Modeling visual clutter perception using proto-object segmentation

Chen-Ping Yu
Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Dimitris Samaras
Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Gregory J. Zelinsky
Department of Psychology, Stony Brook University, Stony Brook, NY, USA
Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

We introduce the proto-object model of visual clutter perception. This unsupervised model segments an image into superpixels, then merges neighboring superpixels that share a common color cluster to obtain proto-objects—defined here as spatially extended regions of coherent features. Clutter is estimated by simply counting the number of proto-objects. We tested this model using 90 images of realistic scenes that were ranked by observers from least to most cluttered. Comparing this behaviorally obtained ranking to a ranking based on the model clutter estimates, we found a significant correlation between the two (Spearman's $\rho = 0.814$, $p < 0.001$). We also found that the proto-object model was highly robust to changes in its parameters and was generalizable to unseen images. We compared the proto-object model to six other models of clutter perception and demonstrated that it outperformed each, in some cases dramatically. Importantly, we also showed that the proto-object model was a better predictor of clutter perception than an actual count of the number of objects in the scenes, suggesting that the set size of a scene may be better described by proto-objects than objects. We conclude that the success of the proto-object model is due in part to its use of an intermediate level of visual representation—one between features and objects—and that this is evidence for the potential importance of a proto-object representation in many common visual percepts and tasks.

Introduction

Behavioral studies of visual clutter

Clutter is defined colloquially as "a crowded or disordered collection of things" (http://www.merriam-webster.com/dictionary/clutter). More operational definitions have also been proposed, defining clutter as "the state in which excess items, or their representation or organization, lead to a degradation of performance at some task" (Rosenholtz, Li, & Nakano, 2007; p. 3). Whatever definition one chooses, visual clutter is a perception that permeates our lives in an untold number of ways. It affects our ability to find things (e.g., Neider & Zelinsky, 2011), how products are marketed and sold to us (Pieters, Wedel, & Zhang, 2007), the efficiency in which we interact with devices (Stone, Fishkin, & Bier, 1994), and even whether we find displays aesthetically pleasing or not (Michailidou, Harper, & Bechhofer, 2008). For these reasons, clutter and its consequences have been actively researched over the past decade in fields as diverse as psychology and vision science, marketing, visualization, and interface design. The goal of this study is to apply techniques from computer vision to better quantify the behavioral perception of clutter, not only to make available clutter estimates to these widely varying domains but also to more fully understand this ubiquitous and important percept.

The effects of visual clutter have been studied most aggressively in the context of a search task, where several studies have shown that increasing clutter

Citation: Yu, C.-P., Samaras, D., & Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *Journal of Vision*, *14*(7):4, 1–16, http://www.journalofvision.org/content/14/7/4, doi:10.1167/14.7.4.

doi: 10.1167/14.7.4 Received December 14, 2013; published June 5, 2014 ISSN 1534-7362

Figure 1. What is the set size of these scenes? Although quantifying the number of objects in realistic scenes may be an ill-posed problem, can you make relative clutter judgments between these scenes?

negatively impacts the time taken to find a target in a scene (Mack & Oliva, 2004; Rosenholtz et al., 2007; Bravo & Farid, 2008; Henderson, Chanceaux, & Smith, 2009; van den Berg, Cornelissen, & Roerdink, 2009; Neider & Zelinsky, 2011).[1] Fueling this interest in clutter among visual search researchers is the *set size effect*—the finding that search performance degrades as objects are added to a display. Many hundreds of studies have been devoted to understanding set size effects (e.g., Wolfe, 1998), but the vast majority of these have been in the context of very simple displays consisting of well segmented objects. Quantifying set size in such displays is trivial—one need only count the number of objects. But how many objects are there in an image of a forest, or a city, or even a kitchen (Figure 1)? Is each tree or window a different object? What about each branch of a tree or each brick in a wall? It has even been argued that the goal of quantifying set size in a realistic scene is not only difficult, it is ill-conceived (Neider & Zelinsky, 2008). As the visual search community has moved over the past decade to more realistic scenes (Eckstein, 2011), it has therefore faced the prospect of abandoning its most cherished theoretical concept—the set size effect.

The quantification of visual clutter offers a potential solution to this problem. Given that search performance also degrades with increasing clutter (e.g., Henderson et al., 2009; Neider & Zelinsky, 2011), clutter has been proposed as a surrogate measure of the set size effect, one that can be applied to images of realistic scenes (Rosenholtz et al., 2007). The logic here is straightforward; if it is not possible to quantify the number of objects in a scene, find a correlate to set size that can be quantified and use it instead.

Although models of clutter will be reviewed in the following section, one of the earliest attempts to model visual clutter used edge density—the ratio of the number of edges in an image to the image size (Mack & Oliva, 2004). This edge density model was followed shortly after by the more elaborate feature congestion model, which estimates clutter in terms of the density of intensity, color, and texture features in an image (Rosenholtz et al., 2007). Despite other more recent modeling efforts (Bravo & Farid, 2008; Lohrenz, Trafton, Beck, & Gendron, 2009; van den Berg et al., 2009), the simplicity and early success of the feature congestion model, combined with the fact that the code needed to run the model was available for public download, led to its adoption as a clutter quantification benchmark by the community of visual clutter researchers.

The feature congestion model has been extensively evaluated in studies of visual clutter. Prominent among these was a study by Henderson et al. (2009), who measured the effect of visual clutter on search behavior in terms of manual and oculomotor dependent variables and using images of real-world scenes as stimuli, which marked a departure from previous work that used simpler chart and map stimuli. They found that increasing visual clutter indeed negatively impacted search performance, both in terms of longer search times and a less efficient direction of gaze to targets, thereby supporting the claim that clutter can be used as a surrogate measure of set size in real-world scenes. However, they also found that the feature congestion model was no better than a simpler measure of edge density in predicting this effect of visual clutter on search. Building on this work, Neider and Zelinsky (2011) sought again to quantify effects of clutter on manual and oculomotor search behavior, this time using scenes that were highly semantically related to each other (thereby ruling out semantic contributions to any observed clutter effects). They did this by using the game SimCity to obtain a database of city images that grew visually more cluttered over the course of game play. Their results largely replicated those from the earlier Henderson et al. study (2009), finding that edge density was at least as good as feature congestion in predicting the effect of clutter on search.

Computational models of clutter

In one of the earliest studies relating clutter to visual search, Bravo and Farid (2004) used "simple" stimuli, defined as objects composed of one material, and "compound" stimuli, defined as objects having two or more parts, and found an interesting interaction between

this manipulation and clutter. Search performance for simple and compound stimuli was roughly comparable when these objects were arranged into sparse search displays. However, when these objects were densely packed in displays, a condition that would likely be perceived as more cluttered, search efficiency was found to degrade significantly for the compound objects. This observation led to their quantification of clutter using a power law model (Bravo & Farid, 2008). This model uses a graph-based image segmentation method (Felzenszwalb & Huttenlocher, 2004) and a power law function having the form $y = cx^k$, where x is a smallest segment-size parameter. Setting the exponent k to -1.32, they find the best fitting c and use it as the clutter estimate for a given image. Using 160 "what's-in-your-bag" images (http://whatsinyourbag.com/), they reported a correlation of 0.62 between these clutter estimates and behavioral search time (Bravo & Farid, 2008).

As Bravo and Farid (2004) were conducting their seminal clutter experiments, Rosenholtz, Li, Mansfield, and Jin (2005) were developing their aforementioned feature congestion model of visual clutter. This influential model extracts color, luminance, and orientation information from an image, with color and luminance obtained after conversion to CIElab color space (Pauli, 1976) and orientation obtained by using orientation specific filters (Bergen & Landy, 1991) to compute oriented opponent energy. The local variance of these features, computed through a combination of linear filtering and nonlinear operations, is then used to build a three-dimensional ellipse. The volume of this ellipse therefore becomes a measure of feature variability in an image, which is used by the model as the clutter estimate—the larger the volume, the greater the clutter (Rosenholtz et al., 2007). Using a variety of map stimuli, they tested their model against the edge density model (Mack & Oliva, 2004) and found that both predicted search times reasonably well (experiment 1; $r = 0.75$ for feature congestion; $r = 0.83$ for edge density). However, when the target was defined by the contrast threshold needed to achieve a given level of search performance (experiment 2) the feature congestion model ($r = 0.93$) outperformed the edge density model ($r = 0.83$).

More recently, Lohrenz et al. (2009) proposed their C3 (Color-Cluster Clutter) model of clutter, which derives clutter estimates by combining color density with global saliency. Color density is computed by clustering into polygons those pixels that are similar in both location and color. Global saliency is computed by taking the weighted average of the distances between each of the color density clusters. They tested their model in two experiments: one using 58 displays depicting six categories of maps (airport terminal maps, flowcharts, road maps, subway maps, topographic charts, and weather maps) and another using 54 images of aeronautical charts. Behavioral clutter ratings were obtained for both stimulus sets. These behavioral ratings were found to correlate highly with clutter estimates from the C3 model ($r = 0.76$ and $r = 0.86$ in experiments 1 and 2, respectively), more so than correlations obtained from the feature congestion model ($r = 0.68$ and $r = 0.75$, respectively).

Another recent approach, the crowding model of visual clutter, focuses on the density of information in a display (van den Berg et al., 2009). Images are first converted to CIElab color space, then decomposed using oriented Gabors and a Gaussian pyramid (Burt & Adelson, 1983) to obtain color and luminance channels. The luminance channel of the image is then filtered with difference-of-Gaussian filters to obtain a contrast image, and all of the channels are post-processed with local averaging. It is this local averaging that is hypothesized to be the mechanism of crowding under this model. The channels are then "pooled" by taking a weighted average with respect to the center of the image, resulting in a progressive blurring radiating out from the image's center. Pooled results are compared to the original channels using a sliding window that computes the KL-divergence between the two, thereby quantifying the loss of information due to possible crowding, and this procedure is repeated over all scales and features and finally combined by taking a weighted sum to produce the final clutter score. They evaluated their model on the 25 map images used to test the original version of the feature congestion model (Rosenholtz et al., 2005) and found a comparable correlation with the behavioral ratings ($r = 0.84$; van den Berg, Roerdink, & Cornelissen, 2007).

Image segmentation and proto-objects

Motivating the study of clutter is the assumption that objects cannot be meaningfully segmented from images of arbitrary scenes, but is this true? The computer vision community has been working for decades on this problem and has made good progress. Of the hundreds of scholarly reports on this topic, the ones that are most relevant to the goal of quantifying the number of objects in a scene (i.e., obtaining a set size) are those that use an unsupervised analysis of an image that requires no prior training or knowledge of particular object classes. Among these methods, the most popular have been normalized cut (Shi & Malik, 2000), mean-shift image segmentation (Comaniciu & Meer, 2002), and a graph-based method developed by Felzenszwalb and Huttenlocher (2004). However, despite clear advances and an impressive level of success, these methods are still far from perfect. Crucially, these methods are typically evaluated against a ground truth of object segmentations obtained from

human raters (the Berkeley segmentation dataset; Martin, Fowlkes, Tal, & Malik, 2001; Arbelaez, Maire, Fowlkes, & Malik, 2011), which, as already discussed, is purely subjective and also imperfect. This reliance on a human ground truth means that image segmentation methods, regardless of how accurate they become, will not be able to answer the question of how many objects exist in a scene, as this answer ultimately depends on what people believe is, and is not, an object.

Recognizing the futility of obtaining objective and quantifiable counts of the objects in scenes, the approach taken by most existing models of clutter (reviewed above) has been to abandon the notion of objects entirely. The clearest example of this is the feature congestion model, which quantifies the feature variability in an image irrespective of any notion of an object. Abandoning objects altogether, however, seems to us an overly extreme conceptual movement in the opposite direction, and that there exists an alternative that finds a middle ground; rather than attempting to quantify clutter in terms of features or objects, attempt this quantification using something between the two—*proto-objects*.

The term proto-object, or pre-attentive object (Pylyshyn, 2001), was coined by Rensink and Enns (1995; 1998) and elaborated in later work by Rensink on coherence theory (Rensink, 2000). Coherence theory states that proto-objects are low-level representations of feature information computed automatically by the visual system over local regions of space, and that attention is the process that combines or groups these proto-objects to form objects. Under this view proto-objects are therefore the representations from which objects are built, with attention being the metaphorical hand that holds them together. Part of the appeal of proto-objects is that they are biologically plausible—requiring only the grouping of similar low-level features from neighboring regions of space. This is consistent with the integration of information over increasingly large regions of space as processing moves farther from the feature detectors found in V1 (Olshausen, Anderson, & Van Essen, 1993; see also Eckhorn et al., 1988).

Since their proposal, proto-objects have appeared as prominent components in several models of visual attention. Orabona, Metta, and Sandini (2007) proposed a model based on proto-objects that are segmented using blob detectors, operators that extract blobs using Difference-of-Gaussian (Collins, 2003) or Laplacian-of-Gaussian (Lindeberg, 1998) filters (Lindeberg, 1998; Collins, 2003), which are combined into a saliency map for their visual attention model. A similar approach was adopted by Wischnewski, Steil, Kehrer, and Schneider (2009), who proposed a model of visual attention that uses a color blob detector (Forssén, 2004) to form proto-objects. These proto-objects are then combined with the Theory of Visual Attention (TVA, Bundesen, 1990) to produce a priority map that captures both top-down and bottom-up contributions of attention, with the bottom-up contribution being the locally grouped features represented by proto-objects. Follow-up work has since extended this proto-object based model from static images to video, thereby demonstrating the generality of the approach (Wischnewski, Belardinelli, Schneider, & Steil, 2010).

The proto-object model of clutter perception

Underlying our approach is the assumption that, whereas quantifying and counting the number of objects in a scene is a futile effort, quantifying and counting proto-objects is not. We define proto-objects as coherent regions of locally similar features that can be used by the visual system to build perceptual objects. While conceptually related to other proto-object segmentation approaches reported in the behavioral vision literature, our approach differs from these in one key respect. Although previous approaches have used blob detectors to segment proto-objects from *saliency maps* (Walther & Koch, 2006; Hou & Zhang, 2007), bottom-up representations of feature contrast in an image (Itti, Koch, & Niebur, 1998; Itti & Koch, 2001), or applied color blob detectors directly to an image or video (Wischnewski et al., 2010), this reliance on blob detection likely results in only a rough approximation of the information used to create proto-objects. Blob detectors, by definition, constrain proto-objects to have an elliptical shape, and this loss of edge information might be expected to lower the precision of any segmentation. The necessary consequence of this is that approaches using blob detection will fail to capture the fine-grained spatial structure of irregularly shaped real-world objects. It would be preferable to extract proto-objects using methods that retain this spatial structure so as to better approximate the visual complexity of objects in our everyday world. For this we turn to image segmentation methods from computer vision.

We propose the proto-object model of clutter perception, which combines superpixel image segmentation with a clustering method (mean-shift, Comaniciu & Meer, 2002) to merge featurally similar superpixels into proto-objects. These methods from computer vision are well-suited to the goal of creating proto-objects, as they address directly the problem of grouping similar image pixels into larger contiguous regions of arbitrary shape. However, our proto-object model differs from standard image segmentation methods in one important respect. Standard methods aim to match extracted segments to a labeled ground truth segmentation of objects, as determined by human observers, where each segment corresponds to a complete and (hopefully) recognizable object. One example of this is the Berkeley Segmentation Dataset (Arbelaez et al., 2011), a currently

Figure 2. Left: one of the images used in this study. Right, top row: a SLIC superpixel segmentation using 200 (left) and 1,000 (right) seeds. Right, bottom row: an entropy rate superpixel segmentation using 200 (left) and 1,000 (right) seeds. Notice that the superpixels generated by SLIC are more compact and regular, whereas those generated by the entropy rate method have greater boundary adherence but are less regular.

popular benchmark against which image segmentation methods can be tested and their parameters tuned. However, proto-objects are the fragments from which objects are built, making these object-based ground truths not applicable. Nor is it reasonable to ask observers to reach down into their mid-level visual systems to perform a comparable labeling of proto-objects. For better or for worse, there exists no ground truth for proto-object segmentation that can be used to evaluate models or tune parameters. We therefore use as a ground truth behaviorally obtained rankings of image clutter and then determine how well our proto-object model, and the models of others, can predict these rankings. Our approach is therefore interdisciplinary, applying superpixel segmentation and clustering methods from computer vision to the task of modeling human clutter perception.

Methods

Computational

The proto-object model of clutter perception consists of two basic stages: A superpixel segmentation stage to obtain image fragments, followed by a clustering and merging stage to assemble these fragments into proto-objects. Given that proto-objects are then simply counted to estimate clutter, the core function of the model is therefore captured by these two stages, which are detailed in the following sections.

Superpixel segmentation

We define an image fragment as a set of pixels that share similar low-level color features in some color space, such as RGB, HSV, or CIElab. This makes an image fragment computationally equivalent to an image superpixel, an atomic region of an image containing pixels that are similar in some feature space, usually intensity or color (Veksler, Boykov, & Mehrani, 2010). Superpixels have become very popular as a preprocessing stage in many bottom-up image segmentation methods (Wang, Jia, Hua, Zhang, & Quan, 2008; Yang, Wright, Ma, & Sastry, 2008; Kappes, Speth, Andres, Reinelt, & Schn, 2011; Yu, Au, Tang, & Xu, 2011) and object detection methods (Endres & Hoiem, 2010; van de Sande, Uijlings, Gevers, & Smeulders, 2011) because they preserve the boundaries between groups of similar pixels. Boundary preservation is a desirable property as it enables object detection methods to be applied to oversegmented images (i.e., many fragments) rather than individual pixels, without fear of losing important edge information. The first and still very popular superpixel segmentation algorithm is normalized cut (Shi & Malik, 2000). This method takes an image and a single parameter value, the number of desired superpixels (k), and produces a segmentation by analyzing the eigenspace of the image's intensity values. However, because the run time of this method increases exponentially with image resolution, it is not suitable for the large images (e.g., 800×600) used in most behavioral experiments, including ours. We therefore experimented with two more recent and computationally efficient methods for superpixel segmentation, the SLIC superpixel (Achanta et al., 2012) and the entropy rate superpixel (Liu, Tuzel, Ramalingam, & Chellappa, 2011).

Figure 2 shows representative superpixel segmentations using these two methods. Both methods initially distribute "seeds" evenly over an input image, the number of which is specified by a user-supplied input parameter (k), and these seeds determine the number of superpixels that will be extracted from an image. The algorithms then iteratively grow each seed's pixel

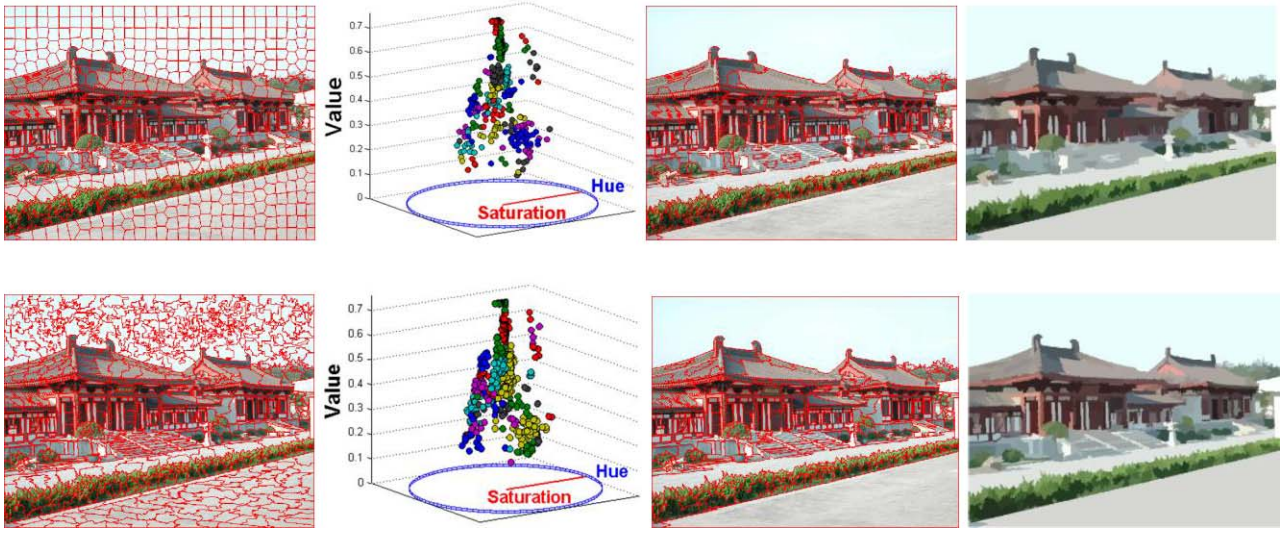

Figure 3. The computational procedure illustrated for a representative scene. Top row (left to right): a SLIC superpixel segmentation using $k = 600$ seeds; 51 clusters of median superpixel color using mean-shift (bandwidth = 4) in HSV color space; 209 proto-objects obtained after merging, normalized visual clutter score = 0.345; a visualization of the proto-object segmentation showing each proto-object filled with the median color from the corresponding pixels in the original image. Bottom row (left to right): an entropy rate superpixel segmentation using $k = 600$ seeds; 47 clusters of median superpixel color using mean-shift (bandwidth = 4) in HSV color space; 281 proto-objects obtained after merging, normalized visual clutter score = 0.468; a visualization of the proto-object segmentation showing each proto-object filled with the median color from the corresponding pixels in the original image.

coverage by maximizing an objective function that considers edge strengths and local affinity until all the seeds have converged to a stationary segment coverage. It is worth noting that this approach of oversegmenting an image also fragments large uniform areas into multiple superpixels, as multiple seeds would likely have been placed within such regions (e.g., the sky is segmented into multiple superpixels in Figure 3). Because superpixel segmentation is usually used as a preprocess, this oversegmentation is not normally a problem, although clearly it is problematic for the present purpose. More fundamentally, because the parameter k determines the number of superpixels that are created, and that the number of proto-objects will be used as our estimate of clutter, this user specification of k makes superpixel segmentation wholly inadequate as a direct method of proto-object creation and clutter estimation. For these reasons we therefore need a second clustering stage that uses feature similarity to merge these superpixel image fragments into coherent regions (proto-objects).

Superpixel clustering

To merge neighboring superpixels having similar features we perform a cluster analysis on the color feature space. Given the singular importance placed on color in this study, three different color spaces are explored: RGB, HSV, and CIElab. In this respect, our approach is related to the C3 clutter model, which groups similar pixels by spatial proximity and color similarity if they fall under a threshold (Lohrenz et al., 2009). However, our work differs from this previous model by using mean-shift (Cheng, 1995; Comaniciu & Meer, 2002) to find the color clusters in a given image, then assigning superpixels to one of these clusters based on the median color over the image fragment in the color feature space. We then merge adjacent superpixels (ones that share a boundary) falling within the same color cluster into a larger region, thereby forming a proto-object.

We should note that the mean-shift algorithm has itself been used as an image segmentation method (Comaniciu & Meer, 2002) and indeed is one of the methods that we evaluate in our comparative analysis. Mean-shift clusters data into an optimal number of groups by iteratively shifting every data point to a common density mode, with a bandwidth parameter determining the search area for the shift directions; the data that converge to the same density mode are considered to belong to the same cluster. This clustering algorithm has been applied to image segmentation by finding a density mode for every image pixel, then assigning pixels that converge to a common mode to the same cluster, again based on spatial proximity and color similarity. Doing this for all common modes results in a segmentation of pixels into coherent regions. Our approach differs from this standard application of mean-shift in that we use the algorithm, not for segmentation, but only for clustering. Specifically, mean-shift is applied solely to the

space of color medians in an image (i.e., using only the feature-space bandwidth parameter and not both the feature-space and spatial bandwidth parameters as in the original formulation of the algorithm), where each median corresponds to a superpixel, and it returns the optimal number of color clusters in this space. Having clustered the data, we then perform the above described assignment of superpixels to clusters, followed by merging, outside of the mean-shift segmentation method. By applying mean-shift at the level of superpixels, and by using our own merging method, we will show that our proto-object model is a better predictor of human clutter perception than standard mean-shift image segmentation.

Summary of the proto-object model

Figure 3 illustrates the key stages of the proto-object model of clutter perception, which can be summarized as follows:

1. Obtain superpixels for an image and find the median color for each. We will argue that our model is robust with respect to the specific superpixel segmentation method used and will show that the best results were obtained with entropy rate superpixels (Liu et al., 2011) using $k = 600$ initial seeds.
2. Apply mean-shift clustering to the color space defined by the superpixel medians to obtain the optimal number of color clusters in the feature space. We will again argue that our model is robust with respect to the specific color space that is used but that slightly better correlations with human clutter rankings were found using a bandwidth of four in an HSV color feature space.
3. Assign each superpixel to a color cluster based on the median color similarity and merge adjacent superpixels falling into the same cluster to create a proto-object segmentation.
4. Normalize the proto-object quantification between zero and one by dividing the final number of proto-objects computed for an image by the initial k number of superpixel seeds. Higher normalized values indicate more cluttered images.

Behavioral

Behavioral data collection was limited to the creation of a set of clutter ranked images. We did this out of concern that the previous image sets used to evaluate models were limited in various respects, especially in that some of these sets contained only a small number of images and that some scene types were disproportionately represented among these images—both factors that might severely reduce their generalization to realistic scenes. Specifically, these image sets were: 25 depictions of U.S. maps and weather forecasts (Rosenholtz et al., 2005); 58 depictions of six map categories in varying resolution, including airport terminal maps, flowcharts, road maps, subway maps, topographic charts, and weather maps (Lohrenz et al., 2009); 25 depictions of 6, 12, or 24 objects arranged into organized arrays (Bravo & Farid, 2004; van den Berg et al., 2009); 60 images of real-world scenes with embedded T and L stimuli (Henderson et al., 2009), 90 images of synthetic cities obtained from the game SimCity (Neider & Zelinsky, 2011), and 160 images depicting the contents of handbags (Bravo & Farid, 2008). We were also concerned that two different ground truths were used by these studies. Whereas some studies evaluated models against a ground truth of explicit clutter rankings of images (Rosenholtz et al., 2005; Lohrenz et al., 2009; van den Berg et al., 2009; Neider & Zelinsky, 2011), other studies adopted an indirect clutter ground truth consisting of manual and oculomotor measures of search efficiency (Rosenholtz et al., 2007; Bravo & Farid, 2008; Henderson et al., 2009; Beck, Lohrenz, & Trafton, 2010). Although a search-based ground truth is closer to the goal of using clutter as a surrogate measure of set size, it raises the possibility that these models were predicting something specific to the search behavior and were not estimating clutter perception per se.

Our goal in this study was to model human clutter perception, leaving the task of modeling the relationship between clutter and search to future work. To accomplish this we used a relatively large set of images of random-category scenes that were explicitly rank ordered by observers for visual clutter. We then determined how well models of clutter, including the proto-object model introduced here, could predict these clutter rankings.

Stimuli

Stimuli consisted of 90 800 × 600 images of real-world random-category scenes from the SUN09 image collection (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010). Image selection was restricted to those scenes for which there existed behaviorally obtained segmentations of the objects, and consequently, object counts. We did this so as to have another point of comparison for our evaluation—how well does the number of objects in a scene predict clutter perception, and how does this compare to predictions from the models? Object count was also used to ensure that selected scenes spanned a reasonable range of image complexity, at least with respect to the number of objects in the scenes. This was done by selecting images so as to fill six object count bins, with each bin having 15 scenes. Bin 1 contained images with object counts in the 1–10

Figure 4. Object segmentations from human observers for 4 of the 90 scenes used in this study. Segmentations were provided as part of the SUN09 image collection. To the right of each are lists of the object segment labels (object counts), in matching colors. Top left: three objects. Top right: 17 objects. Bottom left: 48 objects. Bottom right: 57 objects.

range, Bin 2 contained images with object counts in the 11–20 range, up to Bin 6 that contained images with object counts in the 51–60 range. Other than these constraints, and the resolution restriction (800 × 600 pixels), image selection was random from the targeted subset of SUN09 images. Given that the accuracy and detail of the object segmentations vary greatly in the SUN09 collection, selected images were also visually inspected to make certain that they were not under-segmented. Figure 4 shows these behaviorally obtained object segmentations for 4 of the 90 images used in this study.

Procedure

Fifteen undergraduate students from Stony Brook University (ages 18 to 30) participated for course credit. All had normal or corrected-to-normal vision, by self-report, and were naïve with respect to how their clutter judgments would be used. Participants were told that they would see 90 images, one at a time, and would have to rank order these images from least to most in visual clutter. No definition of clutter was suggested to participants; they were instead instructed to come up with their own definition and to use it consistently throughout their task. Participants were asked to rank order a 12-image list of practice scenes, so as to help them formulate their idiosyncratic clutter scale and to familiarize them with the interface used to assign these rankings (described below). Immediately following the practice block was the experimental block, in which participants had to rank order the 90 images selected using the above-described criteria. Each testing session, practice and experimental rankings combined, lasted between 60 and 90 min, and an experimenter remained in the room with participants so as to observe whether they were making thoughtful deliberations about where each image should be inserted into the rank-ordered clutter sequence.

Participants made their clutter rankings using an interface written in MATLAB and running on a Windows 7 PC with two LCD monitors. The monitors were arranged vertically, with one on top and the other below. The bottom monitor displayed the scene that was currently being ranked for clutter, which subtended a visual angle of approximately 27° horizontally and 20° vertically. Images were presented to each participant in a different randomized order so as to remove

potential bias. The top monitor displayed pairs of scenes that had already been ranked, with the less cluttered scene on the left and the more cluttered scene on the right. Importantly, the program kept an ongoing record of the images that had been rank ordered by the participant. This means that, although only two images were displayed at any one time on the top monitor, participants were able to scroll through the entire set of images that they had ranked when deciding where in this list the new image (displayed on the bottom monitor) should be inserted. Once an image was inserted into the rank ordered set, that image would be displayed on the top monitor (left position) and a new image would appear on the bottom monitor. This procedure repeated until all 90 images were ranked for clutter, with the leftmost image in the set having the lowest clutter ranking and the rightmost image having the highest. If at any time during the ranking procedure participants felt that they had made a mistake, or if their clutter scale simply evolved and they wanted to rerank the images, the opportunity existed for them to do so (and they were encouraged to avail themselves of this opportunity). This could be accomplished by selecting and removing an image from the rank-ordered set, then reinserting it in the new desired location.

Results and discussion

We evaluated the proto-object model of clutter perception in three ways. First, we evaluated the model against the ground truth behavior that it was intended to explain, in this case the clutter ranking of images from the behavioral task. This tells us whether the model is a valid measure of clutter perception. Second, we evaluated how the model's predictive power depends on implementation details and parameters—can it be easily broken? This tells us whether the model is robust. Third, we evaluated the proto-object model against other models of clutter. This tells us which model should be used when estimating clutter perception in the context of realistic visual scenes.

Evaluating the proto-object model against ground truth clutter perception

How well does the proto-object model predict clutter perception? Our use of a clutter ranking task enables this question to be answered by straightforward correlation. First we computed a Spearman's rank-order correlation (ρ) over all pairs of participants, then took the mean of these correlations to obtain a $\rho = 0.692$ ($p < 0.001$). This tells us that there was good agreement among our participants in their ranking of the 90 test images from least to most visually cluttered. This also tells us that this ranking constitutes a reliable ground truth for clutter perception in this task, against which the model's behavior can be meaningfully compared. To obtain a single ordering collapsed across participants we found the median of each image's ranked position in the ordered set, as medians are less sensitive to outliers than means and this was found to be preferable in previous work (Neider & Zelinsky, 2011). Ranking these medians from least to most cluttered gave us a single ordering of image clutter for comparison to our model. To obtain a comparable ordering of images from the proto-object model, we computed a proto-object segmentation for each of the 90 images (see Figure 5 for some examples), counted the number of proto-objects in each, normalized this count by dividing it by the number of initial superpixels, and then rank ordered these estimates of image clutter from least to most, paralleling the behavioral task. Correlating this ranking from the model with the median ranking from participants produced a Spearman's $\rho = 0.814$ ($p < 0.001$). This indicates that the proto-object model is a very good predictor of how behavioral participants rank order scenes for visual clutter; the scenes that were ranked as least (most) cluttered by participants tended also to be the scenes that were ranked as least (most) cluttered by the model (Figure 6). More generally, this also suggests that the proto-object model may be a good predictor of human clutter perception, at least in the context of random-category realistic scenes.

The previous analysis demonstrated good agreement between the rankings from the proto-object model and the median clutter rankings from our behavioral participants, but will the model's predictive success generalize to new scenes? Models of visual clutter have largely neglected this question of generalization. The approach has instead been to tune model parameters so as to find settings that best fit the behavioral data and to report this value as the model's performance.[2] Note, however, that this approach, one that we also adopted to produce Figures 5 and 6, really reflects only training accuracy and not true model prediction—true prediction requires generalization to unseen data that was not used to set model parameters.

Two practices exist to deal with this problem. One is to simply use different training and testing data sets. This, however, is not always practical in studies where the ground truth is based on behavior, as it would require conducting two behavioral experiments (one for tuning the model's parameters and the other for evaluating its performance). A second practice developed to address this problem is known as *cross validation*. Cross validation refers to the division of a single data set (e.g., the results of a single behavioral experiment) into separate training and testing subsets.

Figure 5. Representative examples of 3 of the 90 images used in this study (left column), shown with their corresponding proto-object segmentations (middle column) and reconstructions created by filling each proto-object with its median color (right column). Top row: Clutter score = 0.430 (ranked 41st). Middle row: Clutter score = 0.540 (ranked 65th). Bottom row: Clutter score = 0.692 (ranked 87th). Corresponding rankings from the behavioral participants were: 38th, 73rd, and 89th, respectively. Proto-object model simulations were based on entropy rate superpixel segmentations (Liu et al., 2011) using 600 initial seeds and a mean-shift clustering bandwidth of four within an HSV color feature space.

In the context of the present study, some proportion of the behaviorally ranked scenes would be used to tune the model's parameters, with this tuned model then used to predict clutter in the remaining "unseen" scenes. We performed 10-fold cross validation on our clutter-ranked images, meaning that 90% of the images were used for training and 10% were used for testing, and this procedure was repeated 10 times using different random splits. Averaging over these 10 tests produced a correlation of .74, which was predictably lower than the previously reported correlation of .81 (highlighting the importance of performing cross validation). More importantly, this correlation is still very high, and now indicates true prediction rather than training accuracy. This demonstrated generalization suggests that, as researchers conduct new experiments and obtain new data, the proto-object model will successfully predict clutter perception in these unseen data sets. Note also that this .74 correlation is closer to the .69 level of agreement found among the observers, which in some sense places an upper bound on prediction success. This finding suggests that the proto-object model was predicting clutter perception as well as could be expected given our observer judgments.

Evaluating the robustness of the proto-object model

Even with cross validation, the high correlation between model and behavioral clutter rankings reported in the previous section in some sense represents the best that the proto-object model could do over its parameter space, which includes the segmentation method used to obtain superpixels, the number of seeds used by this method, the color space used by the mean-shift algorithm, and the bandwidth parameter that largely determines mean-shift clustering. Specifically, the correlation of .814 was obtained using the entropy rate superpixel segmentation method (Liu et al., 2011),

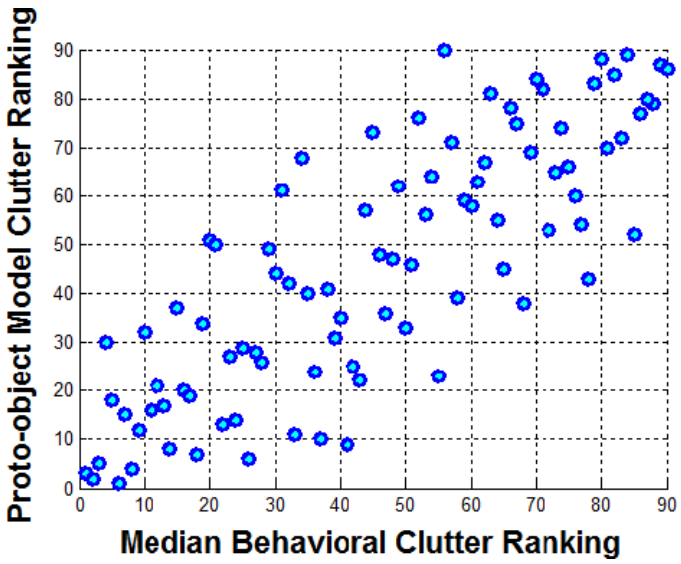

Figure 6. Clutter ranking of the 90 test scenes by the proto-object model plotted as a function of the median clutter ranking by our 15 behavioral participants for the same 90 scenes. Spearman's $\rho = 0.814$.

600 initial seeds, a HSV color space, and a mean-shift clustering bandwidth parameter of four in the feature space (settings used to produce Figures 5 and 6); would the proto-object model remain a good predictor of clutter perception if these parameters were changed?

Table 1 answers this question with respect to the choice of segmentation method, number of seeds, and color space. Reported is a $2 \times 3 \times 6$ matrix of correlations obtained by crossing two segmentation methods (entropy rate, Liu, et al., 2011, and SLIC, Achanta et al., 2012), three color spaces (RGB, HSV, and CIElab), and six seed initializations ($k = 200$, 400, 600, 800, 1000, 1200). An optimal mean-shift clustering bandwidth was computed separately for each of these 36 simulations. The largest difference in correlation between any two cells of this matrix was only .067, contrasting entropy rate superpixels using an HSV color space and 600 seeds with entropy rate superpixels using a CIElab color space and 200 seeds. Within any given dimension, these maximum differences were .067, .051, and .056 for segmentation method, color space, and seed number, respectively. The clear take-away message from this exploration is that, although varying segmentation method, color space, and seed initialization did affect the model's predictive success, our proto-object model was exceptionally robust to changes in these parameter values and performed extremely well across its parameter space. At its worst, the proto-object model still produced a respectable .747 correlation with the behavioral clutter rankings, which we will show in the following section to be comparable to the best performing competing clutter model without cross-validation.

Comparing models of clutter

So far we have shown that the proto-object model is a good and generalizable predictor of human clutter perception and that it is highly robust to changes in its parameters, but how does its performance compare to other models of visual clutter? To answer this question we implemented six methods of quantifying clutter and tested each against our 90 clutter-ranked images.[3] Among the visual clutter models reviewed in the Introduction, we downloaded the publicly available Matlab code for the feature congestion model (Rosenholtz et al., 2007), the power law model (Bravo & Farid, 2008), and the C3 model (Lohrenz et al., 2009; Beck et al., 2010) from the authors' websites and used these implementations for our evaluation. We also implemented a version of an edge density model using canny edge detection (Canny, 1986) with optimal settings to generate edge maps for each of our clutter-ranked images, from which we obtained clutter scores (Mack & Oliva, 2004). For the C3 model (Lohrenz et al., 2009; Beck et al., 2010), the authors kindly provided us with the Python scripts for implementation and information needed to optimize the model so as to obtain the best fit to our behavioral ground truth. The final two models, mean-shift (as described in Comaniciu & Meer, 2002) and a popular graph-based segmentation method (as described in Felzenszwalb & Huttenlocher, 2004), are not models of clutter per se but rather image segmentation methods that merge pixels into larger coherent regions. Like our proto-object model they can therefore easily be applied to the prediction of clutter perception. For these models we simply counted the number of merged segments for each of our images (as we did for the proto-object model), again following parameter optimization. More generally, for all of the models in our evaluation care was taken to explore their parameter spaces (when

# of superpixel seeds	Color space		
	RGB	HSV	CIElab
1200	0.792, 0.760	0.809, 0.796	0.792, 0.769
1000	0.784, 0.757	0.807, 0.795	0.798, 0.783
800	0.790, 0.750	0.806, 0.801	0.786, 0.782
600	0.800, 0.758	**0.814**, 0.812	0.780, 0.785
400	0.799, 0.777	0.813, 0.807	0.785, 0.777
200	0.771, 0.778	0.782, 0.797	0.747, 0.786

Table 1. Spearman's correlations between the proto-object model rankings and behavioral rankings as a function of color space, superpixel segmentation method, and the number of initial superpixel seeds. *Note:* Leftmost correlations were obtained using entropy rate superpixel segmentation; rightmost correlations were obtained using SLIC superpixel segmentation. All correlations used an optimized mean-shift spatial bandwidth parameter. The highest correlation is indicated in bold.

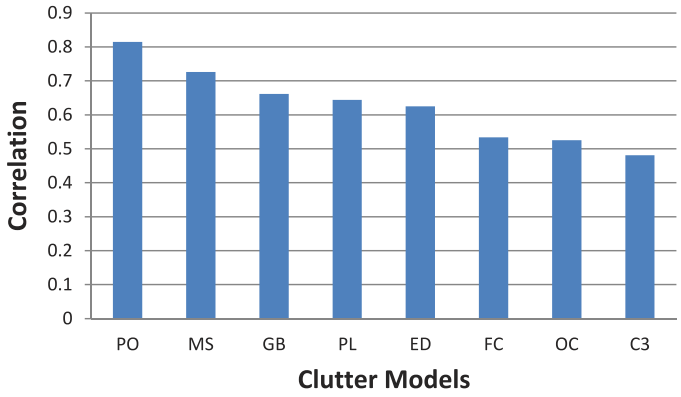

Figure 7. Spearman's correlations between the behaviorally obtained clutter ratings and ratings obtained from eight methods of predicting clutter, shown ordered from highest (left) to lowest (right). PO: Our proto-object clutter model. MS: Mean-shift image segmentation (Comaniciu & Meer, 2002). GB: Graph-based image segmentation (Felzenszwalb & Huttenlocher, 2004). PL: Power law clutter model (Bravo & Farid, 2008). ED: Edge density (Mack & Oliva, 2004). FC: Feature congestion clutter model (Rosenholtz et al., 2007). OC: Object counts provided by the SUN09 image collection (Xiao et al., 2010). C3: Color clustering clutter model (Lohrenz et al., 2009). All *ps* $<$ 0.001 except for C3, which was $p < 0.05$.

applicable) so as to find the settings that optimized prediction of the behavioral clutter rankings. Lastly, we evaluated the potential for behaviorally derived object segmentation counts, as provided by the SUN09 image collection, to predict the clutter rankings from our behavioral task. This evaluation is interesting as it speaks directly to the feasibility of using clutter as a surrogate measure of object set size in realistic scenes.

Figure 7 shows the results of this comparative analysis for the proto-object model and each of the six models tested, as well as for behavioral object counts. Plotted are Spearman's correlations between the behaviorally ranked images' median positions and clutter ratings obtained for each image using each of these methods. Our proto-object model outperformed the other methods of predicting clutter perception, and not by just a little. The second most predictive model, mean-shift image segmentation, produced a correlation of .726, well below the .814 correlation from the proto-object model. This is not to say that the other methods were poor predictors of clutter (indeed, they all predicted clutter better than chance, $p < 0.05$), just that our model was better.

Of all of the methods tested, those that used a superpixel segmentation worked best. In addition to the proto-object and mean-shift models, this included the graph-based model and, unsurprisingly, the power law model, which uses the same segmentation method as in the graph-based model. Simple edge density was the best non-superpixel-based predictor of clutter, followed after a steep drop in correlation by the feature congestion and C3 models.[4] It is worth noting that these less well-performing models were both tested originally on map and chart images that were very different from the realistic scenes used in the present study. It is therefore possible that the poor performance reported here is due to weak generalization to our image set. As for why our model was better able to predict the behavioral clutter rankings compared to the other superpixel-based methods (mean-shift, the graph-based model, and the power law model), we speculate that this is due to the level that each begins the grouping process that ultimately produces the regions of coherent features used to obtain the clutter estimates. The other models all started this grouping process at the level of pixels, and in this sense lacked the two-stage approach adopted by the proto-object model. In contrast, the proto-object model first obtained a superpixel segmentation for each image (Stage 1) and then grouped these superpixels to form proto-objects (Stage 2). The fact that this relatively subtle difference resulted in better behavioral clutter estimates suggests that midlevel human vision may adopt a similar multitiered approach; it may first represent local feature similarity (approximated by superpixels in our model) but then group these representations to form slightly higher level representations that we refer to as proto-objects.

Interestingly, behaviorally obtained object counts were a relatively poor predictor of clutter rankings in our task. Given that care was taken to select only those images that were well segmented into objects (see Figure 4), this finding is not likely due to poor estimates of object count. This suggests that caution should be exerted when attempting to infer clutter perception from the number of objects appearing in a scene, and vice versa. More fundamentally, it suggests that the goal of segmenting the objects in a scene may not only be ill-conceived, it may not even be useful, at least for clutter estimation—the perception of scene clutter may depend on clusters of low-level scene fragments rather than high-level counts of scene objects.

General discussion

To be sure, our visual world consists of features and objects, but there is also something in between—a level of visual representation consisting of proto-objects. Proto-objects are the fragments of our perception, and as such likely mediate many if not all of our everyday percepts and behaviors. In this study we focused on the perception of visual clutter. Clutter perception is interesting in that it colors all of our visual experience; we seem to know without thinking whether a scene is

cluttered or not, and this knowledge impacts our thoughts, our actions, and even our emotions. This largely automatic perception of clutter points to a fairly low-level process operating in the background of our consciousness, one that doesn't require extensive training, expectation, or experience with what is, or is not, a cluttered scene.

We modeled this process by borrowing techniques from computer vision to obtain an unsupervised segmentation of a scene into superpixels, then merged these superpixels based on shared color clusters to obtain what we refer to here as proto-objects—spatially extended regions of coherent features. This proto-object model estimates clutter perception as a simple count of the number of proto-objects extracted from an image, with a larger number predicting a more cluttered percept. We tested this model against a relatively large set of realistic scenes that were behaviorally ranked for visual clutter and found that it was highly successful in predicting this clutter ranking. We showed that this model was generalizable to unseen images and highly robust to changes in its parameters. It also outperformed, in some cases dramatically, all existing models of clutter, making it the new standard against which future models of clutter perception should be compared.

Future work will apply the proto-object model of clutter to a visual search task. The authors of previous clutter models have done this with the hope of using clutter estimates as a surrogate measure of the number of objects in a scene. Underlying this motivation is the assumption that if one were to actually know the number of objects in a scene that this object count would predict how cluttered that scene would appear (but see Rosenholtz et al., 2007). We directly tested this assumption and found that, while a measure of object count did predict clutter, it did so far less successfully than did the proto-object model, and indeed any model that used superpixel segmentation as a preprocess. This raises the possibility that the number of proto-objects, and not the number of actual objects, might be a better surrogate measure of search set size effects in realistic scenes. In addition to exploring further this possibility we will also ask how the distribution of fixations over a scene might be predicted by the distribution of proto-objects—would targets be found more quickly if they were embedded in regions of high proto-object density? This might be the case if the fixations made during a search or a free viewing task were biased to proto-object clusters.

In future work we will also seek to extend the proto-object model by considering features other than just color when merging superpixels into proto-objects. The features used in proto-object formation are still largely unknown, and indeed the concept of what a proto-object is, and is not, is currently evolving. The term *proto-object*, as originally formulated by Rensink and Enns (1995), relied heavily on a simple grouping of visual features into small clusters, each having a very limited, approximately 2° of visual angle, spatial extent. Although our proto-objects are not spatially constrained in the same way, and are able to grow and merge with neighboring image patches depending on their preattentive feature similarity, this early definition seems aligned most closely with the conception of proto-objects used in our model.[5] However, more recent usages of the term *proto-object* (Rensink, 2010) have assumed the additional contribution of 3-D features to obtain local estimates of scene structure, making these more complex proto-objects very different from the simpler entities proposed in the present study.[6] Starting with basic features (e.g., intensity, orientation, texture) and working up to more complex, we will explore how the systematic addition of features to the proto-object model affects its performance. This effort will give us a richer understanding of the midlevel vision features, and visual statistics computed from these features, that are useful for predicting clutter perception and potentially related visual behaviors (e.g., Franconeri, Bemis, & Alvarez, 2009).

Underlying these efforts is the belief that our perception of the world is a piecemeal construction; it starts with pixels and features, but these quickly become merged into locally (superpixel) and globally (proto-object) coherent regions, and, eventually, object parts and objects. The fact that superpixel-based approaches were found to outperform feature-based approaches in this study is telling and speaks to the potential importance of this intermediate proto-object level of visual representation. The theoretical implications of this observation are profound. It may be the case that the gap between pixels and objects is just too great and that robust computational methods for understanding object perception and detection can only emerge by considering the intermediate fragments of our perception.

Keywords: visual clutter, proto-objects, image segmentation, color clustering, superpixel merging, midlevel visual representation

Acknowledgments

This work was supported by NIH Grant R01-MH063748 to G. J. Z., and NSF Grant IIS-1111047 to G. J. Z. and D. S. We thank Thitapa Shinaprayoon for help with data collection and all the members of the Eye Cog lab for invaluable discussions.

Commercial relationships: none.
Corresponding author: Gregory J. Zelinsky.

Email: Gregory.Zelinsky@stonybrook.edu.
Address: Department of Psychology, Stony Brook University, Stony Brook, NY, USA.

Footnotes

[1]We will not discuss here the extensive literature on visual crowding (Whitney & Levi, 2011), despite its potential relationship to visual clutter (van den Berg et al., 2009). The crowding literature has been dominated by a flanker task and letter stimuli, neither of which was the focus of our study. Moreover, different modeling approaches have been adopted by the crowding and clutter literatures, one focused on segmented objects (crowding) and the other on images of realistic scenes (clutter).

[2]The publically available version of the feature congestion model (Rosenholtz et al., 2007) is an exception in this regard, as it has no parameters that are intended to be used for data fitting.

[3]Missing from this comparative evaluation is the crowding model (van den Berg et al., 2009). Because the code to run this model was not publicly available, we were forced to implement this model from scratch using only the information provided in the published report. However, our implementation produced very poor results when compared to the behavioral clutter rankings, $\rho = 0.162$ ($p = 0.13$), leading us to question whether our implementation was lacking in some respect. So as not to potentially mischaracterize this model's performance, we therefore chose not to include it in our evaluation.

[4]The C3 model (Lohrenz et al., 2009; Beck et al., 2010) was the poorest predictor of clutter perception using our image set, despite optimization. This surprised us, as the C3 model computes clutter in a qualitatively similar way as our proto-object model, attempting to group pixels with similar colors into polygons, then computing visual clutter as a function of the polygon pixel densities weighted by the inter-cluster saliency. However, the C3 model was designed and tested solely on chart and map stimuli, in contrast to the real-world images used in the present study. This suggests that the nonregular shapes of clusters resulting from realistic images might create problems for polygon-based clustering approaches (as opposed to superpixel-based approaches), such as the C3 model (Beck, Lohrenz, & Trenchard, personal communication, March 2013).

[5]Indeed, the early suggestion that proto-objects are spatially limited may have been due to the stimuli used in those initial experiments; a limited spatial extent may not be a defining property of a proto-object (Rensink, personal communication, January 29, 2014).

[6]Despite these differences, and more fundamentally, these evolving conceptualizations of proto-objects share the assumption that midlevel visual representations are highly volatile—quickly replaced by new proto-objects arising from new incoming visual information unless they are further bound by attention into objects. This volatility is captured in the current approach by the proposed relationship between proto-object formation and a process akin to superpixel segmentation, which is presumed to change fairly continuously with each new sample of visual information from the world. Understanding volatility as a necessary consequence of continuous segmentation is a potentially important reconstrual of this core property of proto-objects. As demonstrated by the present work, segmentation can be made computationally explicit and applied to images, meaning that the principles of proto-object formation identified using relatively simple stimuli might now be extended to visually complex natural scenes.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(11), 2274–2282.

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(5), 898–916.

Beck, M. R., Lohrenz, M. C., & Trafton, J. G. (2010). Measuring search efficiency in complex visual search tasks: Global and local clutter. *Journal of Experimental Psychology: Applied; Journal of Experimental Psychology: Applied, 16*(3), 238.

Bergen, J. R., & Landy, M. S. (1991). Computational modeling of visual texture segregation. In M. S. Landy & J. A. Movshon, (Eds.), *Computational models of visual processing* (pp. 253–271). Cambridge, MA: MIT Press.

Bravo, M. J., & Farid, H. (2008). A scale invariant measure of clutter. *Journal of Vision, 8*(1):23, 1–9, http://www.journalofvision.org/content/8/1/23, doi:10.1167/8.1.23. [PubMed] [Article]

Bravo, M. J., & Farid, H. (2004). Search for a category target in clutter. *Perception, 33,* 643–652.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547.

Burt, P., & Adelson, E. (1983). The Laplacian pyramid

as a compact image code. *Communications, IEEE Transactions on*, *31*(4), 532–540.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8,* 679–714.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *17*(8), 790–799.

Collins, R. T. (2003). Mean-shift blob tracking through scale space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Vol. *2*, pp. 234–240). Madison, WI: IEEE.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(5), 603–619.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboeck, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics, 60*(2), 121–130.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5):14, 1–36, http://www.journalofvision.org/content/11/5/14, doi:10.1167/11.5.14. [PubMed] [Article]

Endres, I., & Hoiem, D. (2010). Category independent object proposals. *Lecture Notes in Computer Science: Computer Vision–ECCV 2010, 6315,* 575–588.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Forssén, P. E. (2004). *Low and medium level vision using channel representations* (Unpublished doctoral dissertation), Linköping University, Linköping, Sweden.

Franconeri, S. L., Bemis, D. K., & Alvarez, G. A. (2009). Number estimation relies on a set of segmented objects. *Cognition, 113,* 1–13.

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision, 9*(1):32, 1–8, http://www.journalofvision.org/content/9/1/32, doi:10.1167/9.1.32. [PubMed] [Article]

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, CVPR'07. IEEE Conference on* (pp. 1–8). Minneapolis, MN: IEEE.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *20*(11), 1254–1259.

Kappes, J., Speth, M., Andres, B., Reinelt, G., & Schn, C. (2011). Globally optimal image partitioning by multicuts. *Lecture Notes in Computer Science: Energy Minimization Methods in Computer Vision and Pattern Recognition, 6819,* 31–44.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision, 30*(2), 79–116.

Liu, M. Y., Tuzel, O., Ramalingam, S., & Chellappa, R. (2011). Entropy rate superpixel segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 2097–2104). Colorado Springs, CO: IEEE.

Lohrenz, M. C., Trafton, J. G., Beck, M. R., & Gendron, M. L. (2009). A model of clutter for complex, multivariate geospatial displays. *Human Factors*, *51*(1), 90–101.

Mack, M. L., & Oliva, A. (2004). Computational estimation of visual complexity. Poster presented at the Twelfth Annual Object, Perception, Attention, and Memory Conference, Minneapolis, MN.

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference of Computer Vision (ICCV), IEEE Conference on* (pp. 416–423). IEEE.

Michailidou, E., Harper, S., & Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication* (pp. 215–224). Lisbon, Portugal: ACM.

Neider, M. B., & Zelinsky, G. J. (2011). Cutting through the clutter: Searching for targets in evolving complex scenes. *Journal of Vision, 11*(14):7, 1–16, http://www.journalofvision.org/content/11/14/7, doi:10.1167/11.14.7. [PubMed] [Article]

Neider, M. B., & Zelinsky, G. J. (2008). Exploring set size effects in scenes: Identifying the objects of search. *Visual Cognition*, *16*(1), 1–10.

Olshausen, B., Anderson, C., & Van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13,* 4700–4719.

Orabona, F., Metta, G., & Sandini, G. (2007). A proto-object based visual attention model. *Lecture Notes in Computer Science: Attention in Cognitive Sys-*

tems. Theories and Systems from an Interdisciplinary Viewpoint, 4840, 198–215.

Pauli, H. (1976). Proposed extension of the CIE recommendation on "Uniform color spaces, color spaces, and color difference equations, and metric color terms". *Journal of the Optical Society of America*, *66,* 866–867.

Pieters, R., Wedel, M., & Zhang, J. (2007). Optimal feature advertising design under competitive clutter. *Management Science*, *53*(11), 1815–1828.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1), 127–158.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7,* 17–42.

Rensink, R. A. (2010). Seeing seeing. *Psyche, 16*(1), 68–78.

Rensink, R. A., & Enns, J. T. (1998). Early completion of occluded objects. *Vision Research, 38,* 2489–2505.

Rensink, R. A., & Enns, J. T. (1995). Preemption effects in visual search: Evidence for low-level grouping. *Psychological Review, 102,* 101–130.

Rosenholtz, R., Li, Y., Mansfield, J., & Jin, Z. (2005). Feature congestion: A measure of display clutter. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 761–770). Portland, OR: ACM.

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

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(8), 888–905.

Stone, M. C., Fishkin, K., & Bier, E. A. (1994). The movable filter as a user interface tool. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems: Celebrating Interdependence* (pp. 306–312). Boston, MA: ACM.

van den Berg, R., Cornelissen, F. W., & Roerdink, J. B. (2009). A crowding model of visual clutter. *Journal of Vision, 9*(4):24, 1–11, http://www.journalofvision.org/content/9/4/24, doi:10.1167/9.4.24. [PubMed] [Article]

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, *7*(2):14, 1–11, http://www.journalofvision.org/content/7/2/14, doi:10.1167/7.2.14. [PubMed] [Article]

van de Sande, K. E., Uijlings, J. R., Gevers, T., & Smeulders, A. W. (2011). Segmentation as selective search for object recognition. In *International Conference on Computer Vision (ICCV)* (pp. 1879–1886). Barcelona, Spain: IEEE.

Veksler, O., Boykov, Y., & Mehrani, P. (2010). Superpixels and supervoxels in an energy optimization framework. *Lecture Notes in Computer Science: Computer Vision–ECCV 2010, 6315,* 211–224.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Wang, J., Jia, Y., Hua, X. S., Zhang, C., & Quan, L. (2008). Normalized tree partitioning for image segmentation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). Anchorage, AK: IEEE.

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168.

Wischnewski, M., Belardinelli, A., Schneider, W. X., & Steil, J. J. (2010). Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention. *Cognitive Computation*, *2*(4), 326–343.

Wischnewski, M., Steil, J. J., Kehrer, L., & Schneider, W. X. (2009). Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. *Human Centered Robot Systems, 6,* 93–102.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9,* 33–39.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (pp. 3485–3492). San Francisco, CA: IEEE.

Yang, A. Y., Wright, J., Ma, Y., & Sastry, S. S. (2008). Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, *110*(2), 212–225.

Yu, Z., Au, O. C., Tang, K., & Xu, C. (2011). Nonparametric density estimation on a graph: Learning framework, fast approximation and application in image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 2201–2208). Colorado Springs, CO: IEEE.