Predicting Perceived Screen Clutter by Feature Congestion

Chris Lafleur, Bernard Rummel

User Experience Research, Methods and Communication, SAP AG

Abstract

Adding functionality to a computer user interface frequently results in adding more visual features to the screen. The question of when, exactly, a screen becomes too crowded is often answered arbitrarily, or via expensive usability testing. The present study investigates an algorithm presented by Rosenholtz et al. to calculate "Feature Congestion" as a measure of visual clutter on the screen, and whether or not it can accurately predict users' experiencing a screen as being cluttered (2005, 2007a). Screenshots of websites which were equidistant in Feature Congestion were subject to a psychophysical paired comparison scaling experiment with 29 participants. Results from the experiment indicate that the Feature Congestion model can accurately predict perceptions of visual clutter on a computer user interface. Practical implications of the present findings are also discussed.

1 Introduction: Can Clutter Be Measured?

User interface designers and usability engineers frequently struggle with a common tendency to continually expand or add new features to a computer user interface, known in the industry as 'feature creep' (Elliott, 2008). While designers strive for an orderly and simple appearance of screens, adding features to the user interface to enrich its functionality will invariably lead to increasingly complex and cluttered screens.

Because our short-term memories are limited to holding about 4 items, overly complex and cluttered user interfaces likely constitute a problem for computer application users (Cowan, 2001). This is because a user's short-term memory is affected by the visual information load and the number of objects on a screen, which ultimately leads to a decrease in search performance and an overall degradation of usability (Alvarez & Cavanagh, 2004).

As developers are pushed to add an increasing number of on-screen features, the user is forced to search through more clutter, making it increasingly more difficult and time consuming to complete a required task with the application. When trying to find the right balance between an application's functionality and ease of use, design teams need decision criteria about costs and benefits of the various design solutions also in terms of usability.

Usability testing, which is certainly the most valid way to assess ease of use, is often unfeasible because of budget and time constraints. To focus effort and expenses, a screening tool for automatically detecting overloaded and needlessly complex user interfaces would be highly beneficial. A design team could easily identify critical parts of a user interface for further investigation and design work. Such a screening tool would have to provide a measure of screen complexity that can be derived automatically from any screen, while still correlating reasonably well to the subjective experience of real users.

As Rosenholtz, Li, Mansfield, and Jin point out, clutter can be defined as a state where there is an excess of items in a given space, where their representation and/or organization lead to degradation in task performance (2005). This model of visual clutter which they call "feature congestion," provides a potential framework to evaluate usability independent of usability testing. To characterize a state of clutter, Rosenholtz et al. introduce the concept of a feature space, where the various visual features present in a display create a vector space (2005). Each element in the display occupies a point in this space, whose coordinates are given by the respective values of the features exposed. All display elements together form a cloud of points with a given distribution (see Figure 1). Targets in the display are easy to spot if they are at the fringe of this cloud, or more technically, its covariance ellipsoid. Targets closer to the centroid of the distribution are hard or impossible to spot. Thus, *Feature Congestion* occurs when the covariance ellipsoid in the feature space of an image becomes so voluminous that there is basically no room left to place a salient target.

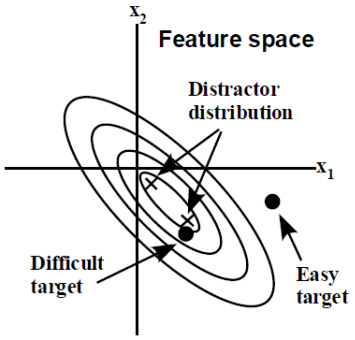

Figure 1: Example feature space of an image (Rosenholz et al., 2005).
The triangle located outside the ellipsoid is comparatively more salient than the square located within the ellipsoid.

The Feature Congestion model of clutter predicts an observer's perception of a display as cluttered whenever there is a high level of Feature Congestion in the display. The Feature Congestion measure indicates how hard it is to add a new, salient item. While Feature Congestion can occur locally at any given point in an image, it is possible to pool local Feature Congestion over the entire display in order to assess its overall level of clutter.

In a subsequent study, Rosenholtz, Li, and Nakano (2007a) implemented the Feature Congestion measure in a computer algorithm. The algorithm operates in 4 stages: 1) compute local feature (co)variance of color, contrast, and orientation for three scales; 2) combine

across scale; 3) combine clutter across feature types; and 4) pool over space to get a single measure of clutter for each input image. The algorithm was validated in a series of search experiments involving maps and experimental "cluttered desk" displays, where it performed well at predicting mean search time, mean contrast thresholds, and subjective ranking of stimuli. This leads to the question of whether the Feature Congestion model will be an effective evaluator of clutter in computer user interfaces. That is the purpose of the present study.

One useful component of the Feature Congestion model is that the maps used by Rosenholtz et al. (2005) provide an excellent means of testing clutter due to the high information density typically present in the map's feature space. However, they differ from computer user interfaces in a number of ways. For example, in computer applications, text is used extensively for labeling purposes and information display. There are also structural conventions such as navigation areas, banners, toolbars etc., which are more or less universally followed. In any case, strict horizontal and vertical alignment of UI elements is considered "good style" and therefore a ubiquitous feature in user interface displays, which is unlikely to be present on any topographical map. At present, we do not know how well the current implementation of Rosenholtz et al.'s Feature Congestion model will evaluate the feature space of a computer user interface, given the increased structure and guidelines that are typically present on them.

Other questions arise to how the user's experience of clutter is actually being measured. Rosenholtz et al. establish subjective judgment of clutter with a simple rank-ordering procedure of stimuli, where they had a series of participants order 25 printed maps from least to most cluttered (2005, 2007a). While both of their experiments established a statistically significant correlation between observer judgments and the Feature Congestion model, the paradigm used by Rosenholtz et al., does not provide us with information on how sensitive humans are to subtle changes in Feature Congestion, which has particularly important practical implications when applying the algorithm to inform real-world design decisions.

2 Methods

2.1 Participants

A total of 29 SAP employees (6 females, 24 males) participated in the present experiment, in exchange for a small gift. Participants ranged in age from 29 to 54 years, with a mean of 39.17 years. All participants were naïve to the purpose of the study, and reported normal or corrected vision.

2.2 Apparatus

Three identical laptops were used to present the participants with the stimuli. Each laptop was equipped with a 15.6 inch colour monitor at resolution 1600*900px, a 2.53 GHz processor, and used a custom coded PXLab script to present the stimuli (Irtel, 2007). The screen was viewed at a distance of roughly 55cm (unrestrained). Responses were made using a

mouse pointer to select the more cluttered stimulus on screen, which was then confirmed by pressing the space bar.

2.3 Stimuli

The stimuli were chosen out of a set of 121 websites (RandomWebsite.net, 2011). From the screenshots, we removed the browser frame and normalized the image size to 989*579px. We then ran the sets through the Feature Congestion model (Rosenholtz et al., 2007b). After computing the clutter ratings we then selected eleven, roughly equidistant colour website screenshots were selected, which ranged from 2.61 (lowest) to 13.15 (highest) in Feature Congestion. The spacing of the items was done by binning all of the items in the stimulus pool in to ranges of one. We then took all of the items in each of the individual bins, producing a mean value. The item in a given bin which was closest to the mean was subsequently taken for our stimulus set (see figure 2).

Figure 2: Stimuli, in order from lowest (top left) to highest Feature Congestion (bottom right), used during the experiment. Note that item 11 was removed from the final analysis.

2.4 Design and Procedure

In order to establish a measure of experienced clutter, we opted for a paired comparison paradigm, which allows us to estimate scale values on a ratio scale level (see section 2.5). Participants were presented with a single block of 55 comparisons ($\frac{n(n-1)}{2}$), which were randomized within block for both the pair presentation order, and right-left presentation of the stimuli during the decision making screen.

Participants were recruited and tested in a secluded area of an office, which had constant artificial lighting. The participants sat in front of the monitor and read the on-screen instruc-

tions, which was reinforced with a verbal explanation of the experiment. In the instructions, the participants were asked to evaluate a series of pairs of website screenshots, judging which website was more cluttered (Figure 3). [1] [2] Furthermore, the participants were encouraged to make a decision as quickly as possible, but not rush their decisions. If the participant did not have any questions at this point, the participant began the first trial with the experimenter looking on, in order to familiarize them with the experimental process. If the participant asked the experimenter if his or her choice was correct, the experimenter would respond, "If this is what you feel is the most cluttered item on the screen, then you may confirm your answer," in order not to bias the participant in one direction or another. A similar response was given to participants if they asked for help or commented on the difficulty of the task during the remainder of the experiment.

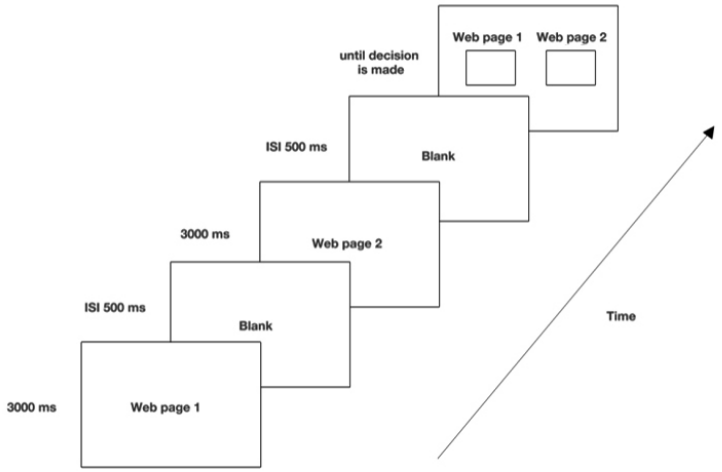

Figure 3: Illustration of the sequence of events in the experiment. Each trial was signaled by a blank screen with the phrase "next pair" at the center. The first webpage appeared for 3000ms, followed by an ISI of 500ms, and then the presentation of the second webpage for 3000ms. Finally both webpages were presented side-by-side to allow the participant to decide which of the two pages was more cluttered. The right-left ordering in the final display was randomly determined as a means of preventing ordering effects.

During each trial, the participants were presented with the first website screenshot for 3000ms, which was then followed by an ISI of 500ms. The second website screenshot was then shown for 3000ms, and again followed by a 500ms ISI. Finally, participants were

1 All participants in the present experiment were native German speakers, and the entire test procedure was administered in German.

2 There is no formal translation in German for the word clutter. Consequently, in order to cover the connotation of clutter to confusion (Rosenholtz et al., 2005), we decided to use the word, *unübersichtlich*, which is roughly translated as something that has been designed in a way that one cannot get a complete overview, and the content is hard to grasp (unübersichtlich, 2011)

shown a decision screen, where two scaled images (500*292px) were presented on screen at the same time. The decision screen remained visible until a decision by the observer was made.

2.5 Data Analysis and Results

From subjects' responses we derived a scale model based on the Elimination By Aspects (EBA) method (Wickelmaier, 2004). The result of the EBA procedure is a ratio scale, where stimuli can be localized with regard to the amount of a certain sensory property assumed to constitute a linear dimension. The paradigm allows us to focus on perceived clutter as a sensory property of screens, and also to assess via model tests the assumption that it actually does constitute a linear dimension. The ratio scale level allows us to draw more quantitative conclusions from the data than mere ranking.

Figure 4 shows estimated EBA scale values (in the following referred to as *perceived clutter*) as a function of the respective stimulus' Feature Congestion measure. One stimulus (11) constituted an outlier, and was hence removed. This stimulus differs from other stimuli in its treatment of, or actually blatant absence of, vertical alignment between colored text blocks. It was argued previously that such alignments are considered good design practice, so subjects might have reacted more sensitively to its violation by "punishing" this design with a higher clutteredness rating. The scale model, excluding stimulus 11, passes the likelihood ratio test for the EBA model ($p>0.12$). The relationship between Feature Congestion and perceived clutter values proves to be nonlinear, but linear in a logarithmic plot as depicted in fig. 5. Here, the linear regression is highly significant (ANOVA $F(1,8)= 44.54$, $p <.000$)) and shows a correlation coefficient of $r=0.92$, $R^2=0.85$.

3 Discussion

Our results indicate that Feature Congestion as calculated from the algorithm presented by Rosenholtz et al. (2007a) can quite accurately predict subjects' perception of a website as cluttered.

The exponential relationship we observe between Feature Congestion and perceived clutter (as being measured by paired-comparison scale values) obviously warrants further discussion and research. Fig. 6 shows a linear plot of perceived clutter by Feature Congestion (the exponential trend line corresponds to the linear one in the logarithmic plot in fig.5). Variations of Feature Congestion below a critical value of about 9 have only limited effect on the perceived clutter. Beyond this value, the perception of clutter raises rapidly. This is in line with results reported by Rosenholtz et al. (2007a), where the relationship of Feature Congestion to search time was also exponential.

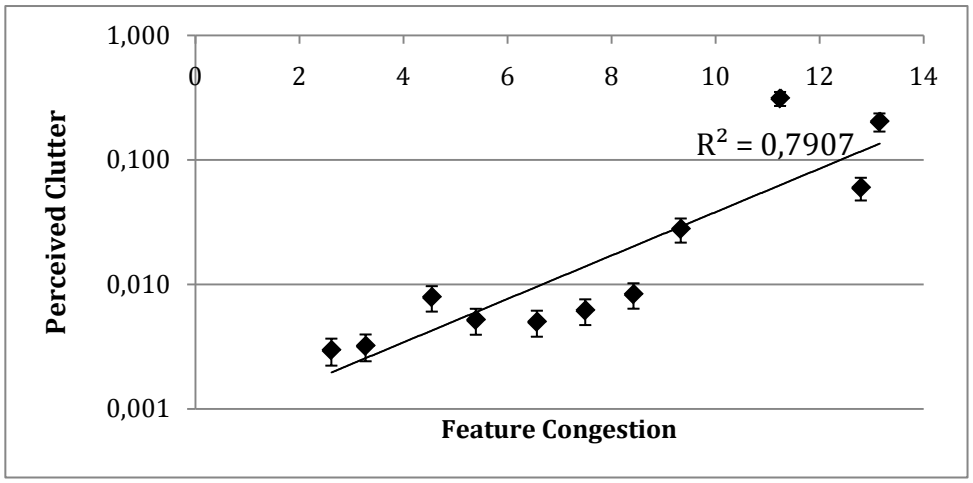

Figure 4: Estimated ratio scale values of perceived clutter by stimulus Feature Congestion (logarithmic plot), including the outlier with FC≥ 11. Error bars represent the standard error of the EBA parameter estimate

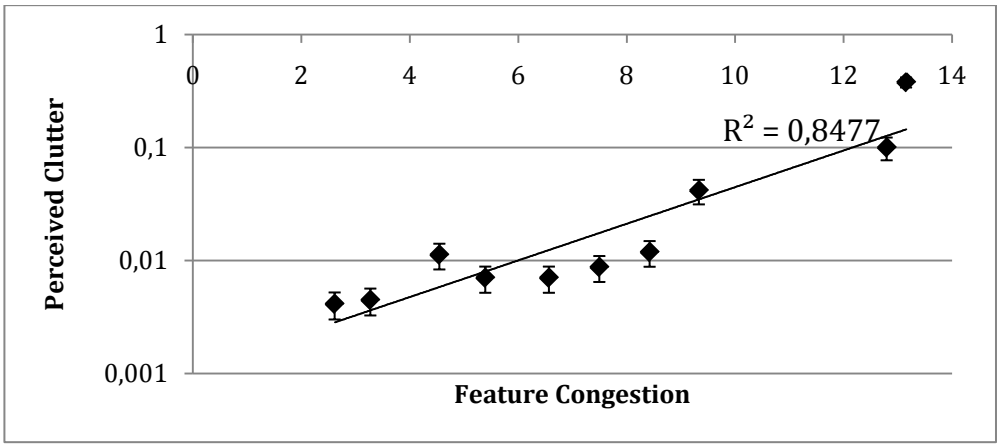

Figure 5: Estimated ratio scale values of perceived clutter by stimulus Feature Congestion, excluding the item with FC≥11. Error bars represent the standard error of the EBA parameter estimate.

For Feature Congestion as a screening measure for perceived clutter on computer user interfaces, there are two direct practical implications of the exponential curve. First, we can observe that there actually is a threshold value of Feature Congestion above which users' perception becomes critical. Fig. 6 further shows a histogram of an additional randomized sample of 333 websites within Feature Congestion intervals of FC=0.5, together with an approximated normal distribution (mean FC=5.90, SD=1.70). Considering this statistic, together with the approximated exponential curve of perceived clutter by Feature Congestion, we can observe that a substantial number of websites reaches or even exceeds the limit where perceived clutter becomes critical.

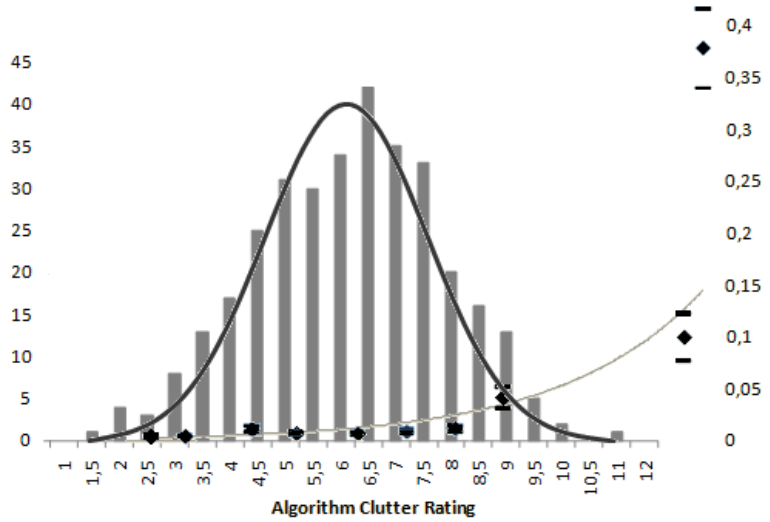

Figure 6: Perceived clutter by stimulus Feature Congestion (item 11 excluded), overlaid with number of websites with this Feature Congestion. Error bars represent the standard error of the EBA parameter estimate. Exponential trend line and normal distribution are approximated.

Second, below this critical value, the slope of the curve depicting the relationship between Feature Congestion and perceived clutter turns out to be rather shallow. In this "low-congestion" domain, it would take a difference of about 2 points of Feature Congestion to generate a difference in perceived clutter greater than the standard error of the EBA scale estimate. While we still can predict the mean scale value with some accuracy, individual variations of clutter perception have the potential to conceal differences, thereby limiting the practical usefulness of the Feature Congestion measure. Also, perceived clutter scale values in this domain still need to be calibrated with regard to the criticality of a given value, i.e. at which scale value exactly a screen becomes unacceptable. Only given those data would we be able to predict reactions in a given user population accurately.

Currently we can only speculate about the causes for the exponential relationship between Feature Congestion and perceived clutter, and its underlying physiological mechanisms. In fact, obviously both Feature Congestion and a person's subjective perception of something as being cluttered must have an upper limit – there certainly is a point where there is simply nothing to be seen on a display, where the curve is bound to reach an asymptote. Given the good exponential fit of our data, it seems like we are rather on the far left hand side of such a theoretical curve, which is the practically relevant branch for user interface design purposes.

Considering these findings, we can conclude that the Feature Congestion model provided by Rosenholtz et al. (2005, 2007a) is an excellent tool to automatically detect severely cluttered screens. Although correlations to perceived clutter are still high, it is less sensitive in the mid-range of Feature Congestion where individual variations in the perception of clutter have greater influence.

Open questions remain about the role of vertical alignment and structural conventions on screens, which we hypothesize might accounted for removing one outlier stimulus from the analysis. Also, the role of text in displays certainly warrants further investigation, since application user interfaces typically consist of text rather than images. In fact, dynamic alphanumeric data may or may not be present on an application screen depending on its state, but they don't indicate the clutteredness of an application per se. We plan to address these questions in follow-up experiments.

Acknowledgements

The authors would like to thank Patrick Fischer his assistance and Mike Dodd for his helpful comments on the paper. The authors also wish to thank the reviewers for their constructive feedback on this paper.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological Science, 15, 106–111.

Awh, E., Barton, B., & Vogel, E.K. (2007). Visual working memory represents a fixed number of items regardless of complexity. Psychological Science, 18, 622–628.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and brain sciences, 24(01), 87–114.

Elliott, B. (2008). Anything is Possible: Managing Feature Creep in an Innovation Rich Environment. IEEE 978-1-4244-2146-6/08.

Irtel, H. (2007). PXLab: The Psychological Experiments Laboratory. Version 2.1.11. Mannheim (Germany): University of Mannheim.

RandomWebsite.net (2011). Retrieved December 9, 2010, from http://www.randomwebsite.net/

Rosenholtz, R., Li, Y., Mansfield, J. & Jin, Z. (2005). Feature Congestion: A Measure of Display Clutter. In: CHI 2005, April 2-7, Portland, Oregon, USA

Rosenholtz, R., Li, Y., & Nakano, L. (2007a). Feature Congestion and Subband Entropy measures of visual clutter. Software. Retrieved March 17, 2011, from http://dspace.mit.edu/handle/1721.1/37593

Rosenholtz, R., Li, Y., & Nakano, L. (2007b). Measuring visual clutter. Journal of Vision, 7(2):17, 1–22, http://journalofvision.org/7/2/17/, doi:10.1167/7.2.17.

Unübersichtlich. (2011). In Farlex, Inc. (Hrsg.): The Free Dictionary. Retrieved February 14, 2011, from http://www.thefreedictionary.com/

Wickelmaier, C. Schmid (2004): A Matlab function to estimate choice-model parameters from paired-comparison data. Behav. Res. Meth. Instr. Comp. 36 29-40.