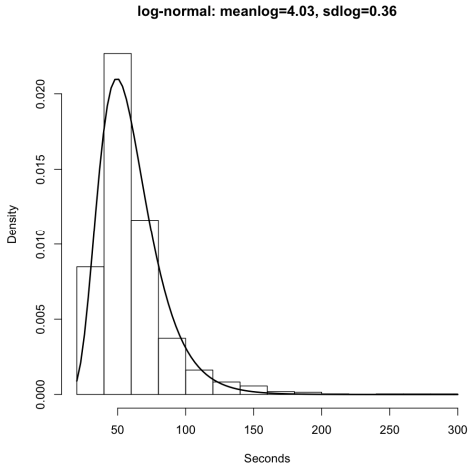


Figure 5. *Log-normal fit of the aggregate number of seconds taken by participants across all experiments*

For the next portion of the analysis we consider individual experiments. The median time in seconds to complete the various experiments is shown in Figure 6. This figure shows the experiments listed on the x-axis and the medians are shown with the corresponding bootstrapped standard errors, with n equal to 30 and replacement used. This plot shows the median time to complete in seconds ranging from 45 seconds for experiment 6 to 87 seconds for experiment ten. The plot also shows that the scale of the experiment, with on the order of 100 games completed per experiment, standard errors for the median ranging from 1 to 4 seconds.

The Wilcoxon-Mann-Whitney rank sum test was used to compare the eleven distributions. At an alpha level of 0.05 it is possible to accept or reject the hypothesis that the distributions are identical or not. Note that with the use of the internet means that each experimental stimuli or game trial will have been played on the order of 100 times.

Specifically, for solid colors the basic-colored stimuli of experiment 1 and the non-basic colored stimuli of experiments 2 and 5 result in p-values which lead us to reject that they are from the same distribution. However, comparing the two non-basic colored stimuli of experiments 2 and 5 result in a large p-value that supports the hypothesis that the distributions are identical. Re-stated the data supports the hypothesis that the resulting distribution of times for basic colored patches is different than that of the two non-basic colored patches but does not support the hypothesis that there is any difference between the two different non-basic colored stimuli.

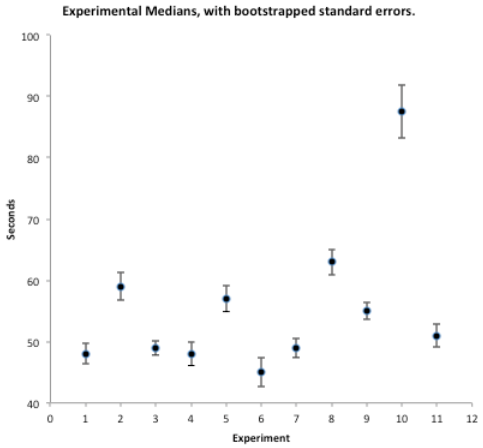


Figure 6. *Individual experimental median time to complete experiments, in seconds, shown with bootstrapped standard errors per experiment.*

For the text or color term experiments however, the test results yield a sufficiently large p-value that supports the hypothesis that the distributions are identical. This is the case even for the Stroop-colored terms. This suggests that while distributions for solid colored patches will have statistically significant differences, the text or color terms will not. There is also no statistically significant difference in the distributions for basic colors as solid patches or as text.

The textured stimuli showed the largest range of results. Textured stimuli sets 11 and 9 did result in p-value test results supporting the hypothesis that they differ. But these two sets differed from both 8 and 10. In addition 8 differed from 10. It is perhaps not surprising that patches of flour make for a harder game, but this result allows a specific quantification of the difference and suggests a means of sorting textures by concentration times. Note however that the fastest results for the colored textures of experiment 11 do not test as different than the solid colored patches. This demonstrates that there are textured stimuli that are comparable to text and solid basic colors.

Figure 7 shows an alternative visualization of the results from Figure 6. In this figure the stimuli sets from table one were plotted directly onto an x-axis. This axis is the median time in seconds to complete the experiment shown in Figure 6 but in this case without error bars and with the different experiments encoded visually. Note that for the results furthest to the left, multiple experiments

plot on top of each other (such as experiment 1) and only one of the results could be shown. Regardless this visualization shows the spread of the data, as well as specific trends. There is clearly a gap between 63 seconds and 87 seconds to be explored in greater detail. There is nearly a factor of two difference between the fastest and slowest stimuli sets.

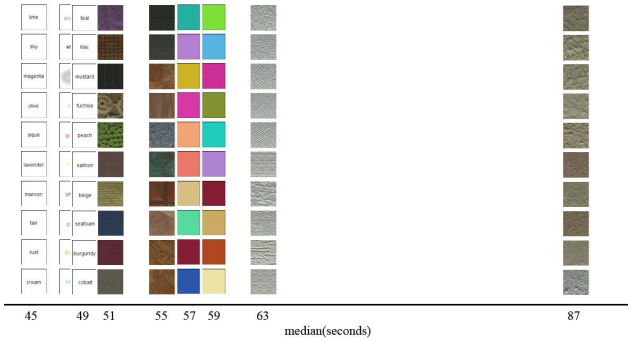


Figure 7. *Sorted median time to completion in seconds for the stimuli shown in table 1 plotted according to the medians shown in Figure 6.*

Discussion

The results in the previous section are encouraging. They suggest specific statistical considerations for visuospatial memory. That is that overall the task follows a log-normal distribution and specific experimental stimuli sets will result in statistically significant differences in distributions. The test case has focused on color, both in text form and as solid and textured patches, This has suggested some interesting similarities and differences between performance for text, solid colors and textures. Work is ongoing to extend this methodology to additional experimental testing of additional stimuli sets, using differing test devices and for a broader range of test domains.

As noted in the introduction, the basic game of concentration in its simplest form can also be implemented as a theoretical simulation based on specific assumptions relating to an ideal or non-ideal game strategy. To illustrate, consider a game-playing algorithm in which the simulated subject has no memory and randomly selects pairs of un-matched stimuli. This random memory-less strategy provides one point of reference for how actual human observers perform.† On the other extreme, consider a game playing algorithm in which the subject has perfect memory and uses an on-repeating search of un-matched stimuli. Both of these simulations were implemented as simulations and run 1000 times assuming ten pairs of arbitrary pairs to be matched. For the random memory-less strategy the result is a median of 188 clicks to complete the game. For the perfect memory non-repeating strategy the result is a median of 28 clicks to complete the game. These simulation results can be compared to those for actual human participants as shown in Figure 2.

Conclusions

A simple game interface has been used to collect over 1000 responses from online participants. The results suggest that statistically significant differences can be found between a subset of the eleven experimental conditions tested. Individuals and games vary by a factor of 3 with respect to their clicks per second for games completed. Solid basic color patches are among the most rapidly completed games while flour texture images are among the slowest of the games completed. Both solid colors and flour textures do not follow Gaussian distributions, although there is overlap in their distributions. Initial results suggest that the use of a game interface with approximately 20 patches is a promising approach to investigating visuospatial memory in an engaging and systematic format. The result is a series of games with hypotheses that provides an intermediate stimulus modality between a solid aperature color or patches and images or saliency maps. A number of future directions are also suggested and described.

References

[1] Luis von Ahn, Ruoran Liu and Manuel Blum, "Peekaboom: A Game for Locating Objects in Images", CHI 2006 Proceedings, pp. 55-64 (2006).

[2] Luis von Ahn and Laura Dabbish, "Designing games with a purpose", *Communications of the ACM*, **51**(8), 2008.

[3] Bei Yuan, Manjari Sapre, and Eelke Folmer "Seek-n-Tag: a game for labeling and classifying virtual world objects", GI '10 Proceedings of Graphics Interface 2010, pp. 201-208.

[4] Ali Dasdan, Chris Drome, and Santanu Kolay, "Thumbs-up: a game for playing to rank search results", Proceeding WWW '09 Proceedings of the 18th international conference on World wide web pp. 1070-1072 (2009).

[5] Renjie Li, Uri Polat, Walter Makous & Daphne Bavelier, "Enhancing the contrast sensitivity function through action video game training", *Nature Neuroscience,* **12**, pp. 549 - 551 (2009).

[6] L. von Ahn, B. Maurer, C. McMillen, D. Abraham and M. Blum, "reCAPTCHA: Human-Based Character Recognition via Web Security Measures", *Science*, vol. **321**, pp. 1465-1468 (2008).

[7] M. Sigman, P. Etchemendy, D. F. Slezak and G. A. Cecchi, "Response time distributions in rapid chess: a large-scale decision making experiment", *Frontiers in Neuroscience*, vol **4,** pp 1-12 (Oct 2010).

[8] R. Schumann-Hengsteler, "Children's and Adult's visuospatial memory: The game Concentration", *Journal of Genetic Psychology*, vol. **157**, no. 1 (Mar 1997) 77.

[9] David A. Washburn and Johnathan P. Gulledge, "A Species Difference in Visuospatial Memory in Adult Humans and Rhesus Monkeys: The Concentration Game", *International Journal of Comparative Psychology*, vol. **15**, pp. 288-302 (2002).

[10] Moroney, N. M. and Beretta, G. B., "Validating large-scale lexical color resources," in Proc. Midterm Meeting of the International Colour Association (AIC) (2011).

[11] Ojala T, Mäenpää T, Pietikäinen M, Viertola J, Kyllönen J and Huovinen S, "Outex - New framework for empirical evaluation of texture analysis algorithms.", Proc. 16th International Conference on Pattern Recognition, Quebec, Canada, 1:701 – 706, (2002).

[12] E. Limpert, W.A. Stahel, and M. Abbt, "Log-Normal Distributions across the Sciences: Keys and Clues", *BioScience*, Vol. **51**, No. 5, pp. 341-352 (2001).

† Interestingly, this is not necessarily the worst case strategy. As Washburn and Gulledge[9] suggest significant perseveration or repetitive selection of the same incorrect mis-matched pair, such as seen with monkeys can result in performance worse than random selection.