Do More Pictures Mean More Effort?

Investigating the Effects of Monocular Depth on Target Detection in a 3D WIMP Pictures Folder

Markos Kyritsis
Computer and Information Science Dept.
Higher Colleges of Technology
Sharjah, UAE
mkyritsis@hct.ac.ae

Stephen Gulliver[1], Eva Feredoes[2] & Winai Nadee[1]
[1]Henley Business School, Dept. of Business Informatics
[2]School of Psychology & Clinical Language Sciences
University of Reading
Reading, UK
s.r.gulliver@henley.ac.uk

***Abstract*— The limited commercial success of 3D WIMP interfaces, despite ongoing efforts, leads us to question whether depth itself is detrimental to task performance due to, e.g., an increase in the amount of clutter, or if the lack of any success can be mostly attributed to unsuitable interactivity with input devices made for 2D interfaces. In this study, we evaluate a common interactive task -browsing a pictures folder- and argue that despite an increase in the number of nontarget distractors available on the screen when introducing monocular depth, there is no decrease in target detection times, nor are there any changes in cognitive load (as measured through pupillometric data). Interestingly, eye tracking data indicates that this is not due to a lack of fixations, as participants tend to spend proportionally less time fixating on pictures in front of them as more items become available in the background. Finally, our participants made significantly more target identification errors when there were only two picture-layers of visible depth, when compared to four picture-layers. We therefore suggest adding monocular depth cues to 3D WIMP photo gallery or desktop pictures folder applications.**

Keywords- Interaction & Interface Evaluation, 3D WIMP; Visual Search; Eye Tracking

I. Introduction

The WIMP interface is undoubtedly the most essential and common method of interaction for the everyday user when it comes to human-computer interaction. Bundled with all the major operating systems, this type of interface is the first thing a newcomer would be expected to use in order to complete everyday computing tasks. Even though there exist variations across the wide range of platforms that host WIMP desktop interfaces, the actual design has remained relatively unchanged for the past 40 years, despite periodic predictions from various research teams of shifts towards alternative methods of interaction [3][4][8]. However, this is not due to a lack of interest from academia or industry, as is evident by theoretical upgrades, the most popular of which is arguably the addition of monocular depth [1][5][10][14][15][17]. Nevertheless, any attempts to include depth have been met with either commercial failure (e.g., project looking glass developed by LG3D) or have not really seen much of a success, such as the fairly recent acquisition of the Bumptop desktop (www.bumptop.com) by Google.

A. *Finding a target picture: a visual search task*

When consider the task of finding a picture in a folder, we are, in essence, conducting a visual search task, a common paradigm used for studying selective attention in the areas of cognitive psychology and neuroscience. In these types of paradigms we are interested in how long it takes for a participant to detect a target amongst nontargets in environments of varying size, as well as how many target identification errors occur. The efficiency of the task is affected by both exogenous bottom-up orientation cues such as colour, size, movement and other features, as well as top-down endogenous orientation cues guided by e.g., working memory [18]. As described thoroughly by the Feature Integration Theory (FIT) [16] concerning exogenous orientation, when the target is surrounded by homogenous nontargets, then the search is considered efficient, preattentive, and not influenced by increases in set size, instead causing a 'pop-out' effect allowing for parallel search. On the other hand, heterogenous nontargets (i.e., the target differs to some nontargets in one feature, but is similar in others) leads to a conjunction search, requiring the binding of features and hence increases in attentional resources and an inefficient serial search.

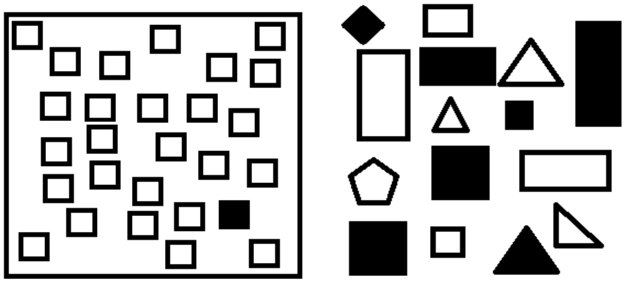


Figure 1. Left - Feature search, finding the black box is effortless and not susceptible to changes in set size; Right - Conjunction Search, the heterogeneous distractors lead to higher attentional resource requirements, making this type of search more inefficient and more susceptible to changes in set size

For complex picture targets and nontargets, as one would expect from pictures in a folder, we can assume that the search will be inefficient, and heavy on attentional resources,

ISBN: 978-1-61208-468-8

and one may therefore be tempted to simply infer that by adding more items in depth we would essentially be increasing the set size and making the search task even less efficient. On the other hand, past studies have shown that depth itself can be considered a bottom-up feature, as well as a guidance property during a visual search task [11][13]. For this reason, target detection may not suffer, as stimuli in depth could efficiently be ignored during the task through the use of a sub-nesting strategy. These strategies have been routinely observed when participants are asked to e.g., find a red ‘k’ amongst red and blue letters, which will lead to blue letters being ignored, although particularly salient stimuli (e.g., a mostly yellow picture amongst otherwise blueish pictures) may still capture attention, as shown by [9]. Regardless, there is, to the best of our knowledge, no compelling evidence to describe how this would affect target detection.

B. *Including Monocular Depth*

Studies have had mixed opinions on whether the addition of depth holds any real advantages in terms of interaction quality. There is evidence, from qualitative data, that suggests a positive attitude by users towards a desktop that includes both depth and physics [1], as well as some limited evidence that supports an increase in performance during certain tasks when depth is included in the system [15]. On the other hand, studies have found a decrease in performance in certain user tasks (such as finding a file in a folder) [6][7][9]. In all of these studies there are several merits as well as pitfalls. The study by [1], being more oriented towards the engineering of a physics-based desktop and less focused on the human cognitive limitations, essentially does not provide evidence as to whether such an interface would be indeed beneficial in terms of usability.

The studies by [6][7] and [9] compared 2D and 2.5D, to a 3D interface, not accounting for the changes in the interaction styles, and whether the 2D input device is unsuitable for this type of interaction (even if the y-axis was constrained to make interaction more simple). In other words, when looking for a target picture in a folder, by not systematically increasing the layers of visible depth, the researchers did not consider whether the increased amount of nontarget distractors was the reason it takes users longer to find the target in these environments, or whether it has more to do with interaction using a 2D mouse in 3D space.

Therefore, rather than a holistic approach (3D WIMP vs 2D WIMP), we instead argue the need to investigate the benefits of including monocular depth to each user-based task independently. As shown in [9], there was no benefit to adding depth in a folder populated with text files, however, when target stimuli were perceptually salient, target detection times decreased significantly in a pictures folder. Therefore, for the scope of this study we only considered the potential benefits of a 3D interface in browsing the pictures folder for a target picture.

II. CURRENT STUDY

In this paper, we investigate the impact of depth on target detection times and errors by developing a 3D WIMP pictures folder, and systematically increasing the number of visible layers of images from two to four. Since we were interested in seeing whether increases in visible depth, and hence set size, would lead to more items being attended during the trial as one would expect from a serial search, or whether participants effectively ignored the increased visible layers as a form of a sub-nesting strategy, we used an eye tracker to explore whether there is a relationship between depth and the number of fixations on each layer of visible depth, as well as measure any changes in pupil size, which has been shown to be a good reflection of mental effort [2].

III. METHOD

A. *Participants*

Having received ethical approval from the University of Reading School of Psychology and Clinical Language Sciences, we recruited 18 participants (15 women, 3 men, age range: 21 - 27, mean age: 24.23), to take part in our experiment. All participants had normal or corrected to normal vision, while none claimed to suffer from colour blindness or any other disorders that would impact the selective attention task. The participants were asked to sign consent forms, asked to read the information sheets, and were debriefed at the end of the experiment.

B. *Materials and Design*

The 3D pictures folder (Figs. 2 & 3) was built using Javascript and the three.js library (a popular retained mode library for 3D development) and was optimized to run well on the Google Chrome web browser. The folder was then populated with 304 pictures of 190x190 pixel resolution made up of people, groups of people, animals, and various objects. Each visible layer was made up of 16 pictures, while the groups themselves were of equal size and placed in the environment following a random uniform distribution.

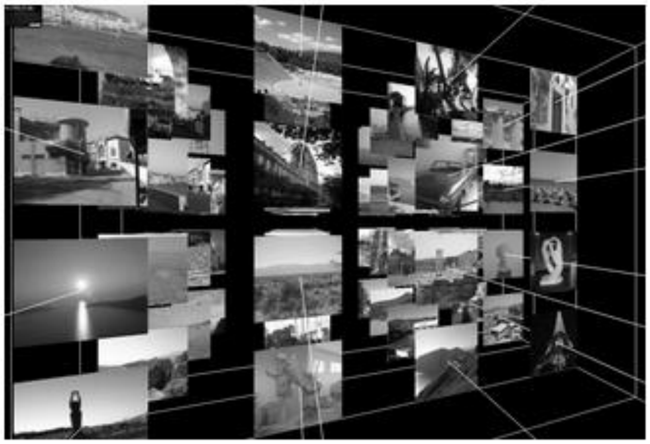


Figure 2. The 3D WIMP pictures folder with four visible layers

The space between each picture on the horizontal axis was set to be 1/10th of the total screen width (which was 192 pixels when the layer was as "close" to the screen as possible), while the space between each picture on the vertical axis was set to 1/20th of the total screen height (which was 60 pixels when the layer was as "close" to the screen as possible). The distance between the layers was set to ten default arbitrary units in 3D "depth" as set by the three.js library.

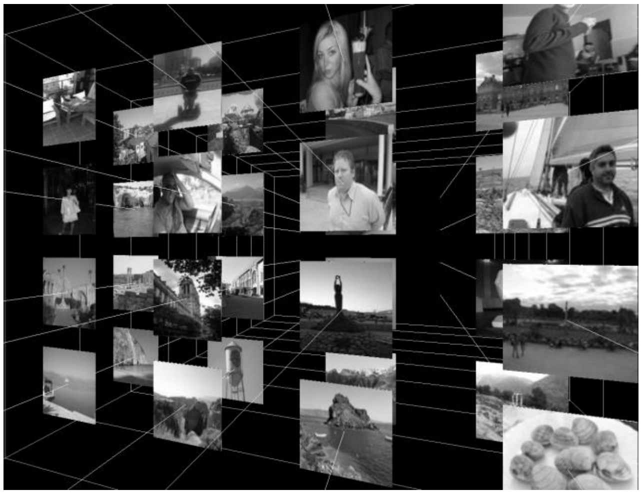


Figure 3. The 3D WIMP pictures folder with two visible layers

Since the target was selected randomly for each trial, differences in low-level (bottom-up) feature complexity and semantic differences with nontargets was not controlled, however, we expect that the random sampling of pictures from multiple categories, as well as the random selection of the target in each trial leads to a decreased likelihood of our results being affected from large differences in salience between target and neighbouring nontargets.

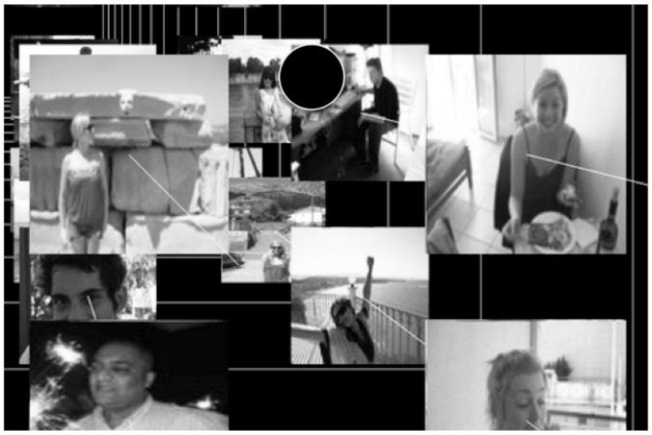


Figure 4. A ray-casting algorithm was used to measure the number of fixations on each layer

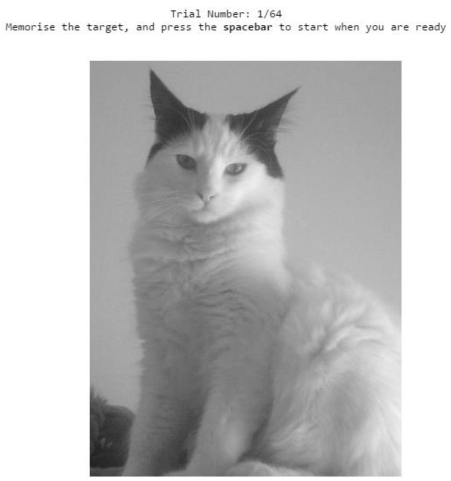
The mouse interaction purposely resembled the classic 2D WIMP interface, even though movement occurred along the z axis (in and out of the environment) using the mouse scroll wheel. This movement was essentially a translation of image on the z-axis, and no transition animations were used. Rotations along the x or y axis were not implemented, in order to decrease the overall complexity of interacting in a 3D environment using a 2D input device [12]. Rotations on the z axis were implemented, however this feature was disabled during the experimental stage in order to facilitate the overlay of eye tracking data to the environment.

An Eyelink 1000 eye-tracker (SR Research, Montreal) was used to record fixations. The chin rest was placed 70cm away from a large 28" monitor (16:10 aspect ratio), with a screen resolution of 1920x1200 pixels, while the sampling rate for the eye tracker was set to 500 samples every second. Calibration was kept at < 0.50 of error, (~ less than half the width of a human thumb at arm's length).

Our software recorded the number of intersections between eye fixations and pictures, as well as pupil size, with iterations occurring asynchronously every millisecond, with ~30ms maximum delay. The fixation measurement was implemented using a ray casting algorithm that would measure a fixation in the same way a mouse click would work when selecting a picture (Fig. 4). The target detection times were extracted from the eye tracking data once a fixation had occurred on the target that subsequently led to its selection using the mouse. Target identification errors were measured by the amount of clicks on a nontarget during each trial. The whole process has been summarised in Fig 6.

C. Procedure

After a small automated tutorial on the user interface, participants were presented with a random target in the beginning of each trial. Once they felt they were ready to begin, participants were instructed through text on the screen to press the spacebar and start the trial (Fig. 5).



Trial Number: 1/64
Memorise the target, and press the **spacebar** to start when you are ready

Figure 5. A random target would appear before the beginning of every trial. Participants were expected to maintain the target in working memory during the visual search task

Upon target detection, the participants would click on the target and proceed to the next trial, if they made an error and clicked on a nontarget, that picture would be coloured red to provide feedback to the participant that they had made a recognition error. Furthermore, if the participant felt they were unable to find the target image they could choose to skip the trial (this was logged as a failed trial). At the end of either 64 trials or one hour, the experiment would end. Furthermore, participants were allowed to take short breaks every 10 minutes, hence, the eye tracker was calibrated and validated before each block of trials. Only seven participants managed to finish all the trials, while the range overall was from 32 to 60 (mean = 54.33, $\sigma = 11.31$).

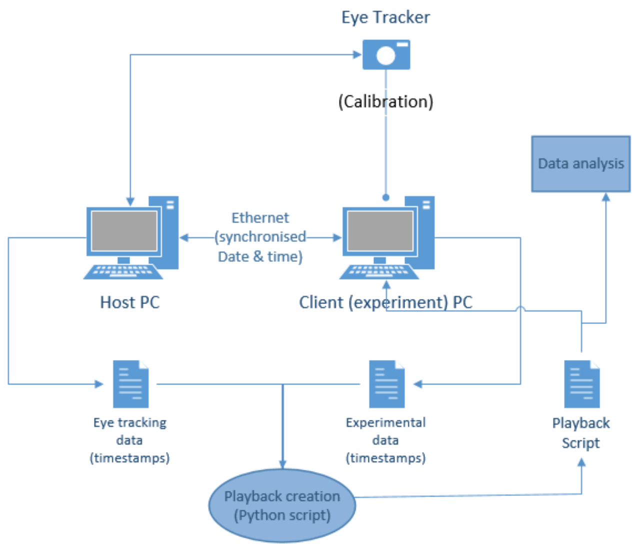


Figure 6. System Architecture illustrating the connections between the eye tracking and experimental machine, as well as the process of extracting playback information and data for the analysis

IV. RESULTS

The target detection time (RT) data did not follow the normal distribution, therefore, the non-parametric Kruskal-Wallis test was used in place of a one-way ANOVA, with visible depth as the independent categorical variable with three levels (two - four layers of visible depth), and RT as the dependent continuous variable. We failed to find any evidence to support that increasing the layers of visible depth lead to an increase in RT, since our results were not significant. However, using Kruskal-Wallis to investigate whether there was a significant main effect of depth on target identification errors; we found a significant result ($H = 6.8$, $p = 0.03$), while pairwise comparisons using Dunnett's procedure (with two layers of visible depth as the control), revealed that trials with four layers of depth produced significantly fewer errors than trials with two layers of depth ($p = 0.05$). This was not the case when comparing trials of two layers with trials of three layers of depth.

To investigate the results further, we took the ratio of fixations from our eye tracking data between the first layer of depth and all other layers. Formally this can be expressed as: $\forall_i \in \{0,\ldots,n\}.\ R_i = F_i / (F_i + O_i)$ where R is the vector of fixation ratios, F is a vector populated with number of intersections on the first layer, and O is a vector populated with number of intersections on other layers. Even though the fixation ratio data did not fit the normal distribution, we successfully transformed the data to satisfy the normality assumption by simply raising all the values to the power of two. This was then confirmed subjectively using QQplots and objectively using Shapiro-Wilk ($p > 0.05$), allowing us to use a one-way ANOVA for the analysis. The test revealed that the layers of visible depth had a significant effect on fixation ratio ($F(2, 132) = 11.63$, $p < 0.001$), while multiple comparisons using the Tukey test reported significantly lower ratio of first layer to other-layers fixations when there were three visible layers of depth compared to two visible layers of depth ($M_{diff} = -0.07$, 95% CI, [-0.13, -0.01], $p = 0.02$), as well as when there were four visible layers of depth compared to two layers of depth ($M_{diff} = -0.12$, 95% CI, [-0.18, -0.06], $p < 0.001$). We did not, however, find a significant difference in fixation ratio when comparing three and four layers (Fig. 7). Finally, average pupil size was also extracted for each level of depth and compared using Kruskal-Wallis (since the data, again, did not fit the normal distribution), however, the results were not significant, indicating no changes in pupil size as a result of increased visible depth. Finally, it is worth mentioning that we did not find an increase in target detection times and errors in relation to trial number (i.e., there was no measurable performance decrease due to fatigue).

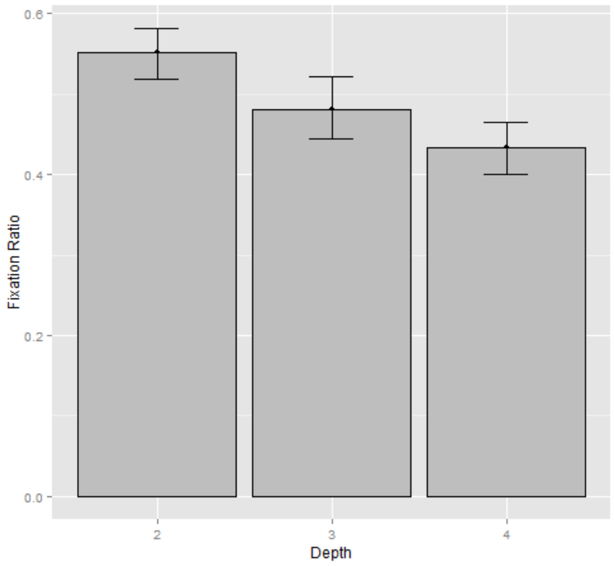


Figure 7. Ratio of fixations between the first layer of depth and all other layers, as visible leayers of depth increase.
(Error bars are 95% CI)

V. DISCUSSION

The result of our study supports that increases in the item set size (in this case, pictures) caused by adding more layers of visible depth does not impact overall target detection times. One may be tempted to assume that this is due to a top-down sub-nesting strategy, where pictures in depth are effectively ignored in order to support the very serial and inefficient search that one would expect from complex pictorial stimuli. This is not entirely true, however, as our eye tracking data revealed that by making more layers of visible depth available to participants, fixations on the first layer decreased significantly (at least when comparing three or four layers with two layers of depth).

We hypothesise that the decreased ratio of fixations on the first layer compared to other layers is either the result of (a) randomly occurring (due to the random distribution of items) highly perceptual feature contrasts in depth that capture attention (bottom-up), or (b) nontargets with features that resemble the target stimulus that may be harder to differentiate when unattended in depth (possibly because of the smaller picture size), capturing attention due to feature similarity to the target in memory (top-down). To further explore this, we conducted an exploratory meta-analysis of the data in order to see if increased depth also led to increased selection of items in depth. Much to our surprise it did not, illustrating that the relationship is very complex and warrants further investigation. Finally, increased layers of visible depth lead to decreased numbers of target identification error, but only when comparing two layers to four layers of depth.

In conclusion, contrary to previous studies, which found depth to decrease performance in a 3D WIMP, our results are more optimistic, and suggest that adding depth does not impact target detection for this particular type of user action (find a picture in a folder). However, there is undoubtedly strong evidence to support that 3D WIMP interfaces do not work well (as can be seen by a plethora of previous studies), probably due to the lack of a suitable input device that can facilitate interaction in three-dimensions. Our results do not contradict past studies, per se, but rather indicate that if a more suitable 3D input device was manufactured, then 3D WIMP picture folders and photo galleries would not lead to a degradation of performance in the task of target detection. This supports the need for further research into novel devices that can perhaps replace the 2D mouse, although past attempts have failed in this regard (e.g., the 3D mouse is hard to use long-term since it leads to fatigue). Finally, we present these results with caution, as our study focused exclusively on usability and performance, rather than user experience. Therefore, we cannot argue that users would find a 3D picture folder compelling, even if it does not lead to performance degradation. In this regard, further research using qualitative methods would be appropriate, as well as a suitable next step for this research topic.

REFERENCES

[1] A. Agarawala and R. Balakrishnan. "Keepin'it real: pushing the desktop metaphor with physics, piles and the pen." Proc. ACM conference on Human Factors in computing systems (SIGCHI 06), Apr. 2006, pp. 1283-1292.

[2] D. Alnaes, M.H. Sneve, T. Espeseth, T. Endestad, S.H. van de Pavert, and B. Laeng. “Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus.” Journal of vision, vol 14, Apr. 2014, pp. 1-1.

[3] J. Biström, A. Cogliati, and K. Rouhiainen. “Post-wimp user interface model for 3d web applications.” Helsinki University of Technology Telecommunications Software and Multimedia Laboratory, Dec. 2005

[4] D.A. Bowman, E. Kruijff, J.J. LaViola, and I. Poupyrev. “An introduction to 3-D user interface design.” Presence: Teleoperators and virtual environments, vol 10, Feb. 2001 pp. 96-108.

[5] J. Boyle, S. Leishman, and P.M. Gray. “From WIMPS to 3D: The development of AMAZE.” Journal of Visual Languages & Computing, vol 7, Sep 1996, pp. 291-319.

[6] A. Cockburn and B. McKenzie. “Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments.” Proc. ACM conference on Human factors in computing systems (SIGCHI 02), Apr 2002, pp. 203-210.

[7] A. Cockburn A. “Revisiting 2D vs 3D implications on spatial memory.” Proc. of the fifth Australian Computer Society conference on Australasian user interfaces, vol 28, Jan 2004, pp. 25-31.

[8] R.J. Jacob, A. Girouard, L.M. Hirshfield, M.S. Horn, O. Shaer, E.T. Solovey and J. Zigelbaum. “Reality-based interaction: a framework for post-WIMP interfaces.” Proc. of the ACM conference on Human factors in computing systems (SIGCHI 08), Apr. 2008, pp. 201-210.

[9] M. Kyritsis, S.R. Gulliver, S. Morar and R. Stevens. “Issues and benefits of using 3D interfaces: visual and verbal tasks.” Proc. Fifth ACM International Conference on Management of Emergent Digital EcoSystems (MEDES 13), Oct 2013, pp. 241-245.

[10] A. Leal, C.A. Wingrave and J.J. LaViola Jr. “Initial explorations into the user experience of 3D file browsing.” Proc. 23rd British Computer Society HCI Group Annual Conference on People and Computers: Celebrating People and Technology, Sep 2009, pp. 339-344.

[11] E. McSorley and J.M. Findlay. “Visual search in depth.” Vision Research, vol 41, Dec. 2001, pp. 3487-96.

[12] J.D. Mulder. “Menu Selection in Desktop Virtual Reality.” Proc. Central European Multimedia and Virtual Reality Conference, 2005, pp. 121–128.

[13] K. Nakayama and G.H. Silverman. “Serial and parallel processing of visual feature conjunctions.” Nature, vol 320, pp. 264-265.

[14] K. Nirmal and N. Mishra. “3D WIMP put into action for 3D GUI.” International Journal of Emerging Technology and Advanced Engineering, vol 3, Apr. 2013, pp. 598-605.

[15] G. Robertson, M. Czerwinski, K. Larson, D.C. Robbins, D. Thiel and M. Van Dantzich. “Data mountain: using spatial memory for document management.” Proc. 11th annual ACM symposium on User interface software and technology, Nov. 1998, pp. 153-162.

[16] A.M. Treisman and G. Gelade. “A feature-integration theory of attention.’ Cognitive psychology, vol 12, Jan. 1980, pp. 97-136.

[17] H. Wenjun. “The Three-Dimensional User Interface.” INTECH Open Access Publisher, 2008, doi: 10.5772/5910.

[18] J.M. Wolfe. “Guided search 2.0 a revised model of visual search.” Psychonomic bulletin & review, vol 1, Jun. 1994, pp. 202-38.