The 13rd International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST
(2012)
D. Arnold, J. Kaminski, F. Niccolucci and A. Stork (Editors)

PileBars: Scalable Dynamic Thumbnail Bars APPENDIX: User study evaluation

Paolo Brivio[1], Marco Tarini[1,2], Federico Ponchio[2], Paolo Cignoni[2], Roberto Scopigno[2]

[1]Università degli Studi dell'Insubria, Varese, Italy
[2] Istituto di Scienza e Tecnologie dell'Informazione (ISTI), CNR, Pisa, Italy

This short doc contains the appendix to the VAST 2012 submission entitled "PileBars: Scalable Dynamic Thumbnail Bars".

1. Introduction

NOTE: this user study is left here for the sake of the anonymous reviewers, but will be removed from the paper (in case of acceptation) and published as a complementary pdf doc on the EG DigLib (additional material linked to the paper).

In this appendix, we present some quantitative and qualitative comparisons among the PileBar interface and the ones of a couple of conventional image-browsers. In our tests, we aim to measure how fast users navigate thousands of images with image-browsers covering only a small part of the screen. In particular, the PileBar browser is compared with a horizontal thumbnail-bar and a browser with a grid layout.

2. Experiments

2.1. Setup

In choosing the most appropriate PileBar contenders, we considered a number of alternatives, selected among several free tools available online. As the browsing interface of most of those browsers was out of our control and as we could not resize them properly, we finally resorted on the FastStone [FS04] thumbnail-bar, and on the Microsoft Windows Explorer grid layout. In particular, we adapted the PileBar, the Explorer, and the FastStone windows to the same $1600x235$ pixels working area, thus covering but a fraction of the whole $1600x900$ screen, which elsewhere was filled with a blank background. All experiments were performed on a laptop with a $17inch$ screen, a $2.5GHz$ dual processor, and $3GB$ of RAM.

All interfaces support browsing through the mouse-wheel. However, the PileBar is the only one implementing a dynamic layout, enabling thumbnail selection, dragging, and preview, while the other browsers interactions are based on scrollbars.

Participants A set of 16 computer science students and young researchers volunteered to participate in two tests. However, due to the long time required to complete the first one, only 10 of them accepted to participate also in it. All had a normal or corrected to normal vision with no color blindness, and they were new both to the PileBar and to the FastStone interface.

Procedure Experiments took place under the same lighting conditions in a silent room. Each participant was allowed a preliminary 5 minutes test-run on each browser, using a different training dataset. During these test-runs, users were carefully instructed using one sheet with illustrated instructions about the complete set of functionalities of either tool.

Then, we asked each user to perform the two experiments, each constituted of a sequence of tasks on a specific dataset. Before performing each task, each browser was restored to its initial configuration: the Explorer and the FastStone browsers figured the thumbnail representing the first image of the dataset, while the PileBar browser focused at the middle of it. Tasks were described in written assignments, and users were totally unassisted while performing them. Timings were taken after each task was read and understood and until the user selected the target image.

2.2. Experiment one

First, we evaluated how much the considered browsers are efficacious for browsing an image dataset in which an explicit total ordering is naturally imposed (i.e. an arbitrary ordering would make the browsing cumbersome).

Dataset The dataset has been generated procedurally: given the set of integers from 1 to $28,000$, 20 disjoint subsets of random size (253 on average) have been randomly extracted. Then, every subset number has been associated an image featuring that number. Finally, the test dataset has been filled with all of these $5,061$ images. Numbers were written in white color on a dark background with the same font and size. Inside the dataset they were ordered from the smallest to the largest and this ordering was respected in each browser.

Task Users were asked to locate, in a random order, each of the first 28 multiples of $1,000$, or, if not present, its two nearest numbers. To further minimize the influence of user knowledge of the dataset on his/her performance, half of the tasks was performed with PileBar first, whereas the other half with Explorer first. In either case, all tasks were performed with FastStone last.

Results and discussion Compared results are shown in Tab. 1. We used the *R* system to compute statistics on the timing data. For each independent variable (i.e. the browser adopted during the experiment), we considered two dependent variables: the time to complete each task, and the order of browser utilization (the latter does not apply for FastStone, of course). All dependent variables resulted normally distributed with respect to each independent variable with Shapiro-Wilk normality test.

A t-Student two-tailed paired test shows that the measured differences of the user performances between PileBar and Explorer and between PileBar and FastStone is significant at least at level $p < 0.01$ (p-value $= 0.0036$ and p-value $= 9.66e-08$, respectively). On average, while using PileBar, participants took 8.5 seconds (s.d. 5.8) to complete, while 10.3 seconds (s.d. 3.2) with Explorer, and 13.1 seconds (s.d. 5.0) with FastStone.

Considering user performances when using Explorer, we notice that there is no significant difference when using Explorer first or second for $p < 0.05$ (p-value $= 0.8817$). On the other hand, the difference between using PileBar first and PileBar second is significant even for $p < 0.001$ (p-value $= 0.0008$). Participants were slower when using PileBar first, taking a mean time of 9.78 seconds (s.d. 1.78), compared to 7.19 seconds (s.d. 0.98).

When all tasks were completed, each participant was asked to score the three tools from 1 (minimum) to 10 (maximum), answering the following questions:

1. how much did you feel comfortable with each tool?
2. how much did you think each tool was helpful for these tasks?

The average scores are shown in Tab. 2. Considering Q1, with a t-Student test no significant difference can be observed between PileBar and Explorer for $p < 0.05$ (p-value $= 0.7670$). Their average scores are similar, too. However, the differences between PileBar and FastStone are very significant (p-value $= 0.0001$). In case of Q2, the measured differences between PileBar and Explorer, and between PileBar and FastStone are significant for $p < 0.001$ (p-value is $9.0e-4$, and $4.29e-9$, respectively).

Table 1: *Experimenting with an explicitly ordered dataset. For each of the* 10 *participants, timings were recorded to complete each of the* 28 *tasks, performed with PileBar, Explorer, and FastStone image browsers. Here, each cell contains the value of the ratio between the time to perform tasks with PileBar over Explorer (first column), and with PileBar over FastStone (second column).*

User	**PileBar/ Explorer**	**PileBar/ FastStone**
User 1	66.9%	60.7%
User 2	69.0%	62.6%
User 3	73.8%	73.8%
User 4	93.8%	74.9%
User 5	98.8%	67.9%
User 6	83.4%	62.8%
User 7	94.3%	70.6%
User 8	85.2%	56.1%
User 9	77.5%	56.5%
User 10	94.6%	65.0%
Average	83.7%	65.1%

Table 2: *User scores for the browsers. After experiment one, users were asked for quantitative evaluation of the browsers. The table reports their average scores (*1 *is minimum,* 10 *is maximum).*

Question	**PileBar**	**Explorer**	**FastStone**
Q1	7.1	6.9	2.9
Q2	8.3	5.7	2.4

Finally, participants were asked for qualitative additional comments about the tools. Three of them reported that they found confusing the PileBar arbitrary image ordering on the vertical direction, while all agreed that the thumbnail-bar of the FastStone inteface was the less effective tool for browsing large image datasets. In addition, most participants argued that, with more training with the PileBar interface, they would have probably performed better with it. This is also partially confirmed by the above questionnaire scores, as users stated that they felt equally comfortable with the PileBar and the Explorer interfaces, but they thought that the former has a higher potential for helping people in locating images.

2.3. Experiment two

After the first experiment, we investigated users performances on a dataset with an explicit image-clustering function defined, but without an explicit semantically significant order. Given the results from the former experiment, we chose to compare only PileBar and Explorer image browsers.

Dataset Images have been retrieved by Google searching for 22 species of different domestic animals and their various races, discarding outliers. The resulting images were joint in a sequence of $3,784$ animals grouped by species and race. In other words, Rottweilers were separated from Chihuahuas, but no cat could appear among dogs. Note that in this dataset tags have no semantic order (i.e. there is no cue to predict if cats come before or after dogs). Image tagging was used by PileBar to cluster images, while the other two browsers showed the plain sequence of animals. Furthermore, for each cluster of images in the PileBar browser, it was computed an image ordering based on image color distribution and color spatial layout.

Task We differentiated between two types of task: *to locate a species of animal* (T1. and T2.), and *to locate a specific animal of a species* (T3. and T4.). The tasks were:

1. to locate a turtle (11 images in total);
2. to locate a Dalmatian (11 images in total);
3. to locate a red Canary pictured on a uniform white background (4 items are present);
4. to locate a cat pictured on a red background (4 items are present);

To prevent the task execution ordering to significantly influence the results, half of the participants performed T1. and T3. with PileBar first, and T2. and T4. with Explorer first. The other half, instead, did the opposite.

Results and discussion The timings to complete each task are summarized in Tab. 3. Notice that in this case data distributions are not normal. Thus, we analyze them with a paired Wilcoxon signed rank test.

Considering each browser separately, there is no significant difference – for $p < 0.05$ – in performing the first two tasks (p-value$_{PB}$ $= 0.4164$, p-value$_{EX}$ $= 0.06084$), nor in performing the last two (p-value$_{PB}$ $= 0.1533$, p-value$_{EX}$ $= 0.5134$). This confirms that the choice of T1. and T2. as exemplars of the first type of task, and of T3. and T4. as exemplars of the second type of task did not influence the results.

Comparing the timings measured with the two browsers, it results that the difference between PileBar and Explorer is significant even for $p < 0.001$ both for the first two tasks (p-value $= 8.315e-7$) and for the last two (p-value $= 5.207e-6$). Overall, the direction of the difference is always clear, as with PileBar the recorded timings are from 2 to 20 times faster than with Explorer.

Table 3: *Results of experiment two. This table reports the average timings (in seconds) to complete each of the four tasks listed above, using PileBar and Explorer image browsers. Significance levels are computed with a Wilcoxon test.*

Task	PileBar		Explorer		p-value
	Time	s.d.	Time	s.d.	
T1	2.4	0.8	49.1	34.8	4.814e-4
T2	2.8	1.3	25.6	14.2	4.803e-4
T3	8.0	5.6	19.1	12.5	2.913e-3
T4	4.8	3.1	15.3	7.6	4.782e-4

3. Concluding remarks

Browsing thousands of images with conventional image browsers has proven to be time consuming. In our experiments, the PileBar interface has been generally appreciated by the users, as it allowed them to locate arbitrary images faster than its contenders and without linearly scanning the whole dataset. Our collected data clearly confirms the improvements that the PileBar novel image arrangements were meant to bring out.

In this study, we did not investigate the usefulness of every PileBar feature. A more extensive user study could evaluate in which measure user performances vary when they are let to select among a pool of image orderings and clusterings. Also, it would be sensible to conduct another study tailored on the graphical settings of the PileBar interface, like the number of piles each column should have, the distance (in screen pixels) among piles, the clustering trend across columns, and the number of piles with exactly one thumbnail. Its results could be used to further optimize the interface design of any application adopting a PileBar.

References

[FS04] FASTSTONE-SOFT: Faststone. http://www.faststone.org/, 2004. 1