The Role of Perspective Cues in RSVP

Joshua Brown[1], Mark Witkowski[1], James Mardell[2], Kent Wittenburg[3] and Robert Spence[1]
[1]Department of Electrical and Electronic Engineering, Imperial College London SW7 2BT, UK;
[2]Arachnys Information Services Ltd., 69 Old Street, London EC1V 9HX, UK;
[3]Mitsubishi Electric Research Laboratories, Inc., Cambridge MA, USA;
[1]{joshua.brown214, m.witkowski, r.spence}@imperial.ac.uk; [2]james@keot.co.uk; [3]wittenbu@merl.com

Abstract - Riffling the pages of a book, perhaps in the search for a specific image, is an example of Rapid Serial Visual Presentation (RSVP). Even at a pace of 10 images per second, successful search is often possible. Interest in RSVP arises because a digital embodiment of RSVP has many applications.

There are many possible 'modes' of RSVP. However, a mode can be especially helpful if, after the appearance of an image, and without delaying the arrival of other images, it can remain in view for a second or two to allow a user to confirm that a desired image has been found. Moreover, if a collection of images is presented in such a way as to be perceived as moving in 3D space, it is thought that the search for an individual image can thereby be enhanced by comparison with a 2D presentation.

To test this conjecture we devise and use the "Deep-Flat" visual illusion whereby a column of moving images magnifying in size is perceived as approaching the viewer as in a 3D space. When the images are presented in an equivalent way horizontally as a row, the viewer tends to see this as images growing in size, but now on a flat (2D) plane. We tested comparable RSVP designs in these two illusions to ascertain the relative effects of 2D and 3D style presentation under precisely controlled conditions. Elicited data included both performance measures (e.g., recognition success), and user preferences and opinions.

We established the effectiveness of RSVP using the illusion. When tested under directly comparable conditions, we concluded that performance is not significantly affected by the illusion of depth, but that the inclusion of certain background cues can have a significantly detrimental effect on performance.

I. Introduction

The term Rapid Serial Visual Presentation (RSVP) is used to describe the rapid sequential presentation, to a human user, of a collection of images. A familiar example is the riffling of a book's pages in order to locate a known image. Even if that riffling proceeds at a rate as high as ten pages per second it is likely that a 'target image' will be successfully spotted (Potter and Levy, 1969, [1]). A significant attraction of this method of image browsing is that it is apparently pre-attentive (Healey et al, 1996, [2]) – recognition occurs within about 100 milliseconds – and is believed to involve no conscious cognitive effort (Potter, 1999, [3]).

A digital embodiment of RSVP has many applications (see Spence and Witkowski, 2013, [4]) ranging from the picture search facility on a smart phone or digital camera, to the fast-forwarding of recorded TV programmes (e.g. Figure 1, Wittenburg et al, 2003, [5]). RSVP can occur in many different modes. 'Slide-show' mode is analogous to page riffling: one image appears at a time and is rapidly replaced by the next image in the collection. One drawback of this mode, however, and which has prompted the invention of many other modes, is that each image only appears for a very short time, offering no opportunity to confirm target recognition. The use of screen space to allow an image to remain in view for a second or so without delaying the appearance of a subsequent image, characterizes many RSVP modes.

Figure 1: Video Fast-Forwarding as RSVP

With some RSVP modes the image collection is intended to be perceived by a user as existing in 2-dimensional or flat space. Other modes allow the viewer to form an impression of images in 3-dimensional (3D) space (Wittenburg et al, 2000, [6]), by exploiting visual cues such as overlap (occlusion), relative image sizes, or the magnification of moving images. There are many cues that give rise to the perception of depth, though only some are applicable to a conventional flat screen display (Gibson (1979) [8] ; Ware (2012) [9]).

If the presented images are moving, their movement can be arranged to instill in the viewer a perception of movement in 3D space. For instance, Figure 2 (left) (Wittenburg et al, 2003, [5]) shows the 'Floating RSVP' mode in which a collection of images representing available products appears to move towards, and then past a user, much as a motorway sign 'moves past' a driver. This design is explicitly

attempting to create an illusion of depth on the display screen, coordinating image size, trajectory and occlusion between images to reinforce the depth effect.

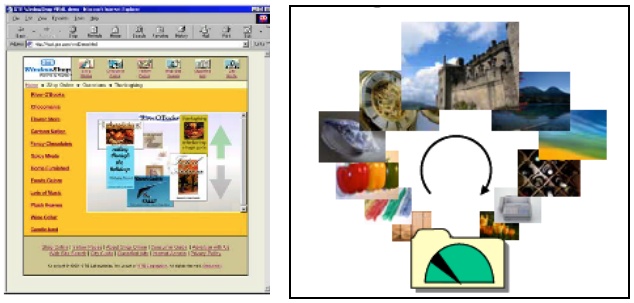

Figure 2: Floating (l) and Carousel (r) RSVP

In the fast-forwarding and rewinding recorded video application (Figure 1, Wittenburg et al, 2003, [5]) variations in image size, movement, perspective cues and overlap combine to provide an illusion of depth. Other examples include 'Carousel' mode RSVP, designed to support a query as to the content of a folder (Figure 2, right). Here, images move in a circular trajectory, appearing from one side of the folder and returning to the other. The illustrated size change and overlap are optional and were not originally intended to instill a sense of perspective in the user.

II. Related Work

Applications of RSVP began to emerge some twenty years ago (for an extensive review see Spence and Witkowski, 2013, [4]). RSVP modes can usefully be classified as static or moving, and flat or perspective. The simplest form of RSVP, and one that has been most extensively researched, is referred to as static because each image appears in the same fixed location, remains there for a fraction of a second and then disappears from view.

An early example is 'slide-show' RSVP, in which each image is simply replaced by the next at a given pace (images per second). Slide-show mode offers a trade-off between image space and presentation time and has found application to the browsing – and subsequent selection – of news stories on a small hand-held device (de Bruijn and Tong, 2003, [10]). In a separate application, the rapid sequential presentation of key frames from a video allows its story line 'gist' to be comprehended within about two seconds (Tse et al, 1998, [11]). Schoeffmann et al (2014) [12] investigated the potential advantages of a 3D view when searching image collections presented in grid formations. Corsato et al (2008) [7] compared various moving RSVP designs for both performance and user preference.

Of particular interest, and the subject of our investigation, is the potential advantage of presenting an image collection in such a way that the images appear to be moving in 3-dimensional space. For example, Wittenburg et al (2003) [5] surmised that "*presentations that rely heavily on movement of images in a 2D plane (scrolling) are going to be more demanding to process than ones that move images forward or back in a depth dimension in a virtual 3D model. The basic psychology of human visual perception tells us that humans are wired to process images in a 3D world. In particular, rapid visual processing of approaching objects is an import survival skill*". Infants as young as 8 days old show defensive reactions when objects were moved toward their faces (Bower et al 1970 [13]).

III. Comparing 2D and 3D RSVP Modes

To directly compare the effects of flat (2D) and perspective (3D) RSVP presentation we devised the Deep-Flat illusion (Figure 3) in which moving image sequences identical in duration, pace, magnification and trajectory length were presented in two ways: one rotated by 90º with respect to the other. When the images move downwards (Figure 3, left) the image stream generally appears to the viewer to be approaching in a 3D space. When the same stream is presented sideways the image stream generally appears to be at an equal depth (Figure 3, right).

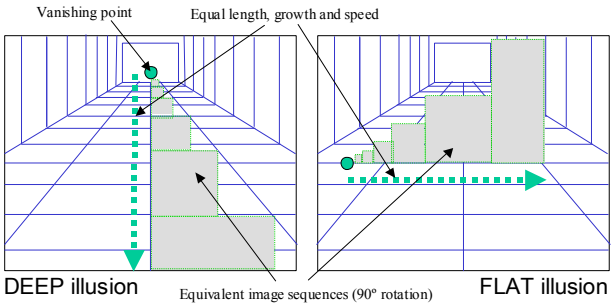

Figure 3: The Deep-Flat Illusion

To further reinforce the illusion of depth or flatness each illusion was presented both with and without a wireframe perspective background having an explicit vanishing point. In this case the vertical stream is consistent with a depth view and in the horizontal case is consistent with a flat view[1]. In all cases we specifically avoided the use of overlap (occlusion) depth cues as this would be incompatible with the horizontal or flat (2D) illusion. Additionally we created a 'control' RSVP mode that would conventionally and unambiguously be described as "flat" (no image magnification, Figure 4, Design 5).

A. Topics Investigated in This Study

As there is little published evidence regarding the relative benefits of apparent depth perception (3D) compared with flat (2D) RSVP designs we carried out an exploratory investigation. The results are intended to inform interaction designers who have chosen to use RSVP to

[1] Several participants reported that the vertical stream was longer than the horizontal (Figure 3). They were not, this is an additional illusion, independent of movement.

support image recognition tasks. We present an experiment in which two established depth cues - image size and background design - are exploited to produce directly comparable perspective depth (3D) and flat (2D) designs triggering the Deep-Flat illusion (Figure 3). We use a "category" recognition task (as in Corsato, 2008, [7]) in which participants are required to identify an image of a type of thing (e.g. "dog").

The investigation reported here addresses specific questions relating to the advantages or otherwise of the deep view over the flat. The primary question is whether the image recognition task is enhanced in one design or the other. We address questions concerning the effect of various RSVP features on user performance and preference:

a) Is overall image recognition better in the perspective (3D) or flat (2D) presentation?
b) Do users make more mistakes in perspective or flat presentations?
c) Is the illusion of depth effective?
d) Are deep or flat presentations preferable?
e) What pace of presentation is acceptable?

IV. The Five Test Designs

Figure 4, left hand column, shows how the five RSVP designs appeared to the viewer during our experiment. Designs 1 and 3 are intended to explicitly emulate the 3D RSVP view. Design 1 shows image magnification combined with the background. In this instance the magnification and background cues are intended to be cumulative, each enhancing the other. Design 3 has an identical image path to Design 1, retaining the magnification, but without the background.

Designs 2 and 4 are intended to explicitly emulate an equivalent flat 2D RSVP view for direct comparison. Design 2 incorporates an image path of identical length and magnification as in Design 1, now intended to be interpreted by the viewer as moving in a single plane across their field of view. Design 4 maintains identical path length, speed and image magnification, but without the background.

Design 5 is a control, devoid of all depth cues. The images travel from left to right at a constant size. This is intended to remove any suggestion of travel in depth. The constant image size is calculated to match the median size of the equivalent image during magnification.

To compensate for differing aspect ratios all images were fitted into notional square "bounding boxes", so the images never overlap their neighbours.

Figure 4 (right hand column) presents representative gaze "heat maps" showing the density of fixation durations for each of the five Designs (min: blue to max: red).

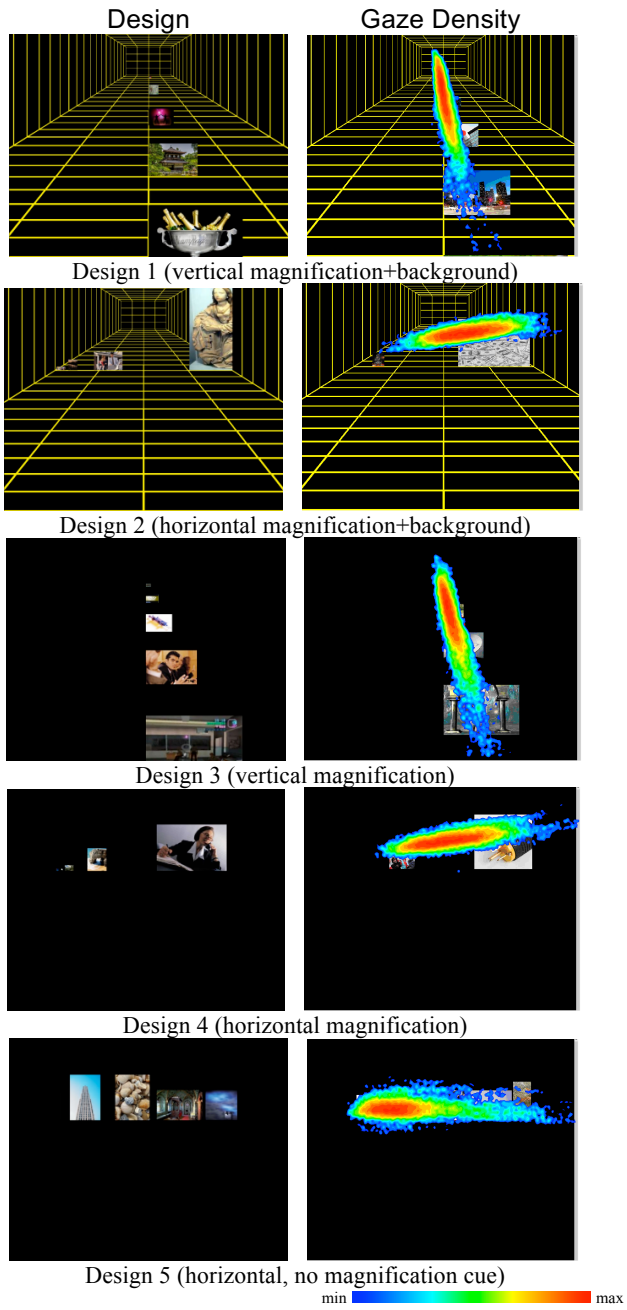

Figure 4: The Five RSVP Test Designs

V. Experiment Design

The five designs described above were employed in a category recognition task as used in Corsato et al. (2008)[2] [7] to directly compare correct recognition and false positives over equivalent conditions relating to the depth (3D RSVP) and flat (2D RSVP) views.

[2] The authors thank the team at *Dip. di Informatica e Sistemistica, Università di Pavia* for making available the image sets used in this investigation.

Volunteer participants were presented with five sequences of 500 images at a pace of either three images per second or seven images per second. Embedded within each sequence were 10 images relating to one of five target categories ("car", "cat", "dog", "plane" and "ship").

Participants were required to press the keyboard space bar whenever they identified an image that fell into the stipulated target category. Key presses were automatically recorded for later analysis. Participants were not told how many targets were embedded. Sufficient time was allowed between target appearances to avoid any potential attentional blink phenomena (Raymond et al, 1992, [14]). Otherwise target order and appearance timing within an image presentation sequence was generated at random.

No sequence, design nor category was shown to any participant twice. A random block design was used to ensure that all combinations were presented in a balanced manner across all participants. The order of sequence presentations was planned to minimise the potential consequences of any learning effects. We also recorded every participant's eye gaze behaviour while conducting the task using a Tobii T60 gaze tracker.

VI. Experimental Procedure

A total of 25 participants undertook the task, 44% were female. Ages ranged from under 19 to 89, with 48% of participants' ages falling between 30 and 49. All participants who took part reported normal or corrected to normal vision.

Each participant was welcomed, seated, and shown a brief video introduction to the experiment, describing the experiment and making it clear that participation was voluntary and that they may withdraw at any point without explanation. A list of the five possible categories was given at the start of the experiment.

Presentation of each sequence lasted either 71 or 166 seconds depending on presentation pace (7 or 3 images per second). The target category and instruction was displayed as text for 19 seconds prior to the presentation beginning ("hit the space bar when you see a picture of a <category>"). A progress bar filled to indicate the time remaining until the presentation began. Participants did not have sight of any Design prior to using it.

At the end of each sequence presentation, participants completed an on-screen questionnaire, designed to elicit aspects of their overall experience answered on a five point Likert scale. Additionally, questions about fatigue and preference between the different interfaces were asked at the end. Participants were thanked but not rewarded.

VII. Results and Analysis

Target recognition was determined by correlating participant keystroke responses with the appearance of the required category. "Correct responses" were recorded when a keystroke occurred within 2500ms of the initial appearance of the target image. The software disregarded keystrokes repeated within 50ms. This is consistent with the criteria used in previous work (e.g. Mardell, 2015 [15]), and allows for the time required for the image to grow to a usable size and resolution, visual processing within the brain, decision time and physical reaction time. We note that target images that are less obvious give rise to extended decision times. Any keystroke not preceded by a target image within 2500ms was recorded as a false positive. Any appearance of a target image not associated with a keystroke was recorded as a miss.

A. Accuracy of Target Recognition

Figure 5 shows the correct recognition rates for the different designs at the higher and lower presentation rates. Overall the recognition rate at 3 images per second is 54.2% compared to 12.4% for the faster pace of 7 images per second. This confirms that a pace of 7 images per second is too fast for this task. Only data for the results at pace 3 will be considered in the rest of this section.

Crucially, and excluding the control Design 5, the effect of image sequence direction under the directly comparable conditions (Designs 1 and 3 vs. Designs 2 and 4) was found not to be significant with the Kruskal-Wallis rank sum test ($p = 0.604$, $H = 0.269$). If this comparative result is taken to be representative of other possible designs, interaction designers need not concern themselves with the distinction between horizontal and vertical image trajectories in the context of producing flat (2D) or perspective (3D) views. However, designers are cautioned that introducing other changes, such as those in Design 5, can have a substantial effect on recognition performance.

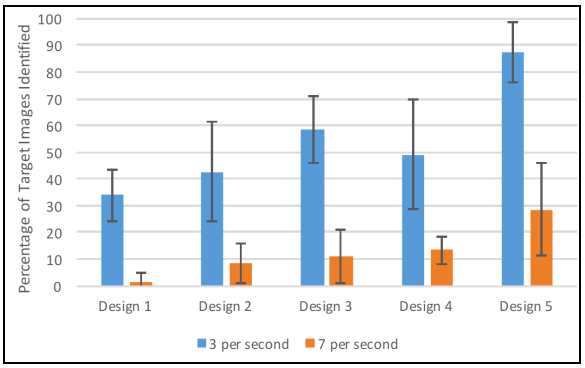

Figure 5: Correct responses by design and presentation rate

As seen from Figure 5, the average recognition without the background (Designs 3 and 4) is 58.2% and 49.1% respectively, whereas with the background included (Designs 1 and 2), 34.0% and 42.7% respectively. Again excluding the control Design 5, the effect of including the background is significant at the 5% ($p < 0.05$) level when analysed with the Kruskal-Wallis rank sum test ($p = 0.0275$, $H = 0.269$). If this is taken as representative, interaction designers should employ static perspective cues with caution. Why this should be so is unclear.

Next, considering the effect of image magnification (Designs 1 to 4) vs. the unmagnified control (Design 5) we find a significant difference in correct recognition using the Kruskal-Wallis test (p= 0.016 , H = 12.248, df = 4). We see two possible explanations for this result. First, the continuous size change between display frames gives rise to a noticeable reduction in image quality compared to simple image translation alone. Designers may want to check carefully for image quality during transitions. Second, images of a size sufficient for the task appear on screen longer in Design 5 compared to the other designs. Interaction designers should consider carefully the allocation of real estate to images of small sizes so as to avoid negative effects on task performance.

B. False Positives

A second measure of performance is the number of times a non-target image was mistaken for one that belonged in the target category. This was done by recording the number of times the space bar was pressed when there had not recently (within 2500ms response window) been a target image on screen.

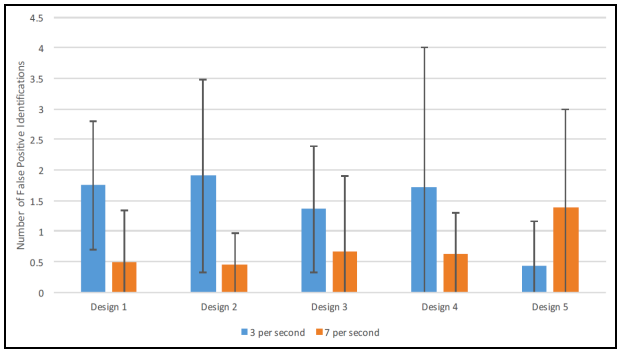

Figure 6: False Positive Responses

Firstly, we note that across all designs false positive responses are very low (mean = 1.421 per sequence presentation, SD = 1.475). One false positive corresponds to 0.002% of the image sequence.

We find design significantly affects the false positive rate using the Kruskal-Wallis test: p= 0.0621, H = 8.961, df = 4. Using the less sensitive Wilcoxon rank sum test between pairs, we note there are no particular pairwise significances. However, due to the very low overall rates, the implication is that the interaction designer need not be overly concerned with variability of false positive responses.

VIII. User Opinion Study

The second of the key aims of the study was to investigate the perceived effects of different cues available to interaction designers wishing to create the illusion that their users are viewing a scene or interface in three dimensional space and related opinions. We sought the opinions of all participants using seven questions, to be marked on a five-point Likert scale (Strongly Agree, Agree, Neutral, Disagree and Strongly Disagree). The statements to be judged were:

a) "I liked the way the images were presented"
b) "It looked like the images were moving towards me"
c) "I needed more time to look at each image"
d) "I found the task challenging"
e) "I thought the interface looked attractive"
f) "I had more time than I needed to look at each image"
g) "I felt confident when I pressed the space bar"

When presented with the "like" Statement (a), all Designs were considered near Neutral, except Design 2 which had a median response of Disagree. Statement (b) is considered in detail later. In response to Statement (c) "I needed more time" all participants either Agreed or Strongly Agreed, more Strongly Agreeing at the faster pace 7. Similarly for the corroborative Statement (f) "I had more time that I needed" all participants Disagreed or Strongly Disagreed, again more Strongly Disagreed at pace 7. The majority of participants Agreed or Strongly Agreed with Statement (d) "I found the task challenging", Design 5 slightly less than 1 to 4. Participants were generally Neutral to Statement (e) "I found the interface attractive" and Neutral to Disagreeing to Statement (g) "I felt confident when I pressed the space bar".

A. The Perception of Depth (Statement b)

In order to evaluate the sense of depth arising from the cues, we presented participants with the phrase “It looked like the images were moving towards me” (Statement 2) after each trial, and asked them to indicate their agreement on the five point Likert scale. The median and interquartile measures for each design are shown in Figure 7 (pace 3 and 7 are combined - we are only concerned with the perception of visual effect here).

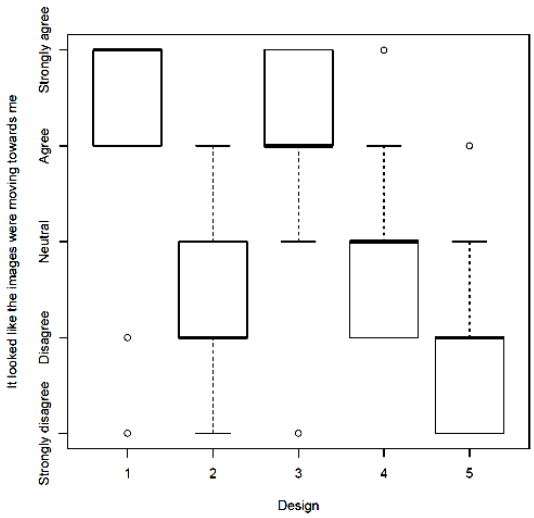

Figure 7: Perception of 3D effect

It can be seen that the combination of vertical magnification with movement and the inclusion of the background (Design 1) gives the strongest sensation of depth. Excluding the background (Design 3) reduces the perspective effect, but it is still strong. When the image sequence is presented horizontally (Design 4) the sensation of depth is reported as neutral. However, when horizontal travel is combined with the background (Design 2) - intended to reinforce the flat (2D) effect - the perception of depth is further reduced. Design 5 - intended to give no sense of depth - was most seen as flat. Using the Wilcoxon test a pairwise ANOVA indicated 7 of the 10 Design combinations were found to be statistically significant at the 5% level.

IX. Eye Gaze Behaviour

For all Designs gaze appears to exhibit a predominantly short tracking behaviour in the direction of image movement interspersed with rapid saccadic episodes in the reverse direction. We noted no unexpected gaze behaviours. However, we observed extended tracking of target and potential targets (Spence and Witkowski, 2013, [4] for a discussion of this effect).

X. Summary and Conclusions

The formal experiment described in this paper was designed to identify differences in recognition between deep (3D) and flat (2D) RSVPs based on our Deep-Flat illusion.

This was achieved by creating an experimental design with image movement trajectory and image magnification properties that were designed to give the illusion of depth when presented vertically (Design 1 and 3) and absence of depth when presented horizontally (Design 2 and 4). To enhance (or reduce) the perceived depth effect each design was also presented with a wireframe background (Designs 1 and 2). We also prepared a control design (Design 5) devoid of all depth cues for comparison. We used a category recognition task in a fully balanced block design and conducted a user opinion survey. We noted the following:

The pace of image presentation severely affects performance. A pace of 3 images per second is generally acceptable for the task investigated (~50% recognition), but 7 per second is not.

There is no evidence on the basis of this investigation that presenting the images in perspective (3D) vs. flat (2D) on a rigidly comparable basis will significantly reduce the recognition rate. This was unexpected by the authors.

However, changes to the design (e.g. removal of all magnification and background cues in control Design 5) did give rise to significant improvement in recognition performance. This might be due to differences in visual clarity inherent in the continuous magnification of moving images. This improvement might equally be explained by the fact that images of a size minimal to the task performance were on screen longer in Design 5 than in the others.

Contrary to our expectations, the inclusion of the wireframe background did have a significant, and detrimental, effect on recognition. While the mechanism is unclear, interaction designers may wish to use such supporting visual design elements with some caution. Other depth cue designs that were not explored here, e.g., incorporating a ground plane, may prove more effective.

Responses to many of the user survey topics were somewhat inconclusive, except where the higher presentation rate was used (disliked by all). However, the user survey strongly indicates that combinations of visual depth cues substantially and significantly reinforces or reduces the 3D effect. Based on this evidence, we encourage exploration of a larger set of 3D design alternatives and further research of the effects of these alternatives on task performance.

References

[1] Potter, M.C. and Levy, E.I. (1969) Recognition memory for a rapid sequence of pictures, *J. Expt. Psychology*, **81**(1):10-15.

[2] Healey, C. G., Booth, K. S., and Enns, J. T. (1996) High-speed visual estimation using preattentive processing, *ACM Transactions on Human Computer Interaction* **3**(2):107-135.

[3] Potter, M.C. (1999) Understanding sentences and scenes: the role of conceptual short-term memory. In V. Coltheart (ed) *Fleeting memories: cognition of brief visual stimuli,* Cambridge, MIT Press.

[4] Spence, R. and Witkowski, M. (2013) *Rapid Serial Visual Presentation: design for cognition*, Springer.

[5] Wittenburg, K., Forlines, C and Lanning, T. (2003) Rapid Serial Visual Presentation techniques for consumer digital video devices, *Proc. ACM Symp. on User Interface Software and Technology*, BC, Canada, November 2003, pp. 115-124.

[6] Wittenburg, K, Chiyoda, C., Heinrichs, M. and Lanning, T. (2000) Browsing through rapid-fire imaging: requirements and industry initiatives, *Proc. Electronic Imaging'2000,* pp. 48-56.

[7] Corsato, S., Mosconi, M. and Porta, M. (2008) An eye tracking approach to image search activities using RSVP display techniques, Proc. *ACM Workshop on Advanced Visual Displays*, Naples, Italy, May 2008, pp. 416-420.

[8] Gibson, J.J. (1979) *The ecological approach to visual perception*, Boston: Houghton Mifflin.

[9] Ware, C. (2012) *Information visualization: perception for design*, Morgan Kaufmann. (3rd edition).

[10] de Bruijn, O. and Tong. C.H. (2004), M-RSVP: mobile web browsing on a PDA, *People and Computers XVII.*

[11] Tse, T., Marchionini, G., Ding, W., Slaughter, L. and Komlodi, A. (1998) Dynamic key-frame presentation techniques for augmented video browsing, *ACM, Proc. Workshop on Advanced Visual Interfaces,* pp.185-194.

[12] Schoeffmann, K., Ahlström, D. and Hudelist, M.A. (2014) 3-D interfaces to improve performance of visual known-item search, *IEEE Trans. on Multimedia*, **16**(7):1942-1951.

[13] Bower, T.G.R., Broughton, J.M. and Moore, M.K. (1970) Infant responses to approaching objects: an indicator of response to distal variables, *Perception and Psychophysics*, **9**(28):193-196.

[14] Raymond, J.E., Shapiro, K.L. and Arnell, K.M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Expt. Psychology. Human Perception and Performance*, **18** (3): 849–860.

[15] Mardell, J.P. (2015) Assisting search and rescue through visual attention, Ph.D. Thesis, Imperial College London.