Evaluating automatic warning cues for visual search in vascular images

Boris van Schooten
Faculty of EEMCS
University of Twente
schooten@ewi.utwente.nl

Betsy van Dijk
Faculty of EEMCS
University of Twente
bvdijk@ewi.utwente.nl

Anton Nijholt
Faculty of EEMCS
University of Twente
anijholt@ewi.utwente.nl

Hans Reiber
Leiden University
Medical Center
j.h.c.reiber@lumc.nl

ABSTRACT

Visual search is a task that is performed in various application areas. Search can be aided by an automatic warning system, which highlights the sections that may contain targets and require the user's attention. The effect of imperfect automatic warnings on overall performance ultimately depends on the interplay between the user and the automatic warning system. While various studies exist, the different studies differ in several experimental variables including the nature of the visualisation itself. Studies in the medical area are relatively rare. We describe an experiment where users had to perform a visual search on a vascular structure, traversing a particular vessel linearly in search of possible errors made in an automatic segmentation. We find that only the case in which the warning system generates only false positives improves user time and error performance. We discuss this finding in relation to the findings of other studies.

Author Keywords

visual search, automatic warning system, magnetic resonance angiography, image segmentation

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—*Graphical user interfaces (GUI)*

INTRODUCTION

Visual search tasks are performed in various areas: finding weapons in x-rayed baggage [4], targets from a moving vehicle [15] or on aerial photographs [10, 11], cancer areas in mammograms [13, 5, 9], polyps in colonoscopy [6], or low-credibility areas in automatic medical image segmentations [8, 7]. In many cases, automatic warning systems have been devised that highlight potential targets. Such systems are imperfect: failure may be either a false positive (false alarm) or false negative (a missed item). A detection system may be tuned to produce either more false positives or false negatives. In some cases it is possible to reach 0% false positives or negatives [6, 9], in other cases it is not [13].

Submitted for review to IUI 2010.

The presence of failures in alarm systems (for both visual search tasks and other tasks) are known to cause problems for users, such as over- or under-reliance. While various studies have been made, experimental variables vary widely among different application areas: the presence or absence of a moving scene or navigation, the prevalence of false positives or negatives, whether the search is self-terminating or not (that is, the search ends when the target is found), task difficulty (examined in [10]), target rarity [4], the level of information about the system given to the users (examined in [2]), and of course the visual task itself, which can be expected to vary widely in nature. While some of these variables have been examined, most have not, and we can expect different applications to have quite different outcomes. These are too many variables to examine all at once, and the research coverage remains as yet spotty. Examining different application areas is still a very meaningful exercise.

We examine a new application involving vascular image analysis, more specifically, 3D magnetic resonance angiography (MRA) segmentation, as performed routinely by radiologists. Vascular segmentation involves determining the thickness of the inside of the vessel (the lumen), which enables analysis of possible pathological narrowings or widenings. While a vessel is tortuous, it can basically be navigated linearly (from one end to the other), as can for example the colon in colonoscopy. So, the task can be characterised as relatively easy, non-self-terminating, involving simple navigation, with users given information about presence of false positives or negatives. We examine in particular the effect of the presence of false positives versus false negatives.

RELATED WORK

Studies of generic self-terminating target finding tasks with target highlighting found that imperfect highlighting often increased rather than decreased overall user response time, due to suboptimal increase in response time for the cases where the wrong target was highlighted [3, 12]. For some non-self-terminating tasks, users were also found to spend more time double-checking the data in case of false positives, resulting in increased response times in the presence of warnings [1].

Wickens et al. [14] found that distinction of visual elements by highlighting helps focussed attention (attention to one target) but hinders integrative attention (where all targets need to be interpreted in an integrated way). Another detrimental effect is called *attention tunneling*, which means the highlights distract the user from seeing other elements in the scene. Yeh et al. [15] found that, even if highlighting of one target served to predict with 100% accuracy a target in the vicinity rather than the highlighted target itself, performance worsened.

Studies on the reliance (or trust) of users on (visual and non-visual) automatic warnings as related to the failure rate of the warning system has been studied fairly extensively. One common finding is that false positives are more damaging to trust and hence performance than false negatives [10]. Maltz et al. [10] also finds that target cueing works best if the targets are otherwise very difficult to detect.

None of these studies were conducted in the medical domain. One of the rare medical studies in this area, done by Freer et al. [5], seems to contradict some of these findings. It indicates a positive effect on clinical outcome in a mammogram-reading study with as much as 97.4% false positives. Freer et al. use a double-reading scheme, taken from medical practice, but used by none of the other studies: each mammogram is first examined as a plain image, before the warning highlights are shown, reducing any possible effect of attention tunneling. Additionally Freer et al.'s task is difficult (experts miss 50% or more of targets), unlike most of the other experiments.

This shows that studies in the medical domain may have different outcomes due to differences in experimental variables, which are implicitly assumed in the other domains. This makes it worthwhile to study other medical tasks more closely.

EXPERIMENTAL DESIGN

Our task consists of checking the correctness of automatic segmentations of vessels in MRA scans. A typical segmentation algorithm determines a vessel's location by drawing a line through the (density) center of the vessel, called the centerline. Then, it determines the thickness of the inside of the vessel (the lumen) based on the centerline.

We used a software phantom approach. The MRA data is artificially generated, along with segmentations with artificially generated segmentation errors. This way it is easy to generate dozens of cases with a clear distinction between correct and erroneous, an unambiguous ground truth, and similar difficulty levels. A vessel is constructed using a sum of sine waves. Three distractors vessels were added in each phantom. Thickness of the vessel was varied in a stylized manner with thinner and thicker areas. When looking at a cross-section, density in the center of the vessel was highest, gradually lowering towards the boundaries of the vessel, and zero outside of the vessel. No noise or other distractors were added, neither were bifurcations present. See figure 1.

Errors are simply defined as a deviation between the segmentation and the densest parts of the volume. Only three error types exist: a veering away of the centerline and segmentation from the vessel, the segmentation being thinner than the vessel, and the segmentation being thicker.

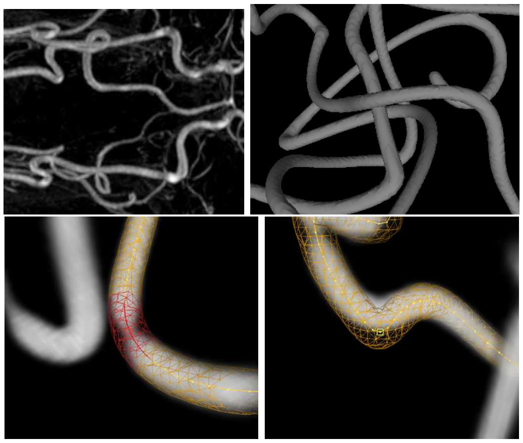

Figure 1. Illustration of the visual stimuli used
Top left: real-life data. Top right: typical software phantom as used in our experiment. Bottom: stimuli as presented to the users. Bottom left: with thickness error in the center and marked as potential error. Bottom right: with veering error in the center but not marked.

We use direct volume rendering (DVR) to visualise the volume data, with a yellow line indicating the centerline, and a brown mesh indicating the segmentation. The warning system highlights parts of the centerline and mesh in red to indicate possible errors.

We chose controls to be as simple as possible without sacrificing user control. Control is with the mouse only. One major choice we made is to base navigation on the centerline. The camera is always centered around a point on the centerline, and rotates so that the vessel is viewed from the side. The centerline is navigated by rolling the mouse wheel, or by clicking on a centerline point with the middle mouse button (MMB). The user can specify relative rotation using a two-axis valuator scheme controlled with the right mouse button (RMB). The camera is zoomed in close to the vessel so details can be seen clearly. The user can simply click on a section of the vessel with the left mouse button (LMB) to indicate a segmentation error. The appropriate section is highlighted in green.

We compare user performance (time taken and error rate) for the following four conditions:

1. NONE - no suspicious areas (baseline)
2. PAR (paranoid suspicious areas) yields only false positives - the user only has to search within the suspicious areas
3. CON (conservative suspicious areas) yields only false negatives - the user can simply click the suspicious areas but has to search the rest for missed errors
4. PER (perfect detection) - while not realistic, this indicates an upper limit to performance of suspicious areas. It is basically an interaction task rather than an interpretation task.

Note that it is not easy to compare a false positives condition with a false negatives one in absolute terms, because the situation is asymmetric. What we can do is compare if either are faster than NONE. We chose conditions to have 6-8 errors with 1-2 false positives or negatives.

We used a within-subjects design. All users received the same set of software phantoms in the same order, but with different, randomly ordered and counterbalanced, conditions. The users had to complete 6 trials per condition, totaling 24 trials. Total duration of the main experiment was 10-20 minutes. A short subjective survey was conducted at the end. We explicitly asked the users whether they actually used the suspicious areas, and which type they preferred. The survey questions we asked are the following:

usedsuspar (did you use the paranoid-mode suspicious areas to find errors?) {4:All the time, 3:some of the time, 2:learned to ignore them during the session, 1:ignored them}
usedsuscon (did you use the conservative-mode suspicious areas to find errors?) {4:All the time, 3:some of the time, 2:learned to ignore them during the session, 1:ignored them}
suspar (5-point scale, from strongly prefer PAR (5) to strongly prefer NONE (1))
suscon (5-point scale, from strongly prefer CON (5) to strongly prefer NONE (1))
susparcon (5-point scale, from strongly prefer PAR (5) to strongly prefer CON (1))

Because we used somewhat stylised models, medical laypersons could easily do the task. Since our research concerns usability involving novel interaction techniques, we asked experts on user interfaces rather than medical experts to perform our experiment. We recruited 8 subjects from the Human Media Interaction department of our CS faculty. They were not paid. They had already done a similar experiment several days earlier, involving DVR visualisation, with and without suspicious area highlighting, along with two other visualisations. This meant they already had experience with the visualisation and controls. Training for this experiment consisted of a 4-minute interactive tutorial, explaining the difference between the four conditions. Users were not told how many false positives or negatives they could expect. The sessions were conducted in a quiet room, with the users seated at a distance of about 70 cm from the 24” display. An experimenter was seated behind them.

RESULTS

We shall begin with time performance. We expect time performance effects to be multiplicative rather than additive, so we transformed the data using the log transform. We used a second transformation to increase statistical sensitivity. It is based on the fact that the sequence of software phantoms used for the trials was the same for all users. We divided the time for each trial by the overall average of that trial over all users (note that all conditions occurred equally often for each trial in the sequence). This has the effect of normalising for variations in trial difficulty.

Though PER is meant as a baseline condition, we first used repeated-measures ANOVA with Sidak posthoc analysis over all conditions including PER. The ANOVA yields $F(3, 21) = 214.052, p < 0.0005$. PER is, as we might expect, very significantly different from the others: $p < 0.0005$. It is almost twice as fast as the NONE condition, which shows that there may be quite a lot to gain from suspicious areas. We disregard it from here on.

We performed a second repeated-measures ANOVA on the remaining three conditions, which yields $F(2, 14) = 5.172, p = 0.021$. A Sidak post-hoc analysis reveals that PAR is significantly faster than NONE ($p = 0.038$). The other comparisons (NONE-CON, CON-PAR) are not significant ($p >= 0.391$). This shows that PAR does provide benefit. Mean performance over all users is given below.

condition:	NONE	CON	PAR	PER
mean trial performance:	36.8	35.0	34.0	19.7

We analysed error rate by means of a χ^2 table, assuming that trials are independent events. User errors (mistakes) were very rare events, with a total of 17 mistakes, which makes them difficult to analyse. We found that three mistakes resulted from a cognitive slip, as admitted by the user in question. These involved a confusion of colour coding (red was confused with green), resulting in 2 false positives and 1 false negative in a particular short section of vessel under the PAR condition. These were the only false positives in the dataset.

We classify trials into two classes: trials with one or more mistakes and trials without mistakes. See the table below.

cond.	*total trials*	*total segm. errors*	*total nr. mistakes*	*total nr. trials w/ mistakes*
none	48	336	6	6
con	48	336	7	5
par	48	336	1 (+3)	1 (+1)

It appears that the PAR condition might result in fewer mistakes, but the values are a bit low for a χ^2 analysis. If we include the cognitive slip, a chi-square analysis on trials with mistakes v trials without mistakes results in $\chi^2(2, N = 144) = 2.198, p = 0.333$. If we consider deletion of the cognitive slip valid, the same analysis results in $\chi^2(2, N = 144) = 3.818, p = 0.148$, and one on total number of user mistakes v total number of correctly selected segmentation errors yields $\chi^2(2, N = 1008) = 4.491, p = 0.106$. While we cannot say that PAR produces significantly less mistakes than the other conditions, it appears at least that CON and PAR do not seem to result in *more* mistakes than NONE.

For a summary of the subjective survey results, see table 1. The sample is a bit small for serious statistical analysis, but it is clear that all users used the suspicious areas, and mostly preferred them. We can at least conclude that users did not find the suspicious area marking annoying. There was little difference in preference between PAR and CON, although PAR was preferred more often than CON, and most users would prefer it over CON as well. However, a larger sample

variable	*nr. users:*	*1*	*2*	*3*	*4*	*5*	*average*
usedsuspar		-	-	3	5		3.62
usedsuscon		-	-	1	7		3.88
suspar		-	-	-	4	4	4.50
suscon		-	-	2	1	5	4.38
susparcon		-	2	-	3	3	3.88

Table 1. Subjective variable statistics

Number of users who selected each item on each survey scale, and the average value.

would be required to test if there is a significant difference here.

CONCLUSIONS

We conducted an experiment involving the manual verification of automatic segmentations of MRA images, with help of an imperfect automatic warning system that highlights possible errors in the segmentation. We compared user time and error performance as well as subjective preference for the following conditions: no warning highlights, only false positives (paranoid), only false negatives (conservative), and perfect highlighting.

We found that users perform significantly faster with paranoid highlighting than with no highlighting, and they make insignificantly less errors. There were no other significant differences. Users also prefer suspicious areas over no suspicious areas, and appear to prefer paranoid over conservative highlighting.

This contradicts most previous findings, which generally indicate that especially paranoid highlighting is often detrimental. Our contradictory result cannot be explained by high difficulty or low target prevalence (the task was easy, as is illustrated by the low error rate). While false positive rate was fairly low (about 20%), other experiments demonstrated a detrimental effect for similar rates [15, 1]. The difference may be explained by the level of information given to the users (they were told whether to expect false positives or negatives), possibly combined with other factors, such as low false positive rate. It appears our results more closely follow a rationally based cognitive model: for the false positives case users will have to search only the marked areas, and hence, search space is reduced, in contrast to the false negatives case, where it is not. Alternatively, the difference in outcome may be explained by a difference in visual stimuli. We argue that further experiments will be necessary to more thoroughly cover this research area.

ACKNOWLEDGEMENTS

This work was carried out under the NWO (Netherlands Organization for Scientific research) Multivis project (N 643.100.602), which is part of the NWO VIEW program.

REFERENCES

1. S. R. Dixon, C. D. Wickens, and M. J. S. On the independence of compliance and reliance: are automation false alarms worse than misses? *Human factors*, 49(4):564–72, 2007.
2. N. Ezer, A. D. Fisk, and W. A. Rogers. Age-related differences in reliance behavior attributable to costs within a human-decision aid system export. *Human Factors*, 50(6):853–863, 2008.
3. D. L. Fisher and K. C. Tan. Visual displays: The highlighting paradox. *Human Factors*, 31(1):17–30, 1989.
4. M. S. Fleck and S. R. Mitroff. Rare targets are rarely missed in correctable search. *Psychological Science*, 18(11):943–947, 2007.
5. T. W. Freer and J. M. Ulissey. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology*, 220:781–786, 2001.
6. W. Hong, F. Qiu, and A. kaufman. A pipeline for computer aided polyp detection. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):861–868, 2006.
7. K. Levinski, A. Sourin, and V. Zagorodnov. 3D visualization and segmentation of brain MRI data. In *GRAPP 2009*, pages 111–118, 2009.
8. J. H. Levy, R. R. Broadhurst, S. Ray, E. L. Chaney, and S. M. Pizer. Signaling local non-credibility in an automatic segmentation pipeline. In *Proceedings of the International Society for Optical Engineering meetings on Medical Imaging, Volume 6512*, 2007.
9. F. J. López-Aligué, I. Acevedo-Sotoca, A. García-Manso, C. J. García-Orellana, and R. Gallardo-Caballero. Microcalcifications detection in digital mammograms. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2004), vol.3*, 2004.
10. M. Maltz and D. Shinar. New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, 45(2):281–295, 2003.
11. S. Rice. Examining single and multiple-process theories of trust in automation. *Journal of General Psychology*, 136(3):303–319, 2009.
12. F. P. Tamborello and M. D. Byrne. Adaptive but non-optimal visual search behavior with highlighted displays. *Cognitive Systems Research*, 8(3):182–191, 2007. Cognitive Modeling.
13. Y. Wang, X. Gao, and J. Li. A feature analysis approach to mass detection in mammography based on rf-svm. In *ICIP 07*, pages 9–12, 2007.
14. C. D. Wickens and A. D. Andre. Proximity compatibility and information display: Effects of color, space, and objectness on information integration. *Human Factors*, 32(1):61–77, 1990.
15. M. Yeh and C. D. Wickens. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3):355–365, 2001.