Feature Congestion: A Measure of Display Clutter

Ruth Rosenholtz, Yuanzhen Li, Jonathan Mansfield, and Zhenlan Jin
MIT Dept. of Brain & Cognitive Sciences
3 Cambridge Center, NE20-447, Cambridge, MA 02139, USA
Contact : rruth@mit.edu

ABSTRACT
Management of clutter is an important factor in the design of user interfaces and information visualizations, allowing improved usability and aesthetics. However, clutter is not a well defined concept. In this paper, we present the Feature Congestion measure of display clutter. This measure is based upon extensive modeling of the saliency of elements of a display, and upon a new operational definition of clutter. The current implementation is based upon two features: color and luminance contrast. We have tested this measure on maps that observers ranked by perceived clutter. Results show good agreement between the observers' rankings and our measure of clutter. Furthermore, our measure can be used to make design suggestions in an automated UI critiquing tool.

Author Keywords
Clutter, visualization, feature congestion, visual interfaces, display design, recommender systems, information density

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Clutter is an important phenomenon in our lives, and an important consideration in the design of user interfaces and information visualizations. Many existing visualization systems are designed to reduce clutter by filtering what objects or information the user sees [e.g. 1, 9, 32], or using non-linear magnification techniques so that objects in the center of the screen are allowed more display area [11]. Tips for designing web pages, maps, and other visualizations often focus on techniques for displaying a large amount of information while keeping clutter to a minimum through careful choices of representation and organization of that information. However, in spite of the existence of these tools and tips, we lack a clear understanding of what clutter is, what features, attributes, and factors are relevant, why it presents a problem, and how to identify it. If we can better understand clutter, we can create tools to automatically identify it in a display, and even eventually create systems that will advise a designer when the level of clutter is too high, and suggest techniques for reducing visual clutter. Such a system might also indicate a too sparse and uncluttered display; excessively sparse displays can lead to inefficient use of available display space, and require the user to go to a new page or view in order to acquire the information required for their task.

In the following section, we discuss lay definitions of clutter, and suggest an operational definition. This definition allows us to enumerate areas of knowledge of human perception that may be brought to bear on understanding clutter. Next, we present a more restrictive definition which gives us a first cut at a measure of the level of clutter in a display. This Feature Congestion measure of clutter makes use of extensive modeling of what makes items in a display visually salient. We present the results of an experiment in which observers were asked to rank the level of clutter in a collection of maps. We show that our Feature Congestion measure of clutter does a good job of predicting the relative clutter rankings of these maps.

WHAT IS CLUTTER?

A common wisdom notion of clutter is simply that it is what occurs when one has too many objects. Within the CHI community, a common method for clutter reduction is simply to remove items or information from the central portion of the display [1, 9, 32]. The American Heritage College Dictionary goes a step further, defining clutter as "a confused or disordered state, caused by filling or covering with objects." This clarifies in an important way that clutter is not merely the state of having too many objects; it refers to a state in which a number of objects cause "confusion."

We suggest that clutter is not a purely non-aesthetic result of having too many objects, but rather that its confusion-causing aspect is important, and will enable us to get a handle on how to measure and reduce clutter. For the purposes of a scientific exploration of clutter and its impact on work, we suggest the following operational definition:

Definition: *Clutter is the state in which excess items, or their representation or organization, lead to a degradation of performance at some task.*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
Copyright 2005 ACM 1-58113-998-5/05/0004…$5.00.

The association between clutter and the representation or organization of information in a display has been recognized by a number of designers and researchers. Web page design tips often focus on organization of textual and graphic information to reduce clutter. Map generalization work focuses on changing representations of information at different map scales to reduce clutter and improve usability [4]. Semantic zoom [e.g. 18] and goal-directed zoom [32] interfaces allow for varying both the amount and the representation of information at each scale of the visualization.

The above definition of clutter brings up two key points: the association between clutter and degradation in performance, and the notion that clutter may depend upon the user's task. Note that the idea that clutter leads to a performance degradation in using a display does not mean that we should aim for displays with zero clutter; in a complex task involving multiple displays, too little information in one display will only force the user to switch to another to acquire the necessary information, which has its own performance costs.

The notion of task dependence seems to agree with people's intuitive notions of clutter. In our experiment on ranking clutter in maps, reported later in this paper, a number of subjects spontaneously suggested that they were trying to think of a task in order to judge clutter.

Clutter likely also depends upon the expertise of the user. Consider a weather map of the United States, with labels specifying the names of a large number of U.S. cities and their expected high temperature. To someone fairly unfamiliar with the U.S., with a task of finding tomorrow's forecast temperature for Denver, this map could appear quite cluttered, since the person would have to search through a large number of cities to find Denver. However, for someone familiar with the U.S. the map might not be very cluttered at all, since they know where to look for cities of interest. In our map ranking experiment, a number of the users reported just this intuition. One user volunteered, "This map would seem pretty cluttered, except what would I use it for? I would probably want to look up the temperature for San Francisco, and I know where that is."

As another example of the dependence upon user expertise, Sellen and Harper [24] cite the example of a manager whose desk his employers found so unacceptably cluttered that they insisted he put all his papers in boxes whenever the company had visitors. But the manager was famous for being able to quickly extract exactly the paper he wanted from those stacks of papers. It is difficult to believe that the manager found his office as cluttered as his employers did.

The dependence upon task and user expertise likely provides a partial explanation for why people do not always agree on the level of clutter. On the other hand, we often find a room cluttered when we have no task to complete, and a sense of clutter does not seem to go away completely even when we know the location of every object. Clearly perception of clutter depends upon a complex set of issues, including such amorphous issues as personal aesthetics.

The complexity of clutter perception, and its dependence upon high-level issues like task and user expertise, can make deriving a clutter measure seem daunting. We suggest that, although user knowledge and task are important, we can, as a first step, make a fair amount of headway into understanding clutter by considering general visual tasks, and looking at low-level contributions to clutter.

In a user interface or visualization, designers aim for quick, easy, natural, and veridical access to the visual information present in a display. Such access typically consists of basic visual subtasks. Researchers know a fair amount about areas of perception and cognition in which additional objects or information in a display can cause degradation in performance at a basic visual task (see the following section). Furthermore, the degradation of performance at many basic visual tasks with increasing numbers of objects does not have a strong dependence upon the expertise of the user. The following subsection gives a brief and partial enumeration of basic visual contributors to clutter. In many cases we can say something quantitative about performance degradation, thus enabling meaningful measures of clutter.

A full model of task performance supercedes a measure of clutter, by our definition. However, in the absence of a complete performance model, clutter gets at an important question: for a given situation (information to be displayed, possible tasks, and user expertise), how does performance vary with display design, i.e. the amount of information displayed, and its representation and organization?

Basic Visual Tasks for Which Increasing the Number of Items Degrades Performance

It is well known that additional items can make search for a visually-defined target slower or less accurate (see [31] for a review). It is true that in some cases the target seems to "pop out" at the viewer, and search shows little effect of additional display items. However, this largely occurs in very homogeneous displays, for instance when searching for a blue disk among a display of red disks. For more typical displays of even moderate complexity, one should expect that extra items make search slower.

Extra display items also degrade performance is in what is known as lateral masking. It is more difficult to see and/or identify a low contrast item in the presence of noise when there is "crowding," i.e. when there are other items very nearby. In a related phenomenon, it is more difficult to count how many objects there are, or identify objects, particularly in the periphery, when objects are closely flanked by other objects. Though this is an issue of object density rather than their number, as the number of objects increases, often density also increases. Another related phenomenon is that when a number of dense curves are present in an image, for instance in a complicated maze, it can be difficult to visually trace one of the curves without using one's finger to keep track.

More objects increase the chances of occlusion of one object with another. Occlusion potentially leads to difficulties in correctly recognizing the occluded figures.

More items can also stretch or exceed the limits of short term memory [13]. This can make it difficult to get veridical information from a display. For instance, more objects can make it difficult to comprehend the mapping between a graph and its legend, to compare the information in two spatial locations in the display, or to comprehend the relationship between two different views into an information space, for example a map at two different levels of zoom. The relevant factor seems not to be merely the number of objects, but their features (color, orientation, etc.); the capacity for certain simple features is higher than for more complex features. Researchers have made headway into understanding the relationship between features and capacity, but there is still much work to be done.

The above are some of the most important phenomena in which performance may degrade due to additional items. Sometimes more items lead to performance *benefits*. This tends to occur when the items are low entropy, meaning that the appearance of one item is easily predicted from its neighbors. Under such conditions, it can be easier, for instance, to spot a trend in data when there are a larger number of points contributing to that trend, or to notice a grouping of items with similar characteristics. Given a consistent trend or group of items, more items can aid detection of a deviation – an outlier with features or trend different from the group, or a boundary between two differing groups.

While all of these phenomena have been studied extensively in the human vision literature, some we understand better than others; for some we have predictive models, for others we merely have a lot of experimental data. Phenomena for which we have good predictive models are prime candidates for a starting point for deriving a measure of clutter. Two of these areas are visual search and masking.

Masking models have proven useful in quantifying such things as the visibility of compression artifacts in JPEG images. However, predominantly these models have predicted perception of low-contrast patterns [29], of limited utility in evaluating clutter for typical displays.

On the other hand, many real-life tasks have a strong visual search component to them. We feel that this is our best choice for a starting point in developing a measure of clutter for information visualizations. In the next section we discuss previous work in deriving a measure of clutter. In the section following that, we will discuss a model of visual search that is particularly well suited to creating a measure of clutter, and derive our measure of clutter.

PREVIOUS WORK

As mentioned above, a number of visualization techniques have aided the user in reducing clutter, by allowing them to selectively view detail. Such techniques have not required a measure of clutter, relying instead upon a user's ability to navigate to a view with less clutter, and upon the tendency of a visualization to become less cluttered as the user zooms in on it, if additional items are not added to the display with increasing zoom. The work of Ahlberg & Shneiderman [1], as well as that of Fiskin & Stone [9], for instance, allow the user to filter what information appears in the display using dynamic queries, and also allow zooming to reduce clutter. Non-linear magnification schemes such as fisheye views [11] allow the user to make use of reduced clutter in zoomed-in views, while maintaining context in the periphery of the display. The addition of a measure of clutter could allow such systems to better control the level of clutter, while taking some of the burden off of the user.

Woodruff et al [32] go beyond relying upon the user to filter information or zoom so as to reduce display clutter. Their system, VIDA, helps users to construct visualizations with neither excessive clutter nor sparsity of information, guided by the cartographic principle of constant information density. For this, they require a measure of information density, essentially equivalent to a measure of clutter. Woodruff et al experiment with several measures of information density: the number of visible objects, and the number of vertices. VIDA allows plug-in of alternative measures of information density.

Other researchers have suggested additional measures of information density or clutter. Dynamic Logic's study on the effect of clutter on ads used as a metric the number elements on a web page, where an element consisted of a word, graphic, or "interest area" [2]. Tufte [28] suggested the number of entries in the source data matrix per unit area. Nickerson [14] enumerated a number of density measures, including the number of graphic tokens per unit area, the number of vectors needed to draw the visualization, and the length of program to generate the visualization. Frank & Timpf [10] suggested the amount of "ink" per unit area as a metric for simple black & white maps.

These metrics have a number of difficulties, and few have actually been implemented. Certainly the amount of clutter has some dependency upon the number of objects, graphic tokens, or entries in the source data matrix in the display. The number of objects (or, conversely, the amount of blank space) has been shown to influence task performance at both a directory assistance search task [25], and at a number of tasks with a map [20]. However, column alignment helped in the directory assistance task, and Phillip & Noyes [20] found that "like clutters like," e.g. additional lines on a map cause increased difficulty for tasks involving lines. Counting the number of objects does not take into account the appearance or organization of the objects. Merely counting the number of elements on a web page did not prove to be a good measure of clutter in the sense of correlating with ad effectiveness [2], though human clutter judgments did correlate with ad effectiveness. Furthermore, for complex visualizations, the number of objects can be ill-defined; how many objects make up a texture indicating a mountain range on a map, or the outlines of U.S.

states? A vertex may be similarly ill-defined – how sharp a turn in a curve is required? The amount of ink is ill-defined for color displays. In addition to these difficulties in applying suggested clutter measures when we are given a list of drawing objects in a visualization, the situation is even bleaker if one is given an existing visualization in image form, as with the maps in our empirical study.

In the following section, we discuss relevant research into predictive models of a particularly ubiquitous visual task: visual search. We then introduce our measure of clutter, based upon this research. Our clutter measure, in its current implementation, does not explicitly deal with objects, but it will be a function of the number of objects in the display, as well as of their appearance and organization. Furthermore, in its current implementation, our measure may be applied to any static display, since it takes an image as input, and does not require a list of items in the display.

THE STATISTICAL SALIENCY MODEL AND THE FEATURE CONGESTION MEASURE OF CLUTTER

Visual Search

A vast amount of research has been done on visual search. Treisman's early work [e.g.26] is justly famous and well-known among HCI researchers and thus worth discussion although it is not the most useful research for our purposes. Treisman suggested a dichotomy between so-called parallel vs. serial visual search. Parallel search involved simultaneous processing across the visual field, and thus was insensitive to the addition of more items to the display. Serial search, on the other hand, involved a serial mechanism which visited one or a small number of items at a time, and thus would be slower with additional items. Treisman suggested that search could be used to determine basic features of the visual system: search for a target defined by a basic feature ("feature search") would be parallel, whereas search for a target defined by a combination of basic features ("conjunction search") would be serial. Search asymmetries, in which search for target A among distractors B is easier than search for target B among distractors A, provide, by this account, additional information about the basic feature detectors and their operation in visual search.

One reason this model is of limited utility for our purposes is that it is primarily descriptive of experimental results, rather than predictive. In particular, it is difficult to extract from Treisman's model predictions about search in complex displays. In addition, many of Treisman's early theoretical conclusions are now in question. Few researchers still believe in a simple dichotomy between "feature" and "conjunction" search (e.g., [8, 17, 31]), and her interpretation of search asymmetries has been seriously challenged [22]. Newer predictive models, based on concepts from image processing and statistics, are rising in prominence. These models have proven quite powerful at predicting search performance without the need for explicit asymmetries or separate mechanisms for features vs. conjunctions. These newer models are the basis for our present work.

Researchers have developed a number of predictive models of visual search [30, 17, 12, 22]. Some have aimed at predicting a quantitative measure of search performance, such as reaction time, or percent-correct performance [17, 30]. Other work has aimed at predicting the "saliency" of display items [12, 22]. The saliency of an item corresponds qualitatively to ease of search for that item if it were the target, and correlates with the likelihood that a user makes an eye movement to that item (eye movements also depend upon the user's task, of course). The saliency measures are more versatile for evaluating displays, since they can operate on arbitrary image data. In looking for a qualitative measure of the clutter in a display, qualitative measures of search performance like saliency measures should suffice.

Two saliency models could be appropriate for our purposes: one by Itti et al [12] and one by Rosenholtz [22]. Both are designed to model human performance with visual displays. The Itti model is based on biologically inspired mechanisms, starting with linear filters similar to receptive fields found in visual cortex, and then applying a variety of non-linear neural-like operations. Rosenholtz's model tries to capture human performance at a functional rather than biological level, utilizing the notion that the visual system is designed to characterize various statistical aspects of the visual display. This statistical framework leads to easier intuitions for why a target in a display is or is not salient, and can suggest features for a salient target. Because these aspects will prove useful when developing a clutter model, or making design recommendations, we will use the Rosenholtz model as a starting point in our investigations.

The Statistical Saliency Model

Rosenholtz begins with the premise that the visual system has an interest in detecting "unusual" items. She suggests that an item is "unusual," and thus salient, if its features are outliers to the local distribution of features in the display. Following a long line of visual search researchers, these features are likely to include such things as contrast, color, orientation, and motion [31]. Rosenholtz suggests a measure like a z-score for the degree to which a feature vector, T, is an outlier to the local distribution of feature vectors, represented by their mean, μ_D, and covariance, Σ_D. The saliency, Δ, is given by the following equation:

$$\Delta = \sqrt{(T-\mu_D)'\Sigma_D^{-1}(T-\mu_D)}$$

where ()' indicates a vector transpose.[1] The higher the target saliency, the easier the predicted search. The saliency, Δ, can be thought of as a formalization of Duncan & Hum-

[1] This is a parametric measure of "outlierness." One could of course also use a non-parametric measure, which might seem more appropriate, since feature distributions are rarely Gaussian, as assumed by the parametric measure. However, empirical evidence [23] suggests that the parametric measure better models search performance.

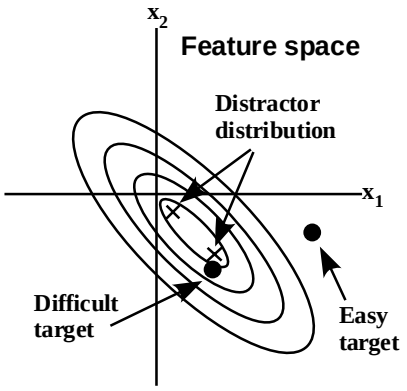

Figure 1. Graphical depiction of the statistical saliency model. Ellipses represent points of equal saliency. Outer ellipses correspond to greater saliency and easier search.

phreys' [7] notion of the different roles of target-distractor versus distractor-distractor similarity in search performance. Rosenholtz' model predicts the results of a wide range of search experiments involving basic features such as those above, including experiments that were previously thought to involve search asymmetries [22].

To better understand the model, and get an intuition for the Feature Congestion measure of clutter described in the following section, consider the graphical interpretation in Figure 1. The statistical saliency model represents the local distribution of features by their mean, μ_D, and covariance, Σ_D. This is equivalent to representing the distribution by a set of *covariance ellipsoids* in the appropriate feature space, as shown. The innermost, 1σ, ellipsoid indicates feature vectors one standard deviation away from the mean feature vector. The 2σ ellipsoid indicates feature vectors that are two standard deviations away from the mean, and so on. A target with a feature vector on the $n\sigma$ ellipsoid will have saliency $\Delta=n$. The farther out the target feature vector lies on these nested ellipsoids, the easier the predicted search.

A designer can also use this model to choose attention-drawing features for a particular item in the display. (If the designer wants *not* to draw attention to an item, the model can help with that as well.) Suppose the relevant feature space is a three-dimensional perceptual color space, like CIELab [6]. If the designer wants to add an item to a given portion of the display, such that the saliency of that item is at least d, then any colors outside of the $d\sigma$ covariance ellipse will suffice. The model also tells us the most efficient direction to move in color space when picking the color of the new item. For a given distance from the mean color, colors in the direction of the shortest principal axis of the covariance ellipsoid will be most salient.

Furthermore, the covariance ellipsoid can inform a designer about the difficulty of drawing attention to a new item in a display. If the covariance ellipsoid is large relative to the set of possible features (e.g. relative to the color gamut), then the saliency model predicts that it will be difficult to add an attention-getting item to the display, because few possible feature vectors will have sufficient saliency. This observation provides the key for our first cut at a measure of clutter, described in the next subsection.

Feature Congestion

As mentioned earlier, we suggest a definition of clutter as a state in which excess items, or their representation and organization, cause degradation at some task. We have enumerated a number of basic visual tasks with known effects of "clutter." However, we do not have good predictive models of many of these basic visual tasks. As a first cut for a measure of visual clutter, we suggest considering the task of visual search, and adding components corresponding to other basic visual tasks as models become available.

Visual search is an important subtask in many real-world tasks. A user must find buttons or other components of a user interface. An alert system must draw attention to a relevant part of a display so that the user can find it easily. Comprehending information visualizations also has a significant visual search component. Because of the ubiquity of the visual search task, even a very limited measure of clutter based only upon this task should be a significant step. The question is how to go from a model of visual search ease to a measure of clutter.

One possible measure is that clutter is inversely proportional to the average saliency of a list of potential search targets in a display. If all possible targets in a display are highly salient, and thus easy to search for, then visual search performance has not been compromised by the number and appearance of items in the display, and thus the display is uncluttered. This measure seems straightforward from our definition of clutter, but it has several disadvantages. The first is that while this may be a worthwhile measure to investigate given a list of potential targets, such a list may not always be readily available. If we guess that every bit of the display consists of a potential target, this measure amounts to computing the inverse of the average saliency per unit display area. This could give counterintuitive results. Consider, for instance, a map of the United States with nothing marked on it but a number indicating the temperature in Chicago. The background of the map will have low saliency, and the temperature a high saliency. Integrating saliency over the entire map will give a low value, indicating, according to this measure, a high level of clutter. Yet clearly this map would be highly uncluttered.

An example illustrates another problem with this straightforward measure of clutter. Consider two displays, each with 50 items, all potential search targets. In the first display, all items have the same color and shape. In the second display, the items vary randomly in both color and shape. The statistical saliency model says that none of the items in the two displays is very salient, but for very different reasons. The uniform items in the first display have low saliency because of their uniformity ($T-\mu_D=0$). The highly variable items in the second display have low saliency be-

cause, while they differ, the difference is small compared to the amount of variability in the display, (Σ_D is quite high). This fact that these two displays will have similar saliency profiles, and thus yield similar measures of clutter, violates our intuitions: based purely upon aesthetic considerations, it seems unlikely that these two displays would be judged equally cluttered.

We suggest an alternate way of characterizing clutter. As more items are added to a display, populating a greater volume of feature space, there is less room in feature space to add new salient items. We refer to this condition as Feature Congestion. Increased congestion leads to degraded performance, e.g. in searching for the new items. We propose that Feature Congestion is one of the major causes of clutter. By analogy, one might consider a person's desk highly cluttered if it is difficult to leave a note that will draw attention by its location or appearance.

We suggest, therefore, a new measure of clutter, based upon predicting the level of feature congestion in an image. This model can operate without knowledge of what constitutes an item or target in the display. Instead, we treat the display essentially as a given background image, and ask how easy it would be to add a new, salient item.

The statistical saliency model indicates the difficulty in adding a new, salient item to a local area of a display: the size of the local covariance ellipsoid represented by Σ_D gives a measure of this difficulty, and thus of the local clutter in the display. Locally measuring the ellipsoid size, and pooling over the relevant display area, gives a measure of clutter for the whole display.

We call this measure the Feature Congestion measure of visual clutter. Displays with high clutter, according to this measure, are cluttered because feature space is already "congested" so that there is little room for a new feature to draw attention. Too many colors, sizes, shapes, and/or motions are already clamoring for attention.

Our discussion leads us to a surprisingly simple measure of clutter – the clutter in a local part of a display is related to the local variability in certain key features. One can draw parallels between this measure and recent work on the related problem of visual complexity. Oliva et al [16] had users hierarchically sort photographs according to their complexity, and indicate at each hierarchical level the basis for the sort. Users indicated that complexity depended upon quantity and variety of objects, detail, and color, as well as upon higher-level, more global concepts like the symmetry, organization, and "openness" of the depicted space.

Implementation of the Feature Congestion clutter measure involves 4 stages: 1) compute local feature (co)variance at multiple 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.

In the current implementation, we use color and luminance contrast as features. Additional features would likely increase the power of our measure, and allow it to deal with a full range of possible inputs: from GUI's, through web pages, to natural scenes. We wanted to start with the simplest model we could; thus we restricted ourselves to color and luminance contrast, since they are essential to so much in pattern perception (for examples of using these features, along with orientation, in modeling perceptual phenomena, see [19, 29]). Color naturally seems important to our sense of clutter. Contrast feature detectors are known to exist early in the visual system, and can not only detect simple luminance contrast, but also serve as a measure of size and shape (see [21] for a review of evidence that such detectors may mediate pre-attentive processing of shape in the human visual system). Future implementations might include features such as motion and orientation, which are also believed to be basic features in visual search and attention. Treisman & Gelade's work [26], might also be used to suggest features, though many of the basic features they suggest might already be handled by luminance contrast, and as mentioned above the notion that visual search may be used to identify basic features has been called into question.

Below we briefly describe our implementation. See [21] for more details on the feature covariance stage of processing, as similar steps are used to segment an image into regions of different texture – a process thought to have much in common with visual search in the human brain.

Feature covariance

We start by converting the input image into the perceptually-based CIELab color space [6]. We then process the image at multiple scales (currently 3) by creating a Gaussian pyramid by alternately smoothing and subsampling the image [3]. For luminance contrast, we then compute a form of "contrast-energy" by filtering the luminance band by a center-surround filter formed from the difference of two Gaussians, and squaring the outputs. Then, for each feature, we compute the local feature (co)variance. This may be done efficiently through a combination of linear filtering (to average over a local area) and point-wise non-linear operations (to compute the variance). This gives us contrast variance (a scalar), and a 3x3 covariance matrix for color. From this covariance matrix it is straightforward to compute the volume of the covariance ellipsoid, our local measure of color feature congestion. The contrast feature congestion is simply the square root of the contrast variance.

Combine across scales

For each feature, we combine feature congestion across scale by taking the maximum at each pixel. We chose the max operation based upon the reasoning that a feature is locally congested if it is congested at any scale.

Combine across features

Next we combine color and contrast clutter at each point. Since color clutter is the volume of the covariance ellipsoid, while contrast clutter is a "length" representing variance in a single dimension, we first take the cube root of the color

clutter measure to make the two measures more equivalent. The two features are scaled so as to approximately equate color & luminance contrast discriminability. We then combine color and contrast clutter at each point by taking the average of the two. The result is a single *clutter map* of local clutter at each display location. Combining by taking an average amounts to saying that an image is very cluttered at a point if it has both color and contrast clutter, but less cluttered if it is cluttered only in contrast or only in color. High contrast clutter with low color clutter corresponds, for instance, to a block of black and white text, where the user would be faced with many "objects" (here, words), yet it would be relatively easy to draw attention to a saturated red word or symbol. High color clutter with low contrast color could correspond to a more unusual situation in visualization design, with a variety of colors present locally, but without much luminance contrast; in this situation the colors will not be very discriminable, and a small amount of luminance contrast will be quite noticeable.

A final model of clutter will almost certainly involve a more complicated combination rule than that used here. The features might, for instance, be combined into a single large feature vector prior to computation of the covariance, as might be suggested by [8]. Some features might have priority over other features; e.g. Callaghan [5] has suggested that color dominates over geometric form in texture segregation. Much basic research needs to be done to adequately model feature interaction. In the absence of such research, our aim was to see how far we could go with a simple measure. Attempts to allow a general linear combination of color and luminance contrast features did not greatly improve performance of our clutter measure.

Combine across space

Finally, we pool over space to get a single measure of clutter for each display. In the current implementation we simply take the average clutter value over the entire image.

Computing this clutter measure is reasonably efficient. Our MATLAB implementation, with virtually no attempt made to optimize code, computes clutter for a 512x512 image in approximately 10.5 seconds.

EXPERIMENT: CLUTTER RANKINGS FOR MAPS

In order to begin to test our measure of clutter, we have conducted an experiment in which we ask users to rank the clutter of a number of maps. Maps are a familiar visualization for many users, and have many features typical of more general visualizations. We were interested in whether users give consistent clutter rankings, and in whether or not we could predict those rankings using our Feature Congestion measure of clutter. Ideally, we would also test task performance, and measure its correlation with the clutter rankings given by our subjects and by our model. We are in the process of doing so; in the meantime, it is worthwhile testing whether our model can predict the human clutter rankings. It is reasonable to assume that the rankings will correlate inversely with performance, and it is known that clutter judgments can be predictive of objective measures of display effectiveness in other settings [2].

Methods

We collected 25 maps at various scales by searching the web for maps of the U.S. and San Francisco Bay Area. This included 9 maps of the U.S., 6 maps of the bay area, 3 maps of all of San Francisco, and 7 maps of a portion of the city. This selection of multiple map scales and cartographers naturally led to a wide range of clutter levels. We asked 20 users to order the maps according to perceived clutter. Users were recruited from a college campus, and ranged in age from 10-48. They were not given a definition of clutter, but rather were told to do their best with their own intuitive definition of what "clutter" meant. Each map was printed on 8½ x 11" paper in full color, and scaled so that its width filled the page. Users indicated their rankings by ordering the 25 maps. They received the maps in random order, and were instructed that this was the case.

Results

The following is the ordering of the map images from low clutter to high, according to their average ranking by human observers: 2, 1, 3, 4, 7, 9, 8, 10, 6, 11, 14, 16, 15, 5, 19, 17, 18, 13, 23, 20, 21, 24, 12, 22, 25. Thumbnails of these images may be seen in Figure 3, where they are ordered according to the clutter measure. Because thumbnails do not allow viewing of high frequencies which can be crucial to judgments of clutter, we have also made these images available at a somewhat higher resolution at https://dspace.mit.edu/handle/1721.1/7508. We computed Kendall's coefficient of concordance, W, from the rankings of the 20 users. This statistical test tests both whether there is agreement in rank data among a set of judges, and whether the images differ significantly in their perceived clutter (agreement in rank and a significant effect of image on ranking are equivalent). It returns a value of W between 0 and 1, where larger values indicate more agreement. This test yielded W=0.72, indicating significant and substantial correlation between the rankings of the different subjects ($\chi^2(24)=345.3$, $p<0.001$). The average Spearman rank-order correlation between pairs of subjects was 0.70. Thus, some images were clearly perceived as more cluttered than others, and there was a fair amount of agreement among subjects, though still, of course, variability in the rankings. Figure 2 indicates, for each map image, the 25^{th} and 75^{th} percentile rankings, as a measure of this variability. Clearly subjects show a great deal of agreement on the clutter level of some of the maps (e.g. #25), but quite a bit of disagreement on others (e.g. #5, #13).

Observers' comments gave some hint as to the reason for some of the agreement and disagreement on clutter rankings. For map #25, the strong agreement seems to have been due to the sense that this map obviously attempts to include too much information, and the resulting difficulty in quickly reading information from the map – two hallmarks

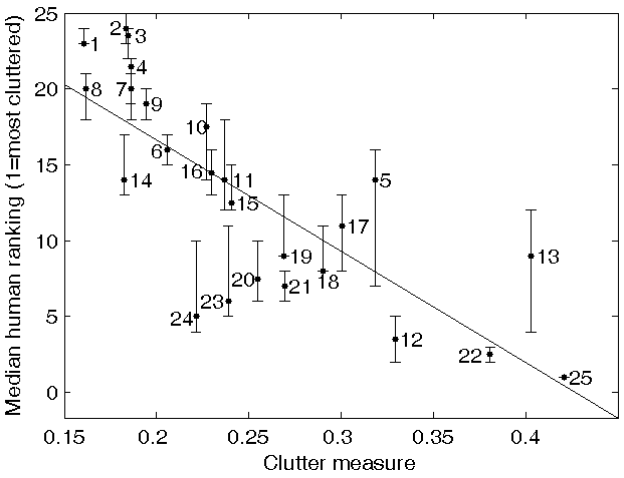

Figure 2: Clutter measure vs. median human ranking, with 25th and 75th percentile ranks shown by error bars. Diagonal line shows best linear fit of measure to rankings. Numbers indicate image numbers, as in Figure 3.

of clutter according to a number of the observers. Much of the disagreement about map #10 seems to focus on whether taking all of the text off of the map per se allowed a large amount of information to be presented without cluttering the map, or whether having the text separated from the map made the map slower to read and thus more cluttered. On the other hand, observers disagreed, e.g. on map #24, on how to weight a low number of colors and a large amount of text; the use of few colors was seen by many observers as reducing clutter, though a couple of observers thought that more colors could actually lead to less clutter.

Modeling with the Feature Congestion Clutter Measure

We compared our clutter measure with the human clutter rankings. Figure 3 shows the maps in order of their clutter as determined by the clutter measure. Calculating Spearman's rank-order correlation between the model and the average subject ranking, we find a significant correlation of 0.83 ($p<0.001$). This is comparable to the correlation between observers, suggesting that the model does as well as could be expected, given the variability among observers.

Figure 2 shows the relationship between human rankings and clutter measure. In general, the fit is good, as expected from the correlation coefficient. The images for which the clutter measure performs worst seem to be those with a large amount of text and little color variation, e.g. #23 and 24. This suggests that we may require a more appropriate weighting between text and color variation, and perhaps that the presence of text needs to be an explicit feature in the model. Note that the distributions of clutter rankings are quite skewed for a number of images, e.g. #18, 19.

Recommendations for drawing attention

In addition to providing a measure of display clutter, our methods can be used to suggest to a designer how to add a new attention-grabbing element to the display. Figure 4 shows examples in which local clutter maps suggested low-clutter candidate locations for a new element, and the local feature covariance suggested what colors of the new element would best draw attention. In principle, we could also use this to select element size and shape, by computing luminance contrast energy for a number of candidate sizes and shapes, and then choosing the most salient, given the local contrast variance. The resulting symbols that we have added seem reasonable, even though map #6, has a wide variety of colors, and map #24 already contains a large number of objects.

CONCLUSIONS AND FUTURE WORK

Extensive experimental work remains to be done, including examining displays other than maps, and comparisons of our clutter measure with task performance data. In addition, experiments are necessary to support better modeling of the effects of expertise and task on clutter, as well as better modeling of basic perceptual phenomena.

In its present implementation, our clutter measure uses only color and luminance contrast as features. It may require additional features, e.g. to deal properly with displays with a large amount of text. Previous work modeling search in text displays (e.g. [2715]) should prove invaluable. For example, Tullis [27] has shown that search in a text display is a function of the number and size of groups, and has demonstrated a method for extracting such structure from a display. Nygren [15], however, has pointed out that grouping, per se, does not help performance unless the groups are sorted so the user can search hierarchically; first finding the right group, then searching within that group. Thus, any measure of clutter will need to make assumptions about the usefulness of text groupings for hierarchical search. Our current measure implicitly understands about grouping by proximity and visual similarity; a display with a red group, a blue group, and a green group will be judged less cluttered than a display with one large multicolored group. However, the current implementation does not know about semantic groupings based either on similarity or on text structure such as headings.

We have suggested an operational definition of clutter as congestion: clutter is when it would be difficult to add a new salient item to a display. Our resulting Feature Congestion measure of clutter can take arbitrary image data as input. We tested this measure by comparing it to rankings of clutter in maps by human observers. Observers were fairly consistent in their rankings, and our measure was consistent with the human data. We have also demonstrated that this clutter measure could be used in an automated way to make design suggestions as to location and features for an item to draw attention.

ACKNOWLEDGMENTS

This work was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-1-0625.

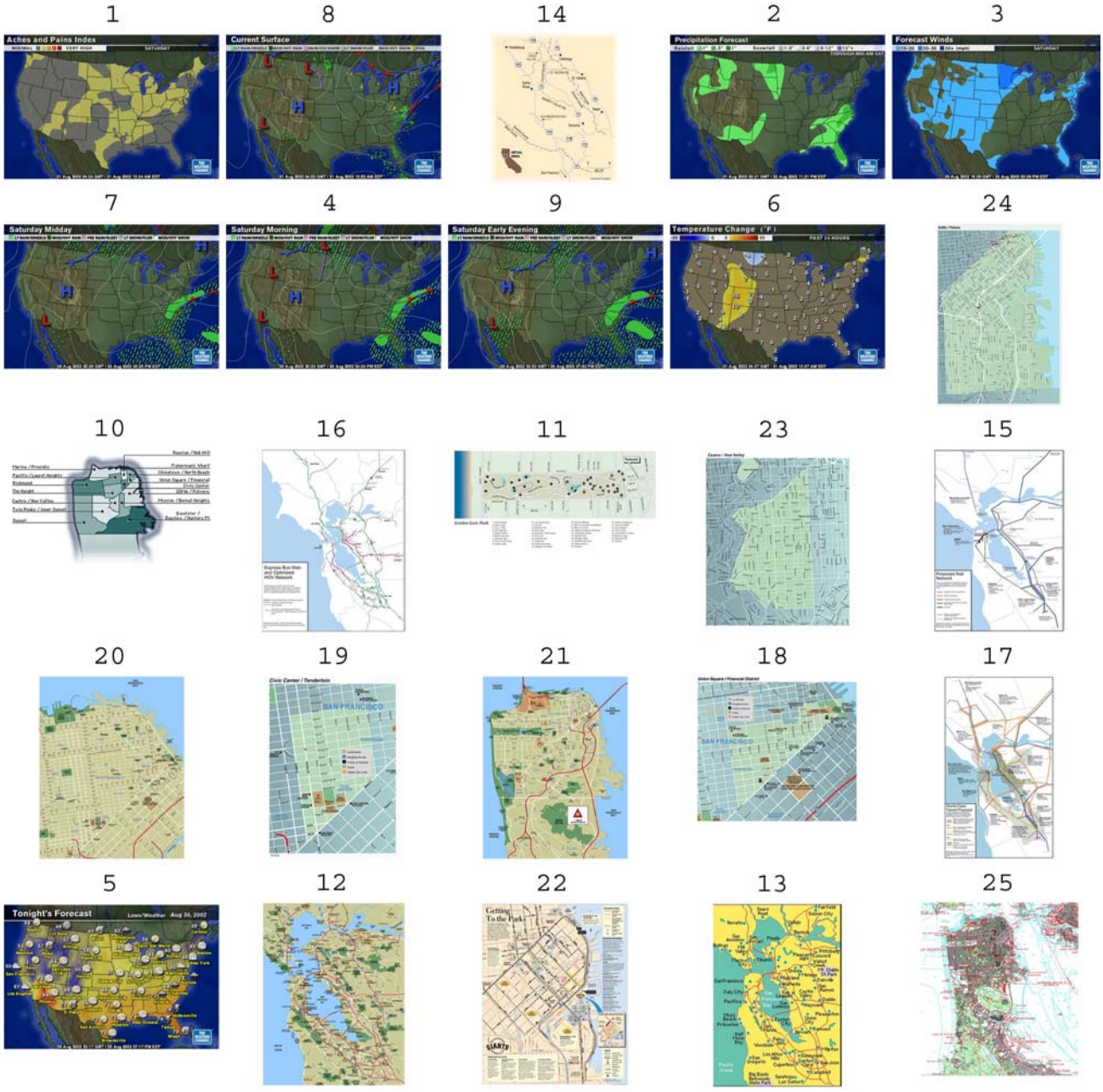

Figure 3. Map images in raster scan order according to their clutter ranking by the Feature Congestion clutter measure.

Figure 4. Maps with added elements designed (using the Statistical Saliency Model and Feature Congestion clutter measure) to draw attention. Left: pink & green symbols added. Right: Red and purple symbols added.

REFERENCES

1. Ahlberg, C. & Shneiderman, B. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proc. CHI 1994*, ACM Press (1994), 313-317.
2. Beyond the Click: Insights from Marketing Effectiveness Research (January 2001). http://www.dynamiclogic.com/na/research/btc/beyond_the_click_0102.html.
3. Burt, P., & Adelson, E.H. The Laplacian Pyramid as a Compact Image Code. *IEEE Trans. on Communication*, COM-31:532-540, 1983.
4. Buttenfield, B.P. & McMaster, R.B. (eds.). *Map Generalization: Making Rules for Knowledge Representation.* Longman, London, 1991.
5. Callaghan, T.C. Interference and domination in texture segregation: Hue, geometric form, and line orientation. *Perception & Psychophysics 46*, 4 (1989), 299--311.
6. C.I.E. Recommendations on uniform color spaces, color difference equations, psychometric color terms. *Supplement No.2 to CIE publication No.15 (E.-1.3.1) 1971/(TC-1.3.)* (1978).
7. Duncan, J. & Humphreys, G.W. Visual search and stimulus similarity. *Psychol. Rev., 96*, (1989), 433-458.
8. Eckstein, M.P., Thomas, J.P., Palmer, J, & Shimozaki, S.S. A signal detection model predicts the effects of set-size in visual search accuracy for feature, conjunction and disjunction displays, *Perception and Psychophysics,62*, 3 (2000), 425-451.
9. Fishkin, K. & Stone, M.C. Enhanced Dynamic Queries via Movable Filters. *Proc. CHI 1995*, ACM Press (1995), 415-420.
10. Frank, A.U. & Timpf, S. Multiple Representations for Cartographic Objects in a Multi-scale Tree – An Intelligent Graphical Zoom. *Comput. & Graphics, 18*, 6 (1994), 823-829.
11. Furnas, G.W. Generalized Fisheye Views. *Proc. CHI 1986*, ACM Press (1986), 16-23.
12. Itti, L., Koch, C., & Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 20*, 11, (1998), 1254-1259.
13. Miller, G.A. The Magical Number Seven, Plus or Minus Two (http://psychclassics.yorku.ca/Miller/). *Psychological Review, 63*, (1956), pp. 81-97.
14. Nickerson, J.V. Visual Programming. Ph.D. dissertation, New York University, New York, 1994. http://www.stevens-tech.edu/jnickerson/.
15. Nygren, E. & Allard, A. "Between the Clicks": Skilled Users Scanning of Pages. In *Designing for the Web: Empirical Studies, Human Factors and the Web.*. Sandia National Laboratories, Albuquerque, NM (March 1996).
16. Oliva, A., Mack, M.L., Shrestha, M., & Peeper, A. Identifying the Perceptual Dimensions of Visual Complexity of Scenes. Proc. 26th Annual Meeting of the Cognitive Science Society, (2004).
17. Palmer, J. Set-size Effects in Visual Search: the Effect of Attention is Independent of the Stimulus for Simple Tasks. *Vision Research, 34*, (1994), 1703-1721.
18. Perlin, K. & Fox, D. Pad: An Alternative Approach to the Computer Interface. *Proc. SIGGRAPH 1993*, ACM Press (1993), 57-64.
19. Perona, P. & Malik, J. Preattentive texture discrimination with early vision mechanisms. *JOSA(A) 7*, 5, (1990), 923-932.
20. Phillips, R.J. & Noyes, L. An Investigation of Visual Clutter in the Topographic Base of a Geological Map. *The Cartographic Journal, 19*, 2 (1982), 122-132.
21. Rosenholtz, R. Significantly different textures: A computational model of pre-attentive texture segmentation. *Proc. European Conference on Computer Vision*, Springer Verlag (2000), 197-211.
22. Rosenholtz, R. Search asymmetries? What search asymmetries? *Perception & Psychophysics, 63*, 3, (2001), 476-489.
23. Rosenholtz, R. Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal detection theory models of search. *J. Experimental Psychology, 27*, 4, (2001), 985-999.
24. Sellen, A.J. & Harper, R.H.R. *The Myth of the Paperless Office*. MIT Press, Cambridge, MA, 2003.
25. Springer, C.J. Retrieval of Information from Complex Alphanumeric Displays: Screen Formatting Variables' Effects on Target Identification Time. *Proc. 2nd Int'l Conf. on Human-Computer Interaction*, Elsevier Science Pub. (1987), 375-382.
26. Treisman, A.M., & Gelade, G. A feature-integration theory of attention. *Cog. Psych., 12*, (1980), 97-136.
27. Tullis, T.S. A Computer-Based Tool for Evaluating Alphanumeric Displays. *INTERACT '84*, B. Shackel (ed.), Elsevier Science (1985), 719-723.
28. Tufte, E.R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.
29. Watson, A.B. Visual detection of spatial contrast patterns: Evaluation of five simple models. *Opt. Express, 6*, (2000), 12-33, http://www.opticsexpress.org/abstract.cfm?URI=OPEX-6-1-12.
30. Wolfe, J.M. Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review, 1*, 2, (1994), 202-238.
31. Wolfe, J.M. Visual search. In H. Pashler, ed., *Attention*. University College London Press, London, U.K., 1998.
32. Woodruff, A., Landay, J., & Stonebraker, M. Constant Information Density in Zoomable Interfaces. *Proc. AVI 1998*, 57-65.