Psychological Review
2013, Vol. 120, No. 1, 85–109

0033-295X/13/$12.00 DOI: 10.1037/a0030779

A Probabilistic Model of Visual Working Memory: Incorporating Higher Order Regularities Into Working Memory Capacity Estimates

Timothy F. Brady and Joshua B. Tenenbaum
Massachusetts Institute of Technology

When remembering a real-world scene, people encode both detailed information about specific objects and higher order information like the overall gist of the scene. However, formal models of change detection, like those used to estimate visual working memory capacity, assume observers encode only a simple memory representation that includes no higher order structure and treats items independently from one another. We present a probabilistic model of change detection that attempts to bridge this gap by formalizing the role of perceptual organization and allowing for richer, more structured memory representations. Using either standard visual working memory displays or displays in which the items are purposefully arranged in patterns, we find that models that take into account perceptual grouping between items and the encoding of higher order summary information are necessary to account for human change detection performance. Considering the higher order structure of items in visual working memory will be critical for models to make useful predictions about observers' memory capacity and change detection abilities in simple displays as well as in more natural scenes.

Keywords: change detection, visual short-term memory, working memory, hierarchical Bayes, probabilistic model

Supplemental materials: http://dx.doi.org/10.1037/a0030779.supp

Working memory capacity constrains cognitive abilities in a wide variety of domains (Baddeley, 2000), and individual differences in this capacity predict differences in fluid intelligence, reading comprehension, and academic achievement (Alloway & Alloway, 2010; Daneman & Carpenter, 1980; Fukuda, Vogel, Mayr, & Awh, 2010). The architecture and limits of the working memory system have therefore been extensively studied, and many models have been developed to help explain the limits on our capacity to hold information actively in mind (e.g., Cowan, 2001; Miyake & Shah, 1999). In the domain of visual working memory, these models have grown particularly sophisticated and have been formalized in an attempt to derive measures of the capacity of the working memory system (Alvarez & Cavanagh, 2004; Bays, Catalao, & Husain, 2009; Cowan, 2001; Luck & Vogel, 1997; Wilken & Ma, 2004; Zhang & Luck, 2008). However, these models tend to focus on how observers encode independent objects from extremely simple displays of segmented geometric shapes.

By contrast to these simple displays, memory for real-world stimuli depends greatly on the background knowledge and principles of perceptual organization our visual system brings to bear on a particular stimulus. For example, when trying to remember real-world scenes, people encode a visual and semantic gist, plus detailed information about some specific objects (Hollingworth & Henderson, 2003; Oliva, 2005). Moreover, they use this gist to guide their choice of which specific objects to remember (Friedman, 1979; Hollingworth & Henderson, 2000), and when later trying to recall the details of the scene, they are influenced by this gist, tending to remember objects that are consistent with the scene but were not in fact present (Brewer & Treyens, 1981; Lampinen, Copeland, & Neuschatz, 2001; M. B. Miller & Gazzaniga, 1998).

In fact, even in simple displays, perceptual organization and background knowledge play a significant role in visual working memory. For example, what counts as a single object may not be straightforward, since even the segmentation of the display depends on our background knowledge about how often the items co-occur. For instance, after learning that pairs of colors often appear together, observers can encode nearly twice as many colors from the same displays (Brady, Konkle, & Alvarez, 2009). Displays where objects group together into perceptual units also result in better visual working memory performance, as though each unit in the group was encoded more easily (Woodman, Vecera, & Luck, 2003; Xu, 2006; Xu & Chun, 2007). Furthermore, observers are better able to recognize changes to displays if those changes alter some statistical summary of the display; for example, if a display is changed from mostly black squares to mostly white squares, observers notice this change more easily than a matched change that does not alter the global statistics (Victor & Conte, 2004; see also Alvarez & Oliva, 2009).

There is thus significant behavioral evidence that even in simple visual working memory displays, items are not treated independently (for a review, see Brady, Konkle, & Alvarez, 2011). However, existing formal models of the architecture and capacity of

This article was published Online First December 10, 2012.

Timothy F. Brady and Joshua B. Tenenbaum, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

See supplemental materials for data and MATLAB code.

Correspondence concerning this article should be addressed to Timothy F. Brady, now at Department of Psychology, Harvard University, 702 William James Hall, 33 Kirkland Street, Cambridge, MA 02138. E-mail: tfbrady@wjh.harvard.edu

visual working memory do not take into account the presence of such higher order structure and prior knowledge. Instead, they most often depend on calculating how many individual items observers remember if the items were treated independently.

Existing Models of Visual Working Memory Capacity

The most common paradigm for examining visual working memory capacity is a change detection task (e.g., Luck & Vogel, 1997; Pashler, 1988). In a typical change detection task, observers are presented with a study display consisting of some number *N* of colored squares (see Figure 1). The display then disappears, and a short time later another display reappears either that is identical to the study display or in which a single square has changed color. Observers must decide whether this test display is identical to the study display or whether there has been a change. Observers are told that at most a single item will change color.

The standard way of reporting performance in such a visual working memory task is to report the "number of colors remembered," often marked by the letter *K*. These values are calculated with a particular model of change detection (a "slot model"), which supposes that the decline in observers' performance when more squares must be remembered is caused solely by a hard limit in the number of items that can be remembered (Cowan, 2001; Pashler, 1988). Such estimates thus assume complete independence between the items.

For example, imagine that an observer is shown a display of *N* colored squares and afterward shown a single square, and asked whether it is the same as or different from the item that appeared at the same spatial location in the original display (Cowan, 2001). According to the slot model of change detection, if the observer encoded the item in memory, then the observer will get the question correct; and this will happen on *K*/*N* trials. For example, if the observer can encode three items and there are six on the display, on 50% of the trials the observer will have encoded the item that is tested and will get those 50% of trials correct. Such

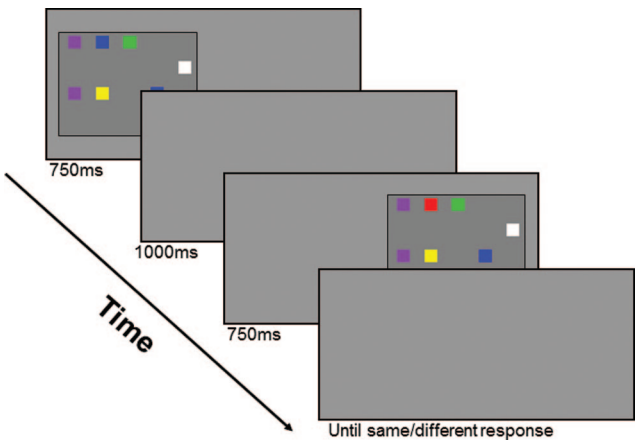

Figure 1. Methods of a change detection task (as used in Experiments 1 and 2). Observers are first briefly presented with a display (the study display) and then, after a blank, are presented with another display where either the items are the same or one item has changed color (the test display). They must say whether the two displays were the same or different.

models suppose no noise in the memory representation: If the item is encoded, it is remembered perfectly. On the other hand, if the observer does not encode the item in memory, then the model supposes that observers guess randomly (correctly choose same or different 50% of the time). Thus, the chance of getting a trial correct is

$$\mathrm{PC} = \frac{K}{N} * 100\% + \frac{(N-K)}{N} * 50\%. \quad (1)$$

By solving for *K*, we can take the percent correct at change detection for a given observer and determine how many items the observer remembered out of the *N* present on each trial (Cowan, 2001). Such modeling predicts reasonable values for a variety of simple displays (e.g., Cowan, 2001, 2005; Vogel, Woodman, & Luck, 2001), suggesting that observers have a roughly fixed capacity of three to four items, independent of a number of factors that affect percent correct (like set size, *N*).

However, nearly all visual working memory articles report such values, often without considering whether the model that underlies them is an accurate description of observers' working memory representation for their particular experimental stimuli. Thus, even in displays where observers perform grouping or encode summary statistics in addition to specific items, many researchers continue to report how many items observers can remember (*K* values) using the standard formula in which each item is treated as an independent unit (e.g., Brady, Konkle, & Alvarez, 2009; Xu & Chun, 2007). This results in *K* values that vary by condition, which would indicate a working memory capacity that is not fixed. In these cases, the model being used to compute capacity is almost certainly incorrect—observers are not encoding items independently.

In addition to the model underlying *K* values, other models have been used to quantify working memory capacity (e.g., Bays et al., 2009; Wilken & Ma, 2004; Zhang & Luck, 2008). However, these models also operate without taking into account the presence of higher order structure and prior knowledge, as they model displays that are sampled uniformly, limiting any overarching structure or gist. It is thus difficult to make claims about observers' capacities with such models. Due to the nature of the models, it is also difficult to expand existing models to account for summary representations, or representations of items that are not independent of one another.

Change Detection as Bayesian Inference

In this article we reformulate change detection as probabilistic inference in a generative model. We first formalize how observers encode an initial study display, and then we model the change detection task as an inference from the information about the test display and the information in memory to a decision about whether a change occurred. Modeling change detection in this Bayesian framework allows us to use more complex and structured knowledge in our memory encoding model (e.g., Hemmer & Steyvers, 2009; Tenenbaum, Griffiths, & Kemp, 2006), allowing us to make predictions about memory capacity under circumstances where items are nonindependent or where summary statistics are encoded in addition to specific items.

To understand the Bayesian model of change detection, it is useful to think of how it might apply to the simplest displays (like

those in Figure 1) with a standard, slot model representation (as in Cowan, 2001). To model this, we can assume that memory takes the form of a discrete number of slots, K, each of which stores which color was present on the display in a particular location. Also in line with standard slot models, we can initially assume that observers choose which K of the N items to encode at random. To model the change detection task in the Bayesian framework, we then consider how observers make a decision about whether there was a change when the test display is presented.

When observers must decide if there has been a change, they have access to all the items in the test display and to the items they encoded in memory from the study display. Using the information that at most a single item can change color between the two displays, the observer can perform an optimal inference to arrive at a judgment for whether the display has changed. In particular, the observer can place probabilities on how likely each possible display is to have been the study display, and then effectively rule out all possible displays that (a) are inconsistent with the items in memory or (b) have more than a single change from the test display. The observer can then arrive at a probability that indicates how likely it is that the study display was the same as the test display. Interestingly, this Bayesian model of change detection reconstructs the standard K slot model (Cowan, 2001; Pashler, 1988).

Importantly, however, by framing the model in terms of probabilistic inference, we make explicit the assumptions about the architecture of working memory the model entails. First, in such a model we are assuming that observers remember information about a specific subset K of the N items. Second, we are assuming that memory for these items is without noise. Both of these assumptions are simply properties of the probability distributions we choose and can be relaxed or generalized without changing the model architecture. Thus, the Bayesian framework we adapt allows a much greater range of memory architectures to be tested and made explicit.

The Current Experiments

In the current article we use such a Bayesian model of change detection to examine the use of higher order information in visual working memory. Although higher order information can take many forms, we begin with two possible representations: (a) a model that encodes both specific items and a summary, texture-like representation of the display (how likely neighboring items are to be the same color), and (b) a model in which observers first use basic principles of perceptual organization to "chunk" the display into perceptual units before encoding a fixed number of items. Furthermore, we consider whether observers might be using both of these representations on different trials or within a single trial. To examine whether such representations can account for human memory performance, we not only look at the overall level of performance achieved by using a particular memory representation in the model, but also examine how human performance varies from display to display.

In Experiments 1A and 1B, we test our proposed memory representations on displays where the items are purposefully arranged in patterns. In Experiment 2, we generalize these results to displays of randomly chosen colored squares (as in Luck & Vogel, 1997). We show for the first time that observers are highly consistent in which changes they find easy or difficult to detect, even in standard colored square displays. In addition, we show that models that have richer representations than simple slot or resource models provide good fits to the difficulty of individual displays, because these more structured models' representations capture which particular changes people are likely to detect. In fact, a model in which observers sometimes chunk a display using perceptual grouping and sometimes encode summary statistics (e.g., the texture of a display) seems to accurately account for a large part of the variance in observers' change detection performance. By contrast, the simpler models of change detection typically used in calculations of visual working memory capacity (e.g., the model underlying K values) do not predict any reliable differences in difficulty between displays. We conclude that even in simple visual working memory displays, items are not represented independently, and that models of working memory with richer representations are needed to understand observers' working memory capacity.

Experiments 1A and 1B: Patterned Dot Displays

Rather than being forced to treat each item as independent, our Bayesian model of change detection can be modified to take into account the influences of perceptual organization, summary statistics, and long-term knowledge. We thus had observers perform a memory task with displays where the items were arranged in spatial patterns. Observers are known to perform better on such displays than on displays without patterns (e.g., Garner, 1974; see also Hollingworth, Hyun, & Zhang, 2005; Phillips, 1974; Sebrechts & Garner, 1981). Because observers' memory representations in these displays are likely to be more complex than simple independent representations of items, such displays provide a test case for modeling higher order structure in visual working memory. To examine the generality of observers' memory representations, we used two similar sets of stimuli (Experiment 1A, red and blue circles; Experiment 1B, black and white squares), which vary basic visual properties of the stimuli but keep the same high-level grouping and object structure.

Method

Observers. One hundred thirty observers were recruited and run with Amazon Mechanical Turk (see Brady & Alvarez, 2011, for a validation of using Mechanical Turk for visual working memory studies). All were from the United States, gave informed consent, and were paid 30 cents for approximately 4 min of their time. Of the total observers, 65 participated in Experiment 1A and 65 in Experiment 1B.

Procedure. To examine human memory performance for patterned displays, we had observers perform a change detection task. We showed each of our observers the exact same set of 24 displays. Each display was presented to each observer in both a "same" and "different" trial, so observers completed 48 trials each. On each trial, the study display was presented for 750 ms, followed by a 1,000-ms blank period; then either an identical or a changed version of this original display was presented for 750 ms in a different screen location (the test display). Timing was controlled by JavaScript. Observers' task was simply to indicate, using a set of buttons labeled *same* and

different, whether the two displays were identical or whether there had been a change. The order of the 48 trials was randomly shuffled for each subject. Observers started each trial manually by clicking on a button labeled *Start this trial*, after which the trial began with a 500-ms delay.

Stimuli. Unlike traditional displays used to assess visual working memory capacity, we used displays where the items to be remembered were not simply colored squares in random locations but also exhibited some higher order structure (as in Phillips, 1974). For stimuli we created 24 displays that consisted of 5 × 5 patterns in which each space was filled in by a red or blue circle (Experiment 1A) or the same patterns were filled with black or white squares (Experiment 1B). The patterns could be anything from completely random to vertical or horizontal lines (see Figure 2). Our displays were thus simple relative to real scenes but were complex enough that we expected existing models, which encode independent items, would fail to predict what observers remember about these displays. Eight of the 24 displays were generated by randomly choosing the color of each dot. The other 16 were generated to explicitly contain patterns (for details of how we generated the patterned displays, see Appendix A). The changed versions of each display were created by taking the initial display and randomly flipping the color of a single item.

The displays each subtended 150 × 150 pixels inside a 400 × 180-pixel black (Experiment 1A) or gray (Experiment 1B) box. On each trial, the prechange display appeared on the left of the box, followed by the (potentially) changed version of the display on the right side of the box. Observers' monitor size and resolution was not controlled. However, all observers attested to the fact that the entire stimulus presentation box was visible on their monitor.

Results

For each display we computed a d', measuring how difficult it was to detect the change in that particular display (averaged across observers). We focus on d' because we are concerned in our modeling effort primarily with what representations might underlie performance, rather than what decision-making process observers use. Decision criteria should primarily affect response bias, and thus not impact the d'.

The stimuli in Experiments 1A were the same as those in Experiment 1B, except that the patterns were constructed out of red and blue dots in Experiment 1A and black and white squares in Experiment 1B. As expected, performance in Experiments 1A and 1B was highly similar: The correlation in the display-by-display d' was .91 between the two experiments. This suggests that observers' representations of these displays are invariant to the low-level properties of the stimuli (e.g., color, spatial frequency) as is typical of visual working memory (Luck, 2008). As a result, we collapsed performance across both experiments for the remaining analyses of d', though the results remain qualitatively the same when considering either experiment alone.

On average, human observers' d' was 2.18 (*SEM* ±0.06), suggesting that observers were quite good at detecting changes on these displays. The false-alarm rate ("same" trials to which observers said "different") was 10%, and the hit rate ("different" trials to which observers said "different") was 74%, suggesting that observers had a tendency to respond "same" more than "different." Since the displays each contain 25 dots, this d' corresponds to a K value of 17.8 dots if the items are assumed to be represented independently and with no summary information encoded (Pashler, 1988).

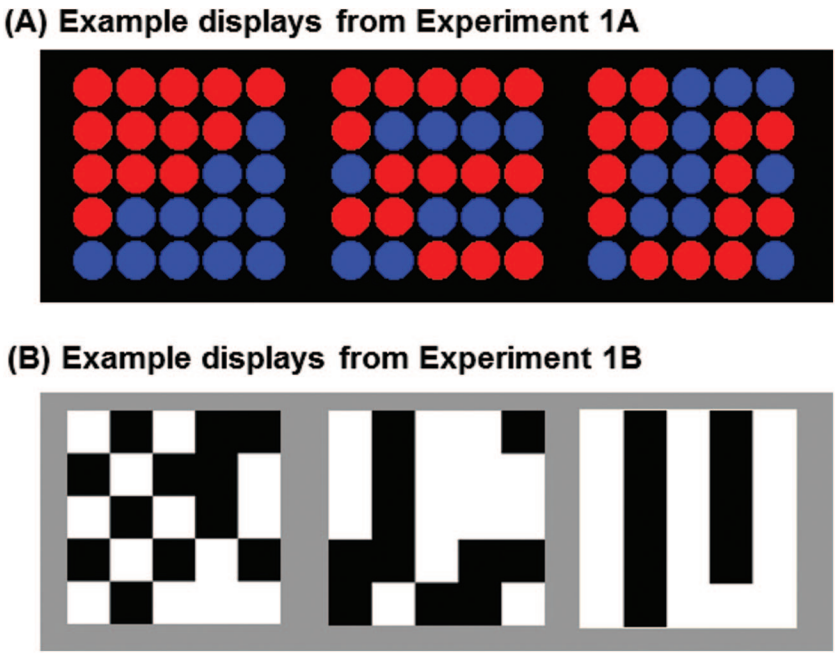

Figure 2. (A) Example study displays from Experiment 1A. (B) Example study displays from Experiment 1B. In both Experiments 1A and 1B, some displays were generated by randomly choosing each item's color, and some were generated to explicitly contain patterns.

In addition, observers were highly consistent in which displays they found most difficult to detect changes in (see Figure 3). We performed split-half analyses, computing the average d' for each display using the data from a randomly-selected half of our observers, and then comparing this to data from the other half of the observers. The same displays were difficult for both groups ($r = .89$, averaged over 500 random splits of the observers in both Experiments 1A and 1B; $p < .001$). Adjusting for the lower sample size of a split-half correlation using the Spearman-Brown formula gives a reliability estimate of the full sample of $r = .94$ (Kaplan & Saccuzzo, 2008).

Computing d' separately for each display and each observer is impossible, as each observer saw each display only once. Thus, to compute standard errors on d' on a display-by-display basis, we used bootstrapping; by resampling from the observers' reports on a particular display, we can get an estimate of the variance in our estimate of the d' for that display. This provides a visualization of the display-by-display consistency (see Figure 3). Some displays, like those on the left of Figure 3, are consistently hard for observers. Others, like those on the right of Figure 3, are consistently easy for observers to detect changes in.

Discussion

In Experiments 1A and 1B, we assessed observers' visual working memory capacity for structured displays of red and blue dots or black and white squares. We found multiple aspects of human performance in this task that conflict with the predictions of standard formal models of visual working memory.

First, we find that observers perform much better in detecting changes to these displays than existing working memory models would predict. Under existing formal models, in which items are assumed to be represented independently with no texture/summary information or perceptual grouping, observers' d' in this task would correspond to memory for nearly 18 dots (Pashler, 1988). This is nearly 5 times the number usually found in simpler displays (Cowan, 2001), and thus presents a direct challenge to existing methods of formalizing change detection and visual working memory capacity.

Furthermore, observers are reliable in which changes they find hard or easy to detect. This consistent difference between displays cannot be explained under any model in which observers treat the items independently. Previous formal models of change detection would treat all our displays as equivalent, since all displays change only a single item's color and all contain an equal number of items. They thus make no predictions regarding differences in difficulty across displays, or regarding which particular changes will be hard or easy to detect.

To account for the high level of performance overall and the consistent differences in performance between displays, it is necessary to posit a more complex memory representation or encoding strategy. We next consider two models for what information observers might be encoding in these patterned displays: a model in which observers encode both an overall summary of the texture of the display (e.g., "vertical lines") in addition to information about particular items, and a model in which observers chunk information by perceptually grouping dots of the same color into single units in working memory. In addition, we consider the hypothesis that observers may use both kinds of representations and combine them, either across trials or because different observers use different strategies. These models formalize particular hypotheses about what rep-

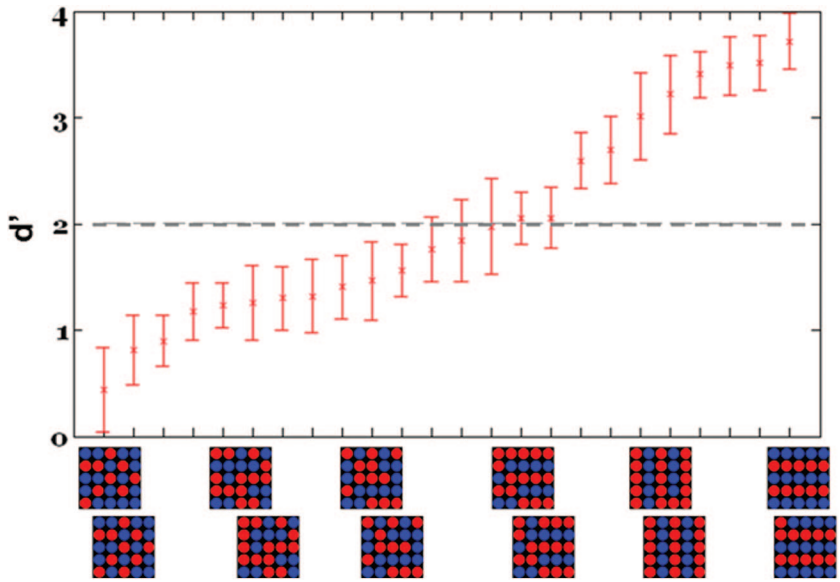

Figure 3. Consistency in which displays are most difficult in Experiment 1A. The *x*-axis contains each of the 24 display pairs, rank ordered by difficulty (lowest d' on the left, highest on the right; for visualization purposes, only a subset of display pairs is shown on the *x*-axis). The top display in each pair is the study display; the bottom is the test display with a single item changed. The dashed gray line corresponds to the mean d' across all displays. The error bars correspond to the standard error of the mean calculated by bootstrapping. The consistent differences in d' between displays indicate that some displays are more difficult than other displays.

resentations observers encode from these displays. They thus allow us to examine whether observers' performance is compatible with a fixed working memory capacity in terms of some format of representation other than a fixed number of independent items.

Summary-Based Encoding Model

In real-world scenes, observers encode not only information about specific objects but also information about the semantic gist of the scene (e.g., Lampinen et al., 2001; Oliva, 2005). In addition to this semantic information, observers encode diffuse visual summary information in the form of low-level ensemble statistics (or global texture) that they make use of even in simple displays of gabors or circles (Alvarez & Oliva, 2009; Brady & Alvarez, 2011). For example, in a landmark series of studies on summary statistics of sets, Ariely (2001) demonstrated that observers extract the mean size of items from a display and store it in memory even when they have little to no information about the size of the individual items on the display (Ariely, 2001; for reviews, see Alvarez, 2011; Haberman & Whitney, 2012). Observers seem to store not only summary information like mean size but also spatial summary information, like the amount of horizontal and vertical information on the top and bottom of the display (Alvarez & Oliva, 2009) and even high-level summary information like the mean gender and emotion of faces (Haberman & Whitney, 2007).

In other words, observers seem to store summary information, giving them a sense of the global texture of the entire display in addition to specific information about individual items (Haberman & Whitney, 2012). Furthermore, observers integrate this texture information with their representation of particular items. For example, Brady and Alvarez (2011) have shown that observers use the distribution of sizes of items on a display to modulate their representation of particular items from that display (e.g., if all items were small, they report items as smaller than they were). There is thus strong evidence that global texture representations are being stored and used in working memory tasks, even when observers are told to remember only individual items. This is contrary to the existing formal models of working memory, which assume independent representation of items.

To examine whether such summary representations could underlie the reliable differences in performance on our patterned displays, we built a model that formalized such a summary-based encoding strategy. We posited that observers might encode both a global summary of the display and particular "outlier" items that did not match this summary. Our modeling proceeded in two stages, mirroring the two stages of the change detection task: a model of how observers encode the study display and a model of how they decide if a change occurred once they have access to the test display.

More specifically, in the summary-based encoding model, we propose that observers use the information in the study display to do two things: First, they infer what global summary best describes the texture of the display; then, using this summary, they select the subset of the items that are the biggest outliers (e.g., least well captured by the summary) and encode these items specifically into an item-based memory. For a simplifying assumption, we use a summary representation based on a global texture representation (Markov random fields [MRFs]) that consists of just two parameters: one representing how likely a dot in this display is to be the same as or different from its horizontal neighbors and one representing how likely a dot is to be the same as or different from its vertical neighbors. This summary representation allows the model to encode how spatially smooth a display is both horizontally and vertically, thus allowing it to represent summaries or textures that are approximately equivalent to "vertical lines," "checkerboard," "large smooth regions," etc.

After a short viewing, the study display disappears and the observer is left with only what he or she encoded about it in memory. Then a test display appears and the observer must decide, based on what the observer has encoded in memory, whether this display is the same as the first display. Thus, at the time of the test display (the change detection stage), the observer has access to the test display and both the item-level and summary information from the study display that the observer encoded in memory. Under the constraint that at most one item will have changed, it is then possible to use Bayesian inference to put a probability on how likely it is that a given test display is the same as the study display and, with these probabilities, to calculate the likelihood that the display changed.

For example, an observer might encode that a particular display is relatively smooth (horizontal neighbors are similar to each other, and vertical neighbors are also similar to each other) but that the two items in the top right corner violate this assumption, and are red and blue, respectively. Then, when this observer sees the test display, the observer might recognize that although both items that he or she specifically encoded into an item memory are the same color they used to be, the display does not seem as smooth as it initially was: There are a number of dots that are not like their horizontal or vertical neighbors. This would lead the observer to believe there was a change, despite not having specifically noticed what items changed.

Importantly, when this model encodes no higher order structure, it recaptures the standard slot-based model of change detection. However, when the displays do have higher order regularities that can be captured by the models' summary representation, the model can use this information both to select appropriate individual items to remember and to infer properties of the display that are not specifically encoded.

Formal Specification of the Encoding Model

The graphical model representation of the encoding model (shown in Figure 4) specifies how the stimuli are initially encoded into memory. We observe the study display (D^1), and we use this both to infer the higher order structure that may have generated this display (G) and to choose the specific set of K items to remember from this display (S).

In the model, any given summary representation must specify which displays are probable and which are improbable under that summary. Unfortunately, even in simple displays like ours with only two color choices and 25 dots, there are 2^{25} possible displays. This makes creating a set of possible summary representations by hand and specifying the likelihood each summary gives to each of the 2^{25} displays infeasible. Thus, as a simplifying assumption, we chose to define the summary representation using MRFs, which allow us to specify a probability distribution over all images by simply defining a small number of parameters about how items tend to differ from their immediate neighbors. Such models have been used extensively in computer vision as models of images and textures (Geman & Geman, 1984; Li, 1995). We use only two summary parameters, which specify how often items are the same or different color than their hori-

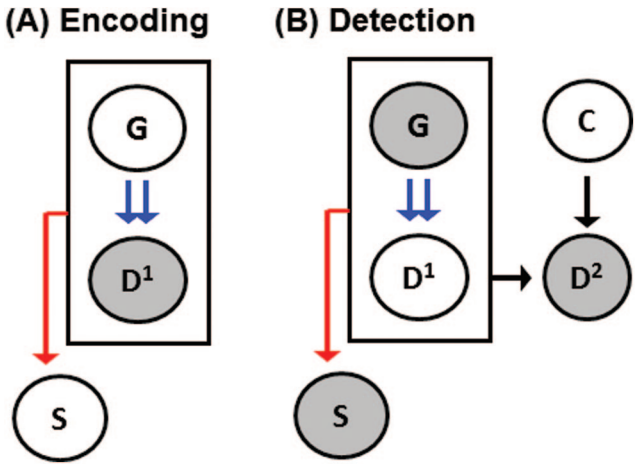

Figure 4. Graphical model notation for the summary-based encoding model at encoding (A) and detection (B). Shaded nodes are observed. The red arrows correspond to observers' memory encoding strategy; the black arrows correspond to constraints of the task (e.g., at most one dot will change between the study display (D^1) and test display (D^2)). The blue arrows correspond to our model of how a display is generated; in this case, how the summary or gist of a display relates to the particular items in that display. At encoding, we observe the study display (D^1), and we use this both to infer the higher order structure that may have generated this display (G) and to choose the specific set of K items to remember from this display (S). At detection, we have access to the summary we encoded in memory (G), the specific items we encoded (S), and the test display (D^2), and must infer what the study display looked like and thus whether the display changed (C) between D^1 and D^2.

zontal neighbors (G_h) and how often items are the same or different color than their vertical neighbors (G_v). Thus, one particular summary representation ($G_h = 1$, $G_v = -1$) might specify that horizontal neighbors tend to be alike but vertical neighbors tend to differ (e.g., the display looks like it has horizontal stripes in it). This summary representation would give high likelihood to displays that have many similar horizontal neighbors and few similar vertical neighbors.

We treat each item in these change detection displays as a random variable D_i^1, where the set of possible values of each D_i^1 is -1 (Color 1) or 1 (Color 2). To define the distribution over possible displays given the summary representation, $P(D \mid G)$, we assume that the color of each item is independent of the color of all other items when conditioned on its immediate horizontal and vertical neighbors.

We thus have two kinds of neighborhood relations (clique potentials) in our model. Our two parameters (G_h and G_v) apply only to cliques of horizontal and vertical neighbors in the lattice (N_h and N_v), respectively. Thus, $P(D^1|G)$ is defined as

$$P(D^1|G) = \frac{\exp\left(-En(D^1|G)\right)}{Z(G)} \tag{2}$$

$$En\left(D^1 \middle| G\right) = G_v \sum_{(i,j)\in N_v} \psi\left(D_i^1, D_j^1\right) + G_h \sum_{(i,j)\in N_h} \psi\left(D_i^1, D_j^1\right), \tag{3}$$

where the partition function

$$Z(G) = \sum_{D^1} \exp\left(-En(D^1|G)\right) \tag{4}$$

normalizes the distribution. $\psi(D_i^1, D_j^1)$ is 1 if $D_i^1 = D_j^1$ and -1 otherwise. If $G > 0$ the distribution will favor displays where neighbors tend to be similar colors, and if $G < 0$ the distribution will favor displays where neighbors tend to be different colors.

The summary representation of the display is therefore represented by the parameters G of an MRF defined over the display. Our definition of $p(D^1|G)$ thus defines the probability distribution p(display | summary). To complete the encoding model, we also need to define p(items | display, summary), $p(S \mid D^1, G)$. To do so, we define a probability distribution that preferentially encodes outlier objects (objects that do not fit well with the summary representation).

We choose whether to remember each object from the display by looking at the conditional probability of that object under the summary, assuming all its neighbors are fixed: $p(D_i^1|G, D_{\sim i}^1)$, where ~i means all items except i. S denotes the set of K specific objects encoded: $S = \{s_1, \ldots, s_k\}$. To choose S, we rank the K most unlikely objects and choose either 0, 1, 2, . . . or K of these objects based on how unlikely they are under the encoded summary representation. The probability of encoding a set of objects (S) from the set of the K most unlikely objects is

$$P(S|G, D^1) = \prod_{j:s_j\in S}\left(1 - p\left(D_j^1|G, D_{\sim j}^1\right)\right) \prod_{j:s_j\notin S} p\left(D_j^1|G, D_{\sim j}^1\right). \tag{5}$$

This defines $p(S|D^1, G)$, which provides the probability of encoding a particular set of specific items in a given display, p(items | display, summary), in our model. The model can encode the K outlier objects or, if there are fewer objects that are outliers (e.g., the display is perfectly predicted by a particular gist, as when it is perfectly smooth), can encode as few as zero specific objects.

To compute the model predictions, we use exact inference. However, due to the computational difficulty of inferring the entire posterior distribution on MRF parameters for a given display (e.g., the difficulty of computing $Z(G)$), and because we do not wish to reduce our summary representation to a single point estimate, we do not compute either the maximum posterior MRF parameters for a given display or the full posterior on G. Instead, we store the posterior in a grid of values for G in both horizontal and vertical directions ($G_h = -1.5, -1, -0.5, 0, 0.5, 1, 1.5$, $G_v = -1.5, -1, -0.5, 0, 0.5, 1, 1.5$). We compute the likelihood of the display under each of these combinations of G_h and G_v and then choose the items to store (S) by integrating over the different choices of G. We store the full posterior over S for each value of G. We choose a uniform prior on the summary representation (e.g., a uniform prior on MRF parameters G). For computational reasons we consider only the K objects that are least likely for a given display for inclusion in S, rather than examine all possible sets of objects.

In summary, to encode a display we first treat the display as an MRF. We then calculate the posterior on possible summary representations by calculating a posterior on G at various (prespecified) values of G. We then use this G and the study display to compute a posterior on which set of $\leq K$ items to encode into item memory (S). At the completion of encoding we have both a distribution on summary representations (G) and a distribution on

items to remember (S), and these are the values we maintain in memory for the detection stage.

Formal Specification of the Detection Stage of the Model

At the detection stage, we need to infer the probability of a change to the display. To do so, we attempt to recover the study display using only the information we have in memory and the information available in the test display. Thus, using the probabilistic model, we work backward through the encoding process, so that, for example, all the possible study displays that do not match the specific items we remembered are ruled out because we would not have encoded a dot as red if it were in fact blue.

More generally, to do this inference we must specify $P(D^1|S)$, $P(D^1|D^2)$, $P(D^1|G)$, $P(S|G, D^1)$. Almost all of these probabilities are calculated by simply inverting the model we use for encoding the display into memory initially with a uniform prior on possible study displays. Thus, $P(D^1|G)$ and $P(S|G, D^1)$ are given by the same equations described in the Encoding section.

Those probabilities not specified in the forward model represent aspects of the change detection task. Thus, $P(D^1|S)$ is a uniform distribution over study displays that are consistent with the items in memory and 0 for displays where one of the items in S differs from the corresponding item in D^1. This represents our simplifying assumption (common to formal versions of the standard slot model of visual working memory) that items in memory are stored without noise and are never forgotten (it is possible to add noise to these memory representations by making $P(D^1|S)$ a multinomial distribution over possible values of each item, or a normal distribution over some perceptual color space as in Zhang & Luck, 2008, but for simplicity we do not model such noise here). $P(D^1|D^2)$ is uniform distribution over all displays D^1 such that either $D^1 = D^2$ or at most one dot differs between D^1 and D^2. This represents the fact that the task instructions indicate at most one dot will change color.

Together these distributions specify the probability of a particular study display given the information we have about the test display and information we have in memory, $P(D^1|G, D^2, S)$. Given the one-to-one correspondence between first displays and possible changes, we can convert this distribution over first displays to a distribution over possible changes. Our prior on whether or not there is a change is 1/2, such that 50% of the mass is assigned to the "no change" display and the other 50% is split among all possible single changes. This allows us to calculate the posterior probability that there was a change in the display, which is also how often we expect observers to respond "change."

Modeling Results and Fit to Human Performance

In Experiment 1, we obtained data from a large number of human observers detecting particular changes in a set of 24 displays. For each display observers saw, we can use the summary-based encoding model to estimate how hard or easy it is for the model to detect the change in that display. The model provides an estimate, for a given change detection trial, of how likely it is that there was a change on that particular trial. By computing this probability for both a same trial and a change trial, we can derive a d' measure for each display in the model.

The model achieves the same overall performance as observers with a K value of only 4, thus encoding only four specific dots in addition to the display's summary (observers' $d' = 2.18$; model's $d' = 1.2, 1.8, 2.05, 2.25$ at $K = 1, 2, 3, 4$). This is because the model does not represent each dot independently: Instead, it represents both texture/summary information and information about specific dots.

Furthermore, this model correctly predicts which display observers will find easy and which displays observers will find difficult. Thus, the correlation between the model's d' for detecting changes in individual displays and the human performance on these displays is quite high ($r = .72$ with $K = 4$; averaging observers' results across Experiment 1A and 1B; see Figure 5). Importantly, this model has no free parameters other than how many specific items to remember, K, which we set to $K = 4$ based on the model's overall performance, not its ability to predict display-by-display difficulty. Thus, the model's simple summary representation captures which changes people are likely to detect and which they are likely to miss without any free parameters set to maximize this correlation.

Comparing the Model to Hit and False-Alarm Rates

We have focused on the d' values of the observers' and the model because we are primarily interested in whether our proposed representation can support observers' performance, rather than being interested in the decision-making process of observers. However, it is also worthwhile to investigate the decision process itself and ask whether our model provides a good fit to the raw hit and false-alarm rates of subjects, in addition to their sensitivity (d').

Interestingly, although d' values are highly correlated between Experiments 1A and 1B ($r = .91$), the response bias (c) is not as well correlated ($r = .46$). In part this reflects the lower reliability

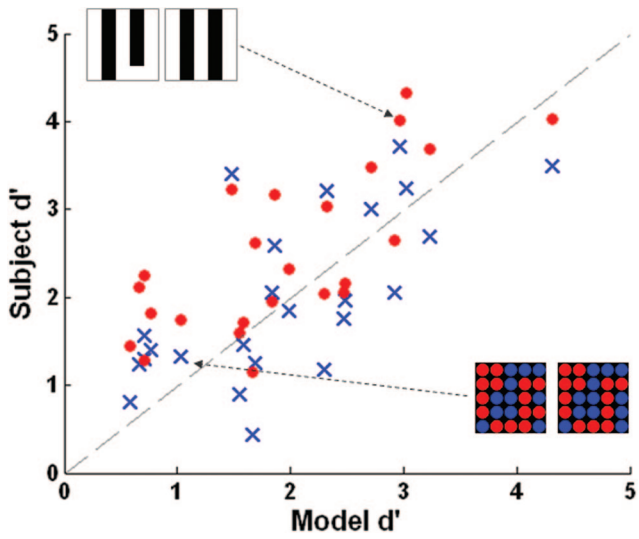

Figure 5. The fit of the summary-based encoding model with $K = 4$ to the observers' data for Experiments 1A (blue Xs) and 1B (red circles). Each point is the d' for a particular display. All the dots appearing on the diagonal would be a perfect fit. Example of both a hard and easy pair of displays is shown.

of the measure of response bias; it is considerably more variable across observers how conservative they are in reporting a change and on which particular displays they are more or less conservative even within an experiment (e.g., the average of 200 split-half correlations for Experiment 1A is $r = .43$, which, adjusted for having half the sample size, gives a within-experiment reliability estimate of only $r = .60$). However, in part this lower correlation between experiments in response bias is likely caused by the different propensity for false alarms and hits in the two experiments, both overall and on particular displays. On average, observers were significantly more likely to report "same" in Experiment 1A than Experiment 1B ($c = 0.41$ vs. $c = 0.15$, $p < .01$). This may be because the connected squares in Experiment 1B cause a larger transient with the appearance of the test display, as more pixels change relative to the background color. As a result of the different response biases in the two experiments, however, we report fits to hit and false-alarm rates separately for Experiments 1A and 1B.

In the model we set the prior probability of a change to 1/2, reflecting the design of the experiment, in which 1/2 of the displays observers see are changed between study and test. However, in the model, not only does this prior serve the role of specifying how likely observers are to believe the display changed a priori; it also effectively serves as a utility parameter, specifying how likely observers are to say "same" or "different" in general. This is because a greater prior on displays being "same" results in observers saying "same" more often in the posterior as well as the prior.

In the particular displays we tested, the default prior of 1/2 results in the model saying "change" slightly more than "same," with a response bias of $c = -0.12$ (hit rate: 88%; false-alarm rate: 18%). This is not in line with human performance, since in change detection tasks in general and our task in particular, observers have a propensity to say "same" unless they actually notice a change. Thus, to fit hits and false alarms separately, it is necessary to modify the prior in the model so that the model says "change" only when the display has a relatively high likelihood of having changed. To do so, we varied the prior as a free parameter in the model and fit this to the data by minimizing the sum of squared differences between the percent correct of the model and of observers (simultaneously for both the same and different displays).

In the case of Experiment 1A, the best fit parameter was a prior of saying "same" 82% of the time. This prior resulted in a hit rate of 80% and a false-alarm rate of 6%, with a correlation between the hit rate of the model and the observers of .50 and between the false alarms of the model and the observers of .59. This model did not perfec tly capture the data: The model significantly deviated from the number of correct responses of observers on same and different displays, $\chi^2(47) = 84.5$, $p < .01$. In Experiment 1B, the best fit parameter was a prior of saying "same" 71% of the time. This prior resulted in a hit rate of 83% and a false-alarm rate of 10%, with a correlation between the hit rate of the model and the observers of .40 and between the false alarms of the model and the observers of .56. Once again this model did not perfectly capture the data: The model significantly deviated from observers' reports, $\chi^2(47) = 82.0$, $p < .01$.

Thus, although this summary-based model captures the data fairly well, it does not capture the data perfectly; there are significant deviations between the model and observers' data. It is therefore useful to examine the specific successes and failures of the model. Thus, we plot the displays with the largest discrepancies between the model and the observers in Figure 6. In general, the summary-based model seems to overestimate how well observers perform on displays with many alternations between colors (Figure 6A) and underestimate performance on displays with large but abnormally shaped blocks of continuous color (Figure 6B). This suggests that although providing a reasonable model of observers' representations, the summary-based model also has systematic differences with the representations used by observers.

We will next consider what aspects of the summary-based model account for its successful fit to the data, and then will propose a different model, based on chunking or perceptual grouping, that might account for the failures of the summary-based model.

Necessity of the Summary Representation

The summary-based encoding model posits that observers encode a summary representation of the display and use this summary to choose outlier items to encode into a specific item memory. This model provides a surprisingly good fit to observers' data. However, it is possible that a single one of the processes used by the model might account for the fit to the data. For example, it is possible that simply choosing outlier items with a summary rep-

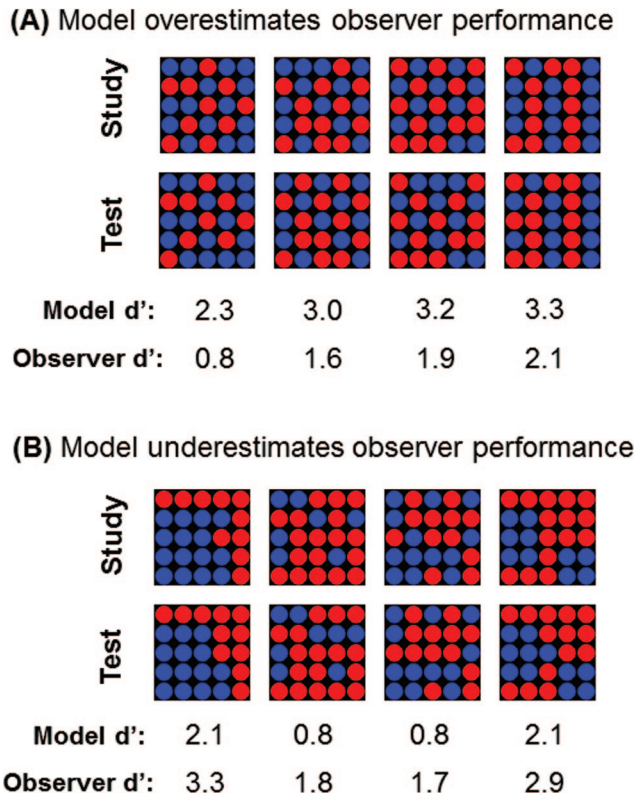

Figure 6. Largest discrepancies between the summary-based model and the observers. In general, the summary-based model seems to overestimate how well observers perform on displays with many alternations between colors (A) and underestimate performance on displays with large but abnormally shaped blocks of continuous color (B).

resentation but not encoding the actual summary representation into memory is sufficient to capture human performance. Alternatively, it is possible that simply encoding a summary representation but not using this representation to encode outlier items is sufficient to explain human performance. To address this and examine the necessity of each component of the model's representation, we "lesioned" the model by looking at model predictions without one or the other of these components.

Choosing outlier items but not remembering the summary representation. Is remembering the summary representation helping us to accurately model human performance, or can we predict human performance equally well by using the summary to choose outliers to encode into memory but then discarding the summary representation itself? Such a model might be appealing because it retains all the elements of the standard slot model (independent representations of K items), while modifying only the process by which these items are chosen by observers.

To examine whether such a model could fit the data, we looked at the fit of a summary-based model that did not have access to the summary representation at the time of change detection, and detected changes solely based on the specific objects encoded. Formally, this model was identical to the model described above, but without conditioning on G when doing change detection. Thus, detection was based only on the probabilities $P(D^1|S)$ and $P(D^1|D^2)$, which are again calculated by using the same equations as used in the encoding model.

We find that such a model does not fit human performance nearly as well as the full summary-based encoding model (see Figure 7A). First, to achieve human levels of performance, such a model must encode as many objects as a model that encodes objects completely at random (human levels of performance at $K = 18$; model $d' = 0.47, 0.92, 1.30, 1.69, 2.27$ at $K = 4, 8, 12, 16, 20$). Furthermore, this model does not accurately predict which specific changes will be noticed, either at $K = 4$ (correlation with d': $r = .30$) or at $K = 18$ ($r = .39$), accounting for at most 28% of the amount of the variance that is accounted for by the full model. In fact, directly comparing the correlation at $K = 4$ in the original model to $K = 4$ in this simplified model reveals that this model fits the data significantly less well than the full model ($z = 1.94$, $p = .05$).

One reason this model does not fit human performance as well as the full model is that it fails to recognize changes that introduce irregular items. For example, if the initial display is quite smooth and/or fits another summary representation very well and thus has no outliers, this model simply encodes items at random. Then, if the "change" display has an obvious outlier item, the model cannot detect it. To recognize this kind of change requires knowing what the summary of the initial display was.

Thus, it is not possible to fit the data with a standard slot-model-like representation, even allowing for the possibility that items are chosen based on a summary representation. This is in line with data that have more directly examined the summary/texture representations of observers and found that such representations are both encoded and used in working memory displays (Ariely, 2001; Brady & Alvarez, 2011; Haberman & Whitney, 2012).

Remembering a summary representation but choosing items at random. In addition to examining whether a model that does not encode a summary representation can fit the data, it is possible to examine a model that encodes both a summary of the display and specific items but does not choose which items to specifically encode by selecting outliers from the summary. Rather than preferentially encoding unlikely items, such a model chooses the items to encode at random. Examining this model is useful because such a representation would result if computing a summary/texture representation is a slow process (e.g., if it was computed by sampling individual items; Myczek & Simons, 2008). In particular, if observers did not have time to choose which items to attend to and selectively encode after computing the summary representation, they would likely end up encoding items nearly at random.

We use S to denote the set of K specific objects encoded: $S = \{s_1, \ldots, s_k\}$. In the full model, it is calculated by choosing objects that are outliers with respect to G. To lesion the model and encode

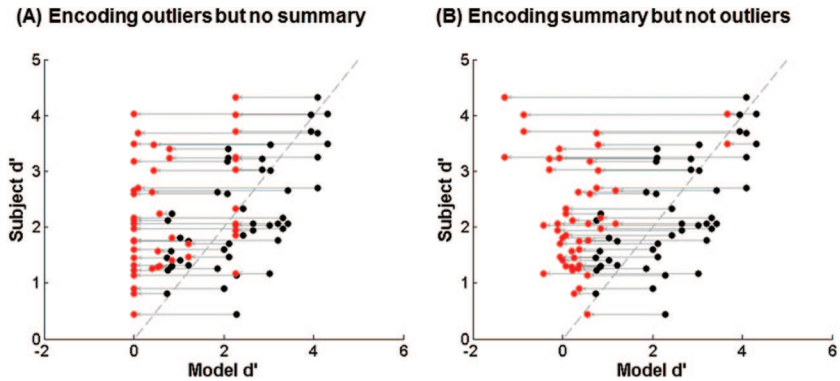

Figure 7. The fit of the model after being "lesioned." Black dots are the predictions of the complete model; red dots are the models' predictions after lesioning. If the red dots move away from the diagonal, this implies the model fits worse after being lesioned. (A) The fit of a model that did not have access to the summary representation at the time of change detection, and detected changes solely based on the specific objects encoded ($K = 4$). (B) The fit of a model that encodes both a summary of the display and specific items, but does not choose which items to specifically encode by selecting outliers from the summary, instead choosing specific items at random ($K = 4$). Removing either component from the model makes the fit considerably worse, suggesting that both are needed to fit human data.

objects at random, we instead choose S by simply sampling K of the N objects in the display at random.

We find that such a model does not fit human performance as well as the full summary-based encoding model (see Figure 7B). To achieve human levels of performance, such a model must encode as many objects as a model that encodes objects completely at random (human levels of performance at $K = 20$; $d' = 0.26, 0.54, 0.91, 1.39, 2.06$ at $K = 4, 8, 12, 16, 20$). Furthermore, it does not do a good job predicting which specific changes will be noticed, either at $K = 4$ (correlation with d': $r = .09$) or at $K = 20$ ($r = .40$), accounting for at most 31% of the variance that is accounted for by the full model. Furthermore, comparing the correlation at $K = 4$ in the original model to $K = 4$ in this simplified model reveals that this model fits the data significantly less well than the full model ($z = 2.6$, $p < .01$).

One reason this model fails to fit human performance is that it fails to recognize changes that remove irregular items. For example, if the initial display is quite smooth but has a single outlier, it will be encoded as a relatively smooth display. Then, if the "change" display removes the outlier item, the model cannot detect it. To recognize this kind of change requires maximizing your information about the first display by encoding specific items that are not well captured by the summary.

Thus, it is not possible to fit the data with a model that encodes the global texture of the display but chooses specific items to encode at random. Instead, a model that uses the summary of the display to guide which specific items to encode provides a significantly better fit. This is in line with data suggesting that summary representations are computed quickly and obligatorily (for reviews, see Alvarez, 2011; Haberman & Whitney, 2012).

Conclusion

Typically, we are forced to assume that observers are representing independent objects from a display in order to calculate observers' capacity. By using a Bayesian model that allows for more structured memory representations, we can calculate observers' memory capacity under the assumption that observers remember not just independent items but also a summary of the display. This model provides a reasonable estimate of the number of items observers are remembering, suggesting that only four specific items in addition to the summary representation must be maintained to match human performance. The model thus aligns with both previous work from visual working memory suggesting a capacity of three to four simple items (Cowan, 2001; Luck & Vogel, 1997) and data from the literature on real-world scenes and simple dot displays suggesting a hierarchical representation with both gist/summary information and item information (Brady & Alvarez, 2011; Lampinen et al., 2001).

Furthermore, because the summary-based model does not treat each item independently, and chooses which items to encode by making strategic decisions based on the display's summary, this model correctly predicts the difficulty of detecting particular changes in individual displays. By contrast, a model that assumes we encode each item in these displays as a separate unit and choose which to encode at random can predict none of the display-by-display variance. This model thus represents a significant step forward for formal models of change detection and visual working memory capacity.

Chunk-Based Encoding Model

Rather than encode both a summary of the display and specific items, it is possible that observers might use a chunk-based representation. For example, a large number of working memory models assume a fixed number of items can be encoded into working memory (Cowan, 2001; Luck & Vogel, 1997). To account for apparently disparate capacities for different kinds of information, such models generally appeal to the idea of chunking, first explicated by George Miller (1956). For example, Miller reported on work that found that observers could remember right decimal digits and approximately nine binary digits. By teaching observers to recode the binary digits into decimal (e.g., taking subsequent binary digits like 0011 and recoding them as 3), he was able to increase capacities up to nearly 40 binary digits. However, observers remembered these 40 digits using a strategy that required them to remember only seven to eight items (recoded decimal digits). Ericsson, Chase, and Faloon (1980) famously reported a similar case where a particular observer was able to increase his digit span from seven to 79 digits by recoding information about the digits into running times from various races he was familiar with, effectively converting the 79 digits into a small number of already existing codes in long-term memory. More recently, Cowan, Chen, and Rouder (2004) have found that by teaching observers associations between randomly chosen words in a cued-recall task, observers can be made to effectively treat a group of two formerly unrelated words as a single chunk in working memory, and that such chunking seems to maintain a fixed capacity in number of chunks even after learning.

In the domain of visual working memory, little work has explicitly examined chunking or what rules apply to grouping of items in visual working memory. In part, this is because visual working memory representations seem to be based particularly on objects and features, and so it may not be possible to recode them into alternative formats to increase capacity without using verbal working memory. However, some work has focused on how learning associations impacts which items are encoded into memory (Olson & Jiang, 2004; Olson, Jiang, & Moore, 2005) and which items are represented as a single chunk (Orbán, Fiser, Aslin, & Lengyel, 2008). Furthermore, it has been shown that learned associations can even result in greater numbers of individual items being encoded into memory (Brady, Konkle, & Alvarez, 2009). However, almost no work has formalized the rules behind which items are perceptually grouped and count as a single unit in a slot model of visual working memory (but see Woodman et al., 2003; Xu, 2006; and Xu & Chun, 2007, for examples of perceptual grouping influencing capacity estimates in visual working memory).

A simple hypothesis is that the basic Gestalt rules of perceptual grouping, in this case grouping by similarity (Koffka, 1935; Wertheimer, 1938), will determine the perceptual units that are treated as single units in visual working memory. Indeed, some work has attempted to examine how observers might group adjacent items of similar luminance together in order to remember more dots in displays much like the displays we use in the current task (Halberda, Simons, & Whetherhold, 2012; see also Hollingworth et al., 2005). However, little formal work has been done examining how well such a model accounts for human change

detection, or whether such a model predicts which displays will be easy or difficult to detect changes in.

To model such a chunking process, we added two components to our basic change detection model. First, rather than encoding K single objects, we encode up to K regions of a display. Second, to select these regions, we use two factors, corresponding to the Gestalt principles of proximity and similarity: (a) a spatial smoothness term that encourages the model to put only adjacent items into the same chunk and (b) a likelihood term that forces the model to put only items of the same color into the same chunk. We thus probabilistically segment the display into M regions and then select which K of these M regions to encode by preferentially encoding larger regions (where chance of encoding is proportional to region size; e.g., we are twice as likely to encode a region of four dots as a region of two dots). This allows us to examine how likely an observer who encoded a display in this way would be to detect particular changes for different values of K (see Figure 8 for a sample of possible region segmentations for a particular display).

In this model, we examine the possibility that observers use the information in the first display to form K regions of the display following the principles of proximity and similarity, and then encode the shape and color of these K regions into memory. Then the second display appears and the observer must decide, based on what he or she has encoded in memory, whether this display is the same as the first display. The observer does so by independently judging the likelihood of each dot in the second display, given the chunks the observer has encoded in memory.

Formal Model Specification

Our formalization of the chunk-based model has three stages. First, we compute a distribution over all possible ways of segmenting the study display into chunks, R. Then, for each value of R, we compute a distribution over all possible ways of choosing K chunks from R to encode into our chunk memory, S. Finally, we calculate how likely the display is to be the same for each possible value of R and each possible value of S given this R. Due to the huge number of possible values of R, we use Gibbs sampling to sample possible segmentations rather than doing a full enumeration. For any given segmentation R, however, we do a full enumeration of assignments of S and thus likelihoods of the display being the same or different.

To compute a distribution over R, we treat the chunk assignment of each item D_i^1 as a random variable R_i. Thus, R_i corresponds to which region D_i^1 is considered a part of, and each R_i can take on any value from 1 to 25 (the total number of items present in the display, and thus the maximum number of separate regions).We then compute a distribution over possible assignments of R_i using a prior that encourages smoothness (such that items D_i^1 that are either horizontal or vertical neighbors are likely to have the same region assignment), and using a likelihood function that is all-or-none, simply assigning 0 likelihood to any value of R where two items assigned the same chunk differ in color (e.g., likelihood 0 to any R where $R_i = R_j$, $D_i^1 \neq D_j^1$) and uniform likelihood to all other assignments of R.

We sample from R using Gibbs sampling. We thus start with a random assignment of values for each R_i, and then sample each R_i repeatedly from the distribution $p(R_i \mid R_{\sim i}, D^1)$ to generate samples from the distribution $p(R|D^1)$. $P(R)$ is calculated with a smoothness prior, where we once again make use of an MRF to prefer assignments of values to R where both horizontal and vertical neighbors are assigned the same value of R. This MRF has a single free parameter, Sm, corresponding to how strongly we prefer smoothness in the chunk assignments. Thus

$$P(R_i|R_{\sim i}) \propto \exp\left(-En(R_i|Sm)\right) \tag{6}$$

$$En\left(R_i|Sm\right) = Sm \sum_{(i,j)\in N} \Psi(R_i, R_j), \tag{7}$$

where, again, $\psi(R_i, R_j) = 1$ if $R_i = R_j$ and -1 otherwise.

For values of $Sm \gg 0$, we prefer larger chunks to smaller chunks, since we more strongly prefer neighboring items to have the same chunk label. The smoothness parameter thus affects how likely adjacent items of the same color are to end up in the same chunk. The model is relatively insensitive to the value of this parameter for values $\geq$1.0. For all simulations, we set this value to 4.0 because this provided a model that created different segmentations of the display fairly often, while still making those segmentations consist of relatively large chunks.

The likelihood function, $P(R|D^1)$, is simply defined such that all chunks or regions must have only a single color within them. Thus, if for any $R_i = R_j$, $D_i^1 \neq D_j^1$, then $P(R|D^1) = 0$, otherwise $P(R|D^1) = 1$. Taken together, this likelihood and the MRF smoothness prior specify the distribution over R.

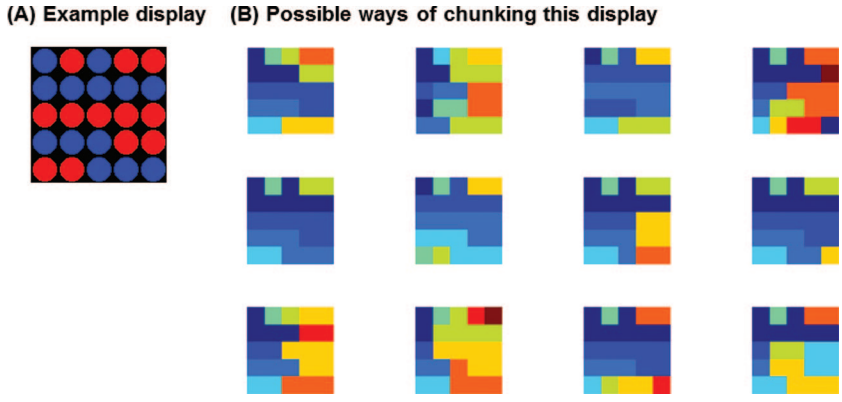

Figure 8. (A) An example display. (B) Several possible ways of chunking this display. These are 12 independent samples from our probabilistic chunking model with the smoothness parameter set to 4. Each color represents a particular chunk.

To compute a distribution over S for a given value of R, we enumerate how many unique chunk assignments are present in R (total number of chunks, M), labeling each of these chunks $L = 1, 2, \ldots M$. We then choose K chunks from this set of M possible chunks for our chunk memory, S, by choosing without replacement and giving each chunk label a chance of being chosen equal to the percentage of the items in the display that belong to that chunk. Thus

$$P(S|R, D^1) = \prod_{i.s_i \in S} \frac{\sum_{j=1\ldots25} (R_j = L_i)}{25} \prod_{i.s_i \notin S} \left(1 - \frac{\sum_{j=1\ldots25} (R_j = L_i)}{25}\right). \quad (8)$$

To calculate the chance of the display being the same given a value of R and S, we use the following logic (similar to Pashler, 1988). The set of items encoded is all the items assigned to any chunk that is encoded. Thus, if $D_i^1 \neq D_i^2$, and D_i is part of a chunk encoded in S, we notice the change 100% of the time. If no such change is detected, we expect the display to be the same in proportion to how many items we have encoded from the total set of items. Thus, the final probability of the display being the same (and thus of observers saying "same") is

$$P(C = 0|R, S) = \frac{1}{2} + \frac{1}{2} * \frac{\sum_{j=1\ldots25} (R_j \in S)}{25}. \quad (9)$$

Modeling Results and Fit to Human Performance

The chunk-based model provides an estimate, for a given change detection trial, of how likely it is that there was a change on that particular trial. By computing this probability for both the "same" trial and a "change" trial that observers saw in Experiment 1, we can derive a d' for each display in the model.

The model achieves the same performance as people with a K value of only 4, thus encoding only four chunks of dots (observers' $d' = 2.18$; model $d' = 0.44, 0.93, 1.49, 2.08, 2.69$ at $K = 1, 2, 3, 4, 5$). This is because the model does not represent each dot independently; instead, it represents grouped sets of dots as single chunks.

Furthermore, because the chunk-based model does not treat each item independently, the model makes predictions about the difficulty of detecting particular changes. In fact, the correlation between the model's difficulty with individual displays and the human performance on these displays was relatively high ($r = .58$; see Figure 9).

At $K = 4$, we can examine the effect of different values of the smoothness parameter on this correlation rather than simply setting this parameter to 4. We find that this correlation is relatively robust to the smoothness preference, with correlations with d' of $r = .35$, $r = .45$, $r = .45$, $r = .58$, $r = .58$ for values of 1, 2, 3, 4, and 5 (with smoothness = 5, the model nearly always segments the display into the largest possible chunks). Thus, the model's simple summary representation captures which changes people are likely to detect and which they are likely to miss independently of the settings of the chunk-size parameter.

In addition to their sensitivity (d'), it is also useful to investigate the fit to the hit and false-alarm rates of subjects. As in the summary-based model, fitting these values requires modifying the prior probability of a change in the model, such that the model reports a change only when there is relatively high evidence for a change.

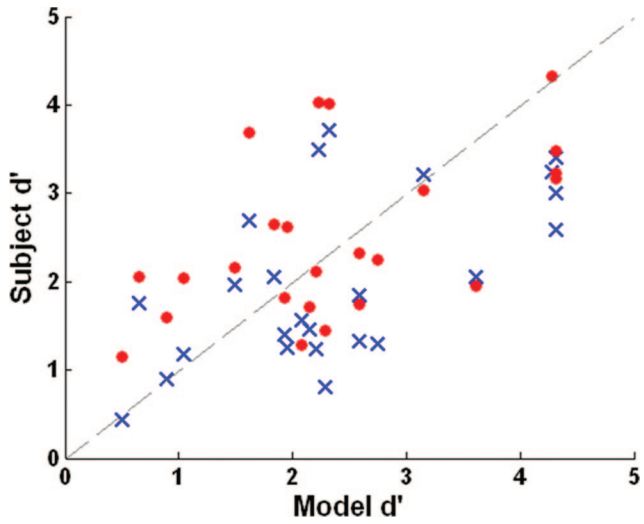

Figure 9. The fit of the chunk-based encoding model with $K = 4$ ($Sm = 4$) to the observers' data for Experiments 1A (blue Xs) and 1B (red circles). Each point is the d' for a particular display.

In the case of Experiment 1A, the best fit parameter was a prior of saying "same" 63% of the time. This prior resulted in a hit rate of 81% and a false-alarm rate of 9%, with a correlation between the hit rate of the model and the observers of .51 and between the false alarms of the model and the observers of .60. As in the case of d', this model did not perfectly capture the data: The model significantly deviated from observers' percent correct on same and different displays, $\chi^2(47) = 84.0$, $p < .01$. In Experiment 1B, the best fit parameter was a prior of saying "same" 69% of the time. This prior resulted in a hit rate of 81% and a false-alarm rate of 7%, with a correlation between the hit rate of the model and the observers of .51 and between the false alarms of the model and the observers of $-.06$.[1] Once again this model did not perfectly capture the data: The model significantly deviated from observers' reports, $\chi^2(47) = 80.1$, $p < .01$.

Thus, although the chunk-based model captures the data fairly well, it does not capture the data perfectly. It is therefore useful to examine the specific successes and failures of the model to examine what it is capturing and failing to capture about the data. The displays with the largest discrepancies between the model and the observers are shown in Figure 10. The chunk-based model seems to overestimate how well observers perform on displays with large but abnormally shaped blocks of continuous color (see Figure 10A) and underestimate performance on displays with clear structure but where the display is divided into a significant number of smaller blocks of color (Figure 10B).

Conclusion

The chunk-based model provides a reasonable estimate of the number of items observers are remembering, suggesting that only four chunks need be remembered to match human performance. The model thus provides evidence that fits with previous work

[1] Note that because of the small number of false alarms, fitting the hit rate is given much more weight by the least squares model fitting procedure.

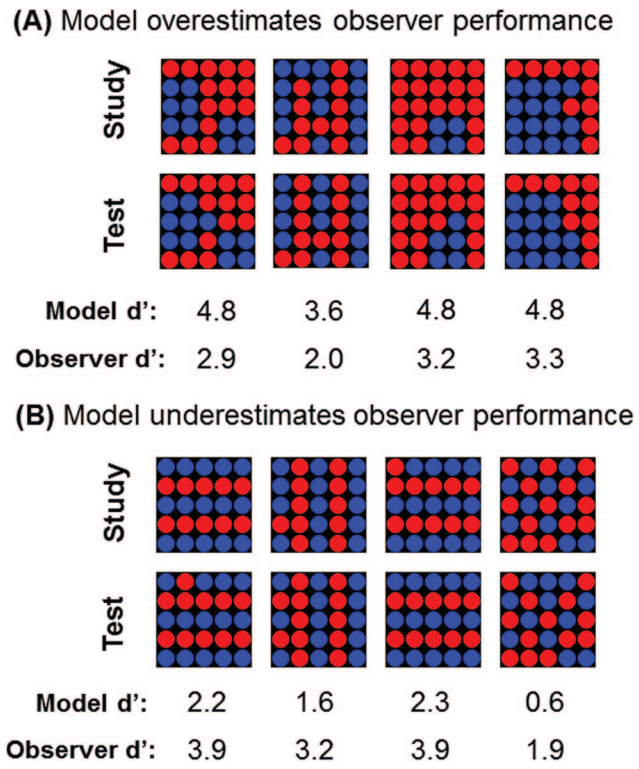

Figure 10. Largest discrepancies between the chunk-based model and the observers. In general, the chunk-based model seems to overestimate how well observers perform on displays with large but abnormally shaped blocks of continuous color (A) and underestimate performance on displays with clear structure but where the display is divided into a significant number of smaller blocks of color (B).

from visual working memory suggesting a capacity of three to four simple items (Cowan, 2001; Luck & Vogel, 1997), with the addition of a basic perceptual organization process that creates chunks before the items are encoded into memory. Furthermore, because the chunk-based model does not treat each item independently, this model makes predictions about the difficulty of detecting particular changes. These predictions coincide well with observers' difficulty in detecting particular changes in particular displays. Together with the summary-based encoding model, this chunk-based model thus provides a possible representation that might underlie human change detection performance in more structured displays.

Combining the Summary-Based and Chunk-Based Models

Both the chunking model and summary-based encoding model capture a significant amount of variance, explaining something about which displays observers find difficult to detect changes in. Do these models explain the same variance? Or do both models provide insight into what kinds of representations observers use to represent these displays? To assess this question, we examined whether combining these two models resulted in a better fit to the data than either model alone.

The summary-based encoding model and chunk-based model's display-by-display d' predictions are almost totally uncorrelated with each other ($r = .03$), despite both doing a reasonable job predicting which displays people will find difficult. In addition, the particular failures of the models are distinct and in some sense complementary (see Figures 6 and 10). It is thus possible that a combination of the two models might do much better at explaining observers' performance than either model alone. A combined model could be beneficial either because different observers tend to use different strategies or because different displays tend to be encoded according to a more chunk-like representation or a more summary-based representation.

The simplest way to examine whether the two models contribute separately to predicting human performance is to use a regression to test how well the display-by-display d' values predicted by the two models predict performance when used together. To do so, we did a linear regression, asking what the best weighting of the two models' predictions was to fit the human data. We found a best fit of $r = .92$, with weights of 0.67 for the summary-based encoding model and 0.45 for the chunk-based encoding model (intercept: -0.13). In fact, even a simple equal-weight averaging of the d' values of the two models resulted in a strong fit ($r = .90$), suggesting that the two models together can account for 81% of the variance in observers' d' across displays without any free parameters set to maximize this correlation (see Figure 11).

An alternative way to examine the contribution of both models is to make a probabilistic model of what decision process observers might undergo in encoding a particular display on a particular trial. To do so, we took the two existing probabilistic models and assumed that observers might sometimes choose to encode a display in a summary-based manner and sometimes choose to encode it in a chunk-based manner. We then probabilistically determined which model to use to encode the display on each trial

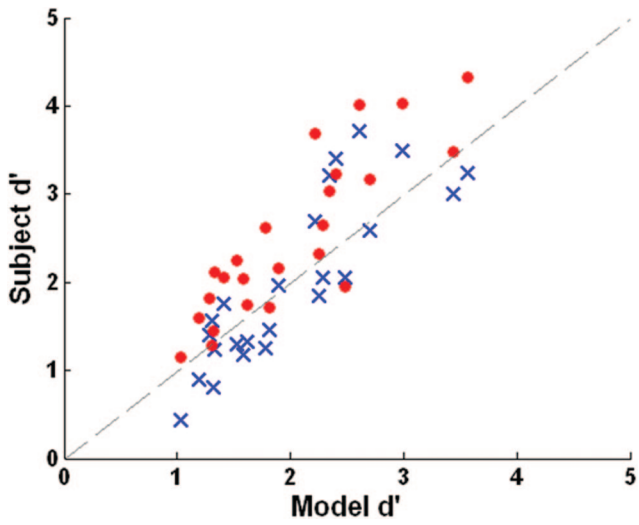

Figure 11. The fit of the combined model (a simple averaging of d' values of the two models) with $K = 4$ in both models to the observers' data for Experiments 1A (blue Xs) and 1B (red circles). Each point is the d' for a particular display. Combining the predictions of the summary-based and chunking models results in a much better fit to the human data than either model alone.

for each subject. This combined model consisted of a Bernoulli random variable M with weight m that determined whether a particular trial was encoded with the summary-based model ($M = 1$) or with the chunk-based model ($M = 0$). The best fit value of m was 0.66, compatible with the idea that 66% of individual trials within participants were encoded in a summary-based manner and the remainder in a chunk-based manner. The predictions of this combined model, converted to d', fit the human data well ($r = .87$).

Using such a model, we can also examine whether the improvement that results from combining the models is better explained by different observers being more likely to use one encoding strategy or another, or whether the benefit of combining the models is more likely to derive from different displays being encoded with different strategies. To do so, we computed a separate value of m, the preference for one model or the other, for each observer. We found a mean value of m of 0.51 across observers, with a standard deviation of 0.15 and a range of 0.33–0.67. Though substantial, this range suggests that every observer used both a summary-based and chunk-based strategy at least one third of the time. This suggests that a large part of the benefit of combining the models comes from variance in strategy within observers.

In addition to d', it is also useful to investigate the fit to the hit and false-alarm rates of subjects. To do so, we combined the models, as above, but using the versions of the summary-based and chunk-based models with the prior probabilities of a change as free parameters. In the case of Experiment 1A, the best fit parameter in the combined model was a prior of saying "same" 83% of the time. This prior resulted in a hit rate of 72% and a false-alarm rate of 7%, with a correlation between the hit rate of the model and the observers of .75 and between the false alarms of the model and the observers of .79. This model did not significantly differ from the data, $\chi^2(47) = 38.9$, $p = .79$. This suggests that the combined model provides a sufficient explanation of the data: Even with a large amount of data (65 observers), we cannot reject the hypothesis that the model represents human performance accurately.

In Experiment 1B, the best fit parameter was a prior of saying "same" 76% of the time. This prior resulted in a hit rate of 81% and a false-alarm rate of 7%, with a correlation between the hit rate of the model and the observers of .68 and between the false alarms of the model and the observers of .35. Once again this model did not significantly differ from the data, $\chi^2(47) = 46.3$, $p = .50$. Thus, in both Experiments 1A and 1B, the model provides a thorough account of human performance.

Conclusion

We examined whether a Bayesian change detection model with more structured memory representations can provide a window into observers' memory capacity. We find that both a summary-based encoding model that encodes the global texture of the display plus specific items and a chunking-based model in which observers first use basic principles of perceptual organization to chunk the display before encoding a fixed number of items provide possible accounts for how observers encode patterned displays. These models can match human levels of accuracy while encoding only three to four items or chunks, and provide a good fit to display-by-display difficulty, accurately predicting which changes observers will find most difficult. Furthermore, the two models seem to capture independent variance, indicating that observers use both kinds of representations when detecting changes in patterned displays. Taken together, the two models account for 81% of the variance in observers' d' across displays and successfully explain the hit and false-alarm rate of individual displays. With both models combined, there is no significant difference between observers' data and the predictions of the model.

By contrast, the simpler formal models of change detection typically used in calculations of visual working memory capacity do not predict any of the reliable differences in difficulty between displays because they treat each item independently. The chunk and summary-based models thus represent a significant step forward for formal models of change detection and visual working memory capacity.

What is the relationship between the perceptual grouping/chunking model and the summary model we have described? Perceptual grouping and chunking are processes by which multiple elements are combined into a single higher order description. For example, a series of 10 evenly spaced dots could be grouped into a single line. In this way, perceptual grouping enables the formation of a compressed representation or chunk of the display (Brady et al., 2011). Critically, perceptual grouping and chunking models like the one we model posit that groups or chunks are the units of representation: If one part of the group or chunk is remembered, all components of the group or chunk can be retrieved. You can never forget the location of the third dot in the grouped line, or the second letter in the verbal chunk "FBI" (Cowan, 2001; Cowan et al. 2004).

By contrast, a model of summary representations assumes a hierarchical view of memory representation. Observers remember information not only about the set as a whole, but also about the items as individual units. For example, Brady and Alvarez (2011) showed that in a working memory task observers remember information not only about the summary of the set of items, but also about individual items, and they combine these two pieces of information when responding. Thus, unlike perceptual grouping and chunking models, a summary-based encoding model represents information at the group level but also maintains separate information about individual items.

In the present experiment, observers' representations seem to reflect in part a chunk-based strategy and in part a summary-based strategy. How can we make sense of this? Both perceptual grouping of items into chunks and the formation of summary representations seem to depend critically on the focus of attention (Brady et al., 2011). For example, perceptual grouping may result when we deploy attention to a single item, and "our attention tends to spread instead across the entire group in which it falls" (Driver & Baylis, 1998, pp. 301–302; see also Driver, Davis, Russell, Turatto, & Freeman, 2001; Scholl, 2001). Thus, perceptual grouping may result from focused attention to an individual item or group (Scholl, 2001).

By contrast, summary representations like the global texture of a display result from diffuse attention (Alvarez, 2011; Alvarez & Oliva, 2008, 2009). When spreading our attentional focus across an entire display, we treat the individual units of the display as a texture and extract a summary representation (Alvarez, 2011; Haberman & Whitney, 2012).

Thus, one possible way of understanding the interaction of these models is that they may represent different focuses of

attention. In other words, the extent to which observers' performance is supported by perceptual grouping versus the extent to which such performance is supported by the representation of summary statistics of the display may depend critically on how diffuse observers' attention to the display is on a particular trial. Focused attention may result in the perception of chunks, whereas a more diffuse attentional strategy may instead lend observers to treat the entire display like a texture and thus encode the overall summary of the display plus the areas of the display that differ from the general texture (Haberman & Whitney, 2012). However, the exact interaction of the models remains to be explored in future work.

Experiments 2A and 2B: Randomly Colored Displays

Using a Bayesian model of change detection together with more structured memory representations allows us to examine observers' working memory capacity in displays with explicit patterns. Can these models also predict which displays are hard or easy on displays without explicit patterns, as in most typical visual working memory experiments (e.g., Luck & Vogel, 1997)? If so, what are the implications for standard *K* values and for simple models of working memory capacity based on these values?

Although most working memory experiments generate displays by randomly choosing colors and placing those color at random spatial locations, this does not mean that there are no regularities present in any given display. In fact, any particular working memory display tends to have significant structure and regularities present even though on average the displays are totally random. Thus, observers may be more likely to get some randomly generated displays correct than others, and this information can be used to examine the representations observers have formed of the displays. This approach is related to the technique of classification images (Ahumada, 1996). In classification images, observers are shown a series of stimuli that differ in the noise that has been added to them, and by averaging together the stimuli that result in correct and incorrect responses, it is possible to determine the representations observers' use to distinguish the stimuli (Eckstein & Ahumada, 2002). However, rather than examine image properties that result in correct or incorrect performance in individual observers, we can instead use our model to examine the full pattern of performance across individual working memory displays.

Variance in observers' encoding or storage in particular displays can have a significant influence on models of memory capacity. For example, Zhang and Luck (2008) used a continuous report task (based on Wilken & Ma, 2004) in which observers are briefly shown a number of colored dots and then asked to report the color of one of these dots by indicating what color it had been on a color wheel. They then modeled observers' responses to partial out observers' errors into two kinds (noisy representations and random guesses), arriving at an estimate of the number of colors observers remember, on average, across all the displays. They found evidence that supported the idea that observers either remember the correct answer or completely forget it, and used this to argue for a model of working memory in which observers can encode at most three items at a time.

Importantly, however, by fitting their model only to the results across all displays rather than taking into account display-by-display variability, they failed to model factors that influence the overall capacity estimate, but average out when looking at many different displays. For example, Bays et al. (2009) showed that many of observers' "random guesses" in this paradigm are actually reports of an incorrect item from the tested display. Reports of the incorrect item tend to average out when looking at all displays, but for each display make a large difference in how many items we should assume observers' were remembering. They used a model in which observers' reports can be not just noisy representations of the correct item or random guesses, but also noisy representations of any of the other items from the display, and showed that such misreports are common. Once these incorrect reports were taken into account, Bays et al. found that the model of Zhang and Luck (2008) no longer provides a good fit to the data (see also Brady, 2011). Although the results of Bays et al. are controversial (see Anderson, Vogel, & Awh, 2011), they nevertheless suggest that display-by-display factors can sometimes significantly influence the degree to which a particular model of working memory is supported, despite a good fit to the average across all displays.

In the current experiment, we sought to examine whether display-by-display variance in encoding particular working memory displays could be formalized with our Bayesian model of observers' memory representations. We applied the same models used in the patterned displays in Experiment 1—the summary-based encoding model and chunk-based model—to displays like those used in typical visual working memory experiments. In Experiment 2A, we generated these displays completely at random, as is typically done in visual working memory experiments. In Experiment 2B, we generated the displays to purposefully contain patterns to provide a better test of the models' representations. We find evidence that observers use such structured representations when encoding these displays, and are able to predict which particular displays observers will find easy or difficult to detect changes in, even in randomly generated displays. This indicates that simple formal models of working memory that encode a small number of independent objects at random do not match the representation observers' use even in relatively simple working memory displays.

Method

Observer. Two hundred observers were recruited and run with Amazon Mechanical Turk. Of the total observers, 100 participated in Experiment 2A and 100 in Experiment 2B. All were from the United States, gave informed consent, and were paid 30 cents for approximately 4 min of their time.

Procedure and stimuli. In Experiment 2A, we randomly generated 24 pairs of displays by selecting eight colors with replacement from a set of seven possible colors (as in Luck & Vogel, 1997) and placing them randomly on a 5×4 invisible grid (see Figure 12). Although it is standard to jitter the items in such displays to avoid colinearities, to facilitate modeling and comparison with the previous experiments we allowed the items to be perfectly aligned. To generate the changed displays, we chose one item at random from each display and changed its color to one of the other six possible colors.

In Experiment 2B, we generated 24 displays to purposefully contain patterns. We generated the displays by creating displays at random and retaining only displays where either the chunk-based

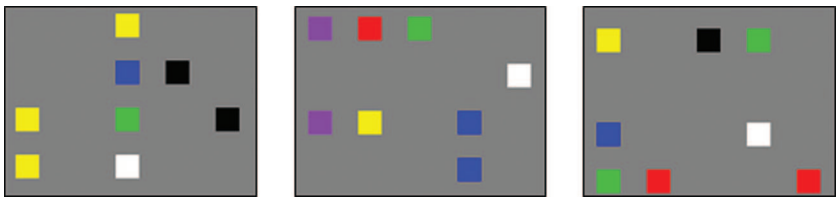

Figure 12. Example displays from Experiment 2. These displays were generated randomly by sampling with replacement from a set of seven colors, as in Luck and Vogel (1997).

model or the summary-based predicted the display would have a d' greater than 2.

The displays each subtended 320 × 240 pixels, with the individual colored squares subtending 30 × 30 pixels. On each trial, the study display appeared on the left, followed by the test display on the right. Observers' monitor size and resolution were not controlled. However, all observers attested to the fact that the entire stimulus presentation box was visible on their monitor.

The method was otherwise the same as Experiment 1.

Results

Experiment 2A. For each display we computed a d', measuring how difficult it was to detect the change in that display (averaged across observers). The mean d' was 1.5 across the displays, corresponding to a K value of 4.38 if we assume all the items are represented independently (Pashler, 1988). The mean hit rate was 60% and the false-alarm rate 11%, reflecting a response bias of $c = 0.48$ (a high likelihood of reporting "same").

However, as in Experiment 1, observers were consistent in which displays they found easy or difficult (see Figure 13). For example, if we compute the average d' for each display using the data from half of our observers and then do the same for the other half of the observers, we find that to a large degree the same displays were difficult for both groups ($r = .68$, averaged over 500 random splits of the observers; reflecting an entire-sample reliability of $r = .83$). By bootstrapping to estimate standard errors on observers' d' for each display, we can visualize this consistency (Figure 13). Some displays, like those on the left of Figure 13, are consistently hard for observers. Others, like those on the right of Figure 13, are consistently easy for observers to detect changes in. Contrary to the assumption of standard working memory models, observers do not appear to treat items independently even on randomly generated displays like those typically used in working memory experiments.

We next fit the summary-based encoding model and the chunk-based model to these data to examine whether these models capture information about observers representations in these displays. To apply the model to the displays from this experiment, we treat the items that are adjacent in the grid as neighbors. Blank spots on the display are ignored, such that the MRF is calculated only over pairs of items (cliques, N_v and N_h) that do not contain a blank location (for details of the inference process, see Appendix B).

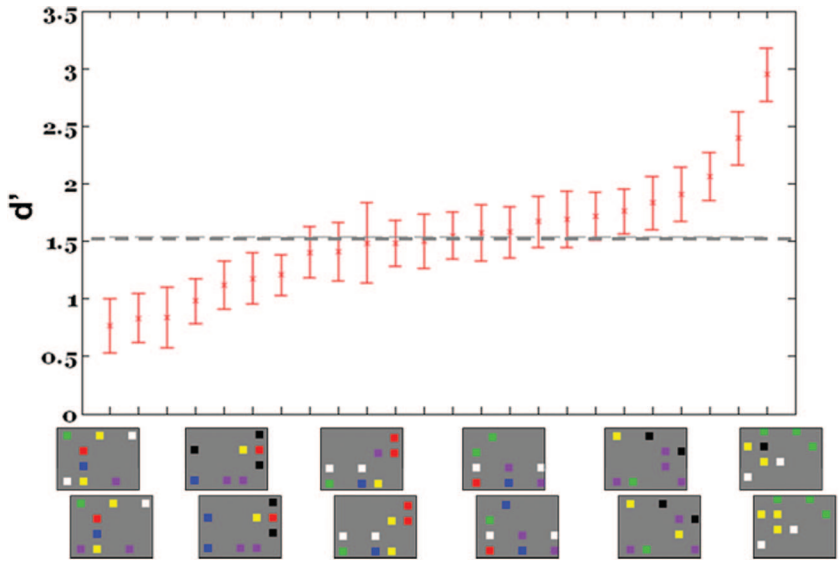

Figure 13. Consistency in which displays are most difficult in Experiment 2. The x-axis contains each of the 24 display pairs, rank ordered by difficulty (lowest d' on the left, highest on the right; for visualization purposes, only a subset of display pairs is shown on the x-axis). The dashed gray line corresponds to the mean d' across all displays. The error bars correspond to across-subject standard error bars. The consistent differences in d' between displays indicate that some displays are more difficult than other displays.

We find that the summary-based model provides a good fit to the data, and in addition correlates with observers' display-by-display difficulty (see Figure 14). The summary-based encoding model equals observers' d' at $K = 4$ ($d' = 1.47$ at $K = 4$, compared with observers' d' of 1.5), and at this K value correlates with display-by-display difficulty well ($r = .63$, $p = .001$). Furthermore, this correlation is not driven by outliers: The Spearman rank-order correlation is also high ($r = .53$, $p = .009$), and if we exclude displays where the model predicts an excessively high d', the correlation remains high despite the decreased range (excluding displays with model $d' > 3$, $r = .61$).

When fitting the hits and false alarms separately (best fit prior: "same" 79% of the time), the model produced a hit rate of 56% and a false-alarm rate of 17%, with a correlation between the hit rate of the model and the observers of .61 and between the false alarms of the model and the observers of .02. As expected, this model did not perfectly capture the data: The model significantly deviated from the number of correct responses of observers on same and different displays, $\chi^2(47) = 144.0$, $p < .01$.

The chunk-based model did not provide as good a fit as the summary model, equaling human performance at $K = 4$ ($d' = 0.88$ at $K = 3$, $d' = 1.32$ at $K = 4$, $d' = 1.81$ at $K = 5$) but only marginally correlating with display-by-display difficulty ($r = .33$ at $K = 3$, $r = .32$ at $K = 4$, $r = .41$ at $K = 5$). When fitting the hits and false alarms separately (prior: "same" 87% of the time), the model had a hit rate of 72% and a false-alarm rate of 9%, with a correlation between the hit rate of the model and the observers of .30 and between the false alarms of the model and the observers of .41. The model also significantly deviated from the number of correct responses of observers, $\chi^2(47) = 150.7$, $p < .01$.

Combining the chunking model with the summary-based model does not significantly improve the fit of the summary-based model to the d' values, with the average of the two models giving a slightly worse fit than the summary-based model alone (with $K = 4$ for both models; correlation with d': $r = .60$). There was some benefit to combining the two models in fitting the full decision process, as fits to the hits and false-alarm rates separately (prior: 87%) resulted in a reduced chi-squared value, $\chi^2(47) = 90.0$. However, this combined model still significantly deviated from observers' ($p < .01$).

Experiment 2B. Generating the displays used in Experiment 2A completely at random means that few displays contained significant enough pattern information to allow for chunking or summary information to play a large role. This allowed us to quantify exactly how well our model representations explained data from truly random displays, as used in most working memory studies (e.g., Luck & Vogel, 1997). However, although we find that even with a sample of just 24 displays some displays are easier than others and this is well explained by our summary-based model, the limited range in observers' d' prevents any strong conclusions about the particular memory representations observers make use of in displays of colored squares (e.g., do observers' representations truly resemble the summary-based model more than the chunk-based model?).

Thus, in Experiment 2B, we generated displays that contain patterns, so that, collapsing across Experiments 2A and 2B, we would have a full range of performances on individual displays. In Experiment 2B, the mean d' was 2.0 across the displays, and observers were once again consistent in which displays were harder and easier (split-half, $r = .64$, suggesting an overall reliability of $r = .76$). The mean hit rate was 65% and the false-alarm rate 6%, reflecting a response bias of $c = 0.53$ (a high likelihood of reporting "same").

Within these new displays, we found that the summary-based model once again provided a strong fit to the d' data ($r = .55$), whereas the chunk-based model provided a considerably worse fit (d': $r = .27$). In addition, when combining the displays from Experiment 2B with the displays from Experiment 2A (see Figure 14), we find that the summary-based model provides a better fit ($r = .64$) than the chunk-based model ($r = .50$).

Experiments 2A and 2B together had a hit rate of 62% and a false-alarm rate of only 8%. The best fit summary-based model to the hit rate and false-alarm rate of the combined experiments required a prior of 84% "same," and resulted in a hit rate and false-alarm rate of 62% and 14%, respectively. This model significantly deviated from the combined data, $\chi^2(95) = 271.7$, $p < .01$.

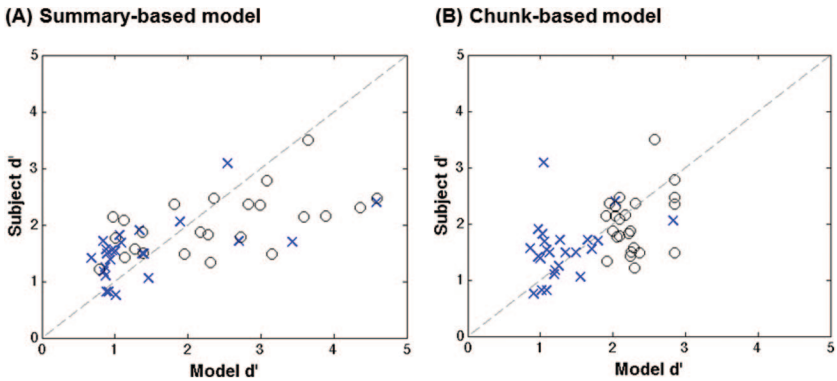

Figure 14. (A) Fit of the summary-based model with $K = 4$. The blue Xs represent the data from Experiment 2A, with randomly generated displays as in typical visual working memory experiments (fit: $r = .63$). The black circles represent data from Experiment 2B, where displays were generated to purposefully contain patterns (fit: $r = .55$). (B) Fit of the chunk-based model with $K = 4$. The blue Xs represent the data from Experiment 2A (fit: $r = .32$). The black circles represent data from Experiment 2B (fit: $r = .27$).

However, the chunk model again provided a considerably worse fit than the summary-based model (prior: 89%; hit rate: 83%; false-alarm rate: 8%), $\chi^2(95) = 412.6$, $p < .01$.

Combining the models resulted in a minor improvement relative to the summary-based model alone in fitting the hits and false alarms (prior: 93%; hit rate: 66%; false-alarm rate: 6%), $\chi^2(95) = 173.8$, but still significantly deviated from observers' data ($p < .01$). In the primary measure of interest, d', combining the two models did not improve the fit of the summary-based model (correlation of d' of combined model: $r = .66$; summary-model alone: $r = .64$).

This suggests that the summary-based model's representation provides a better fit to how observers encode these working memory displays than the chunk-based model does. This could be because the distance between the items prevents low-level perceptual grouping from occurring (Kubovy, Holcombe, & Wagemans, 1998).

Conclusion

Even in standard working memory displays, observers are consistent in which displays they detect changes in and which displays they do not detect changes in. This suggests that the assumption of independence between items that is used in most formal models of working memory does not hold even in these relatively simple displays of segmented shapes. Thus, we need models that take into account basic perceptual grouping and higher order summary representations in order to understand the architecture of visual working memory even when our displays are impoverished relative to real scenes.

Interestingly, even in displays of colored squares—displays that are used in visual working memory in order to minimize the presence of patterns (e.g., Luck & Vogel, 1997)—our summary-based model's representation captures which changes people are likely to detect and which they are likely to miss. By contrast, a model that assumes we encode each item in these displays as a separate unit and choose which to encode at random can predict none of the display-by-display variance. This suggests that observers' representations are more structured than standard models based on independent items would suggest, even in simple working memory displays.

General Discussion

We presented a formal model of change detection that relies upon Bayesian inference to make predictions about visual working memory architecture and capacity. This model allows us to take into account the presence of higher order regularities, while making quantitative predictions about the difficulty of particular working memory displays. In experiments with explicitly patterned displays, we found that both a summary-based representation and a chunk-based representation could successfully explain display-by-display differences in working memory performance. Furthermore, we showed that a model that combines both forms of representation explains a large part of the variance in change detection performance in such patterned displays. In addition, we found that observers were reliable in which displays they found hard or easy even in standard working memory displays composed of colored squares with no explicit spatial patterns, and that our summary-based encoding model could successfully predict this variance.

We thus show that it is necessary to model both more structured memory representations and observers' encoding strategies to successfully understand what information observers represent in visual working memory. We provide a framework for such modeling—Bayesian inference in a model of change detection—and show that it can allow us to understand the format of observers' memory representations. Interestingly, our models converge with the standard visual working memory literature on an estimate of three to four individual objects remembered, even in the patterned displays where simpler formal models massively underestimate observers' performance.

Predicting Display-by-Display Difficulty

Because each item in a typical working memory display is randomly colored and located at a random spatial position, formal models of working memory have tended to treat the displays themselves as interchangeable. Thus, existing models of visual working memory have focused on average memory performance across many different displays. For example, the standard slot model used to calculate K values takes into account only the number of items present and the number of items that change between study and test, ignoring any display-by-display variance in which items are likely to be encoded and how well the items group or how well they can be summarized in ensemble representations. Even modeling efforts that do not focus on slots have tended to examine only performance across all displays (e.g., Wilken & Ma's, 2004, signal detection model where the performance decrement with increasing numbers of items encoded results only from internal noise and noise in the decision process).

However, even when the items themselves are chosen randomly, each display may not itself be "random"; instead, any given display may contain significant structure. Furthermore, by focusing on average performance across displays, existing models have necessarily assumed that each item is treated independently in visual working memory. In the current work, we find that this assumption of independence between items may not hold even in simple displays, but perhaps more importantly, requiring independence between items leaves little room to scale up formal models of working memory to displays where items are clearly not random, as in real-world scenes or even the patterned displays in Experiment 1.

There are two examples of work that fit a formal model that takes into account information about each display in working memory, although neither examines model fits for each particular display as we do in the current work. In the first, Bays et al. (2009) showed that taking into account information about particular displays may be critical to distinguishing between slot models and resource models in continuous report tasks (Bays et al., 2009; Zhang & Luck, 2008). In particular, Bays et al. argued that once trial-by-trial variations are taken into account, the data support a resource model of working memory rather than a slot model of working memory (but see Anderson et al., 2011).

The second example of fitting a working memory model to each display is work done by Brady, Konkle, and Alvarez (2009) on how statistical learning impacts visual working memory. By creating displays where the items were not randomly chosen (partic-

ular colors appear in a pair together more often than chance), they showed that observers can successfully encode more individual colors as they learn regularities in working memory displays. Furthermore, using an information-theoretic model to predict how "compressible" each display was based on how predictable the pairings of colors are, Brady, Konkle, and Alvarez were able to explain how well observers would remember particular displays. For example, displays that have a large number of highly predictable color pairs were remembered better than displays with less predictable pairs.

In the current work, we formalize the encoding of summary statistics and perceptual grouping as possible factors in observers' memory representations. Since the influence of these factors differs on each display, we are able to separately predict the difficulty of each visual working memory display. We thus collected data from large numbers of observers performing the same change detection task on the same displays. This allowed us to examine how well our model predicted performance for each display for the first time. This display-by-display approach could potentially open up a new avenue of research for understanding the representations used in visual working memory, because it allows clear visualizations of what factors influence memory within single displays.

The Use of Ensemble Statistics for Summary-Based Encoding

In our summary-based encoding model, we formalized the idea that observers might store two distinct kinds of memory representations: a set of individual objects plus summary statistics that encode an overall gist of the display. We found evidence that such summary-based encoding can explain human change detection in both patterned displays and simple displays. In addition, we found evidence that a crucial role of summary-based encoding is to guide attention to outlier items.

Our model of summary-based encoding links to both a rich literature on how we encode real-world scenes (e.g., encoding both scene information and specific objects; Hollingworth, 2006; Oliva, 2005) and an emerging literature on the representation of visual information with ensemble statistics (e.g., encoding mean size of a set of items or the distribution of orientations on a display; Alvarez, 2011; Haberman & Whitney, 2012).

When representing a scene, observers encode not only specific objects but also semantic information about a scene's category as well as its affordances and other global scene properties (e.g., Greene & Oliva, 2009a, 2009b, 2010). There is also existing evidence that the representation of such scene and ensemble information influences our encoding of specific objects. For example, observers are better at remembering the spatial position of an object when tested in the context of a scene (Hollingworth, 2007; Mandler & Johnson, 1976), and this effect is stronger when the scene information is meaningful and coherent (Mandler & Johnson, 1976; Mandler & Parker, 1976). In addition, gist representations based on semantic information seem to drive the encoding of outlier objects. Thus, objects are more likely to be both fixated and encoded into memory if they are semantically inconsistent with the background scene (e.g., Friedman, 1979; Hollingworth & Henderson, 2000, 2003). Visual information from scenes also influences our encoding of objects. Thus, observers encoding real-world scenes preferentially encodes not only semantic outliers but also visual outliers ("salient" objects; Fine & Minnery, 2009; Wright, 2005; but see Stirk & Underwood, 2007). In addition, when computing ensemble visual representations in simpler displays, observers discount outlier objects from these representations (Haberman & Whitney, 2010) and combine their representations of the ensemble statistics with their representation of individual items (Brady & Alvarez, 2011). However, it remains unclear whether ensemble statistics and texture representations take up space in memory that would otherwise be used to represent more information about individual items (as argued, for example, by Feigenson, 2008, and Halberda, Sires, & Feigenson, 2006), or whether ensemble representations are stored entirely independently of representations of individual items perhaps analogous to the separable representations of real-world objects and real-world scenes (e.g., Greene & Oliva, 2009b).

Taken together, this suggests that observers' representations of both real-world scenes and simpler displays consist of not only information about particular objects but also scene-based information and ensemble visual information. Furthermore, this summary information is used to influence the choice of particular objects to encode and ultimately influences the representation of those objects.

In the current work, we formalized a simplified version of such a summary-based encoding model. Rather than represent semantic information, we use displays that lack semantic information and used a summary representation based on MRFs (Geman & Geman, 1984). This summary representation represents only spatial continuity properties of the display (e.g., the similarity between items that are horizontal and vertical neighbors), providing a simple model of visual texture. Interestingly, however, a very similar representation seems to capture observers' impression of the subjective randomness of an image patch (Schreiber & Griffiths, 2007), a concept similar to Garner's (1974) notion of "pattern goodness." Pattern goodness is an idea that has been difficult to formalize but qualitatively seems to capture which images are hard and easy to remember (Garner, 1974).

Nevertheless, our summary representation is too impoverished to be a fully accurate model of the summaries encoded in human memory, even for such simple displays. For example, if semantic information like letters or shapes appeared in the dot patterns in our displays, observers would likely recall those patterns well by summarizing them with a gist-like representation. Our model cannot capture such representations. Additional visual summary information is also likely present but not being modeled. For example, if we changed the shape of one of the items in Experiment 1 from a red circle to a red square, observers would almost certainly notice despite the large number of individual items on the display, as such information is well captured by typical texture/ensemble representations (see, e.g., Brady et al., 2011). However, despite the relative simplicity of the formalized summary representation, our model seems to capture a large amount of variance in how well observers remember not only patterned displays but also simple visual working memory displays.

Ultimately, observers' entire store of knowledge can be brought to bear on their memory representations, such that observers in the United States will easily remember the letter string *FBICIAIRS*, but citizens of other countries would not. Thus, our model of the visual structure that observers encode is necessarily too simplified to capture real-world memory performance. Nevertheless, it pro-

vides a first step that allows us to capture significantly more structural complexity than existing models of visual working memory that treat objects as entirely independent units (e.g., Bays & Husain, 2008; Zhang & Luck, 2008). In addition, it provides a framework for examining the role of structure and summary representation in visual working memory that can be expanded upon in future work.

Chunking

In our chunk-based encoding model, we suggested that observers might make use of the Gestalt principle of similarity to form perceptual units out of the individual items in our displays and encode these units into memory as chunks. We found evidence that such chunk-based encoding can explain part of human change detection in patterned displays.

This idea that memory might encode chunks rather than individual objects relates to two existing literatures. One is the literature on semantic, knowledge-based chunk formation. For example, a large amount of work has been done to understand how chunks form based on knowledge, both behaviorally (e.g., Brady, Konkle, & Alvarez, 2009; Chase & Simon, 1973; Cowan et al. 2004; Gobet et al., 2001) and with computational models of what it means to form such chunks; how all-or-nothing chunk formation is; and what learning processes observers undergo (e.g., Brady, Konkle, & Alvarez, 2009; Gobet et al., 2001). The other literature on chunk formation is based on more low-level visual properties, as examined under the headings of perceptual grouping and pattern goodness (e.g., Garner, 1974; Koffka, 1935; Wertheimer, 1938). In the current work, we use nonsemantic stimuli and do not repeat stimuli to allow for learning, and thus it is likely we are tapping a form of chunk formation that is based on grouping properties of low-level vision rather than based on high-level knowledge.

Some previous work has focused on how to formalize this kind of perceptual grouping (Kubovy & van den Berg, 2008; Rosenholtz, Twarog, Schinkel-Bielefeld, & Wattenberg, 2009). For example, Kubovy and van den Berg (2008) have proposed a probabilistic model of perceptual grouping with additive effects of item similarity and proximity on the likelihood of two objects being seen as a group. In the current experiments, our items differ only in color, and thus we make use of a straightforward model of grouping items into chunks, where items that are adjacent and same-colored are likely but not guaranteed to be grouped into a single unit. This grouping model is similar in spirit to that of Kubovy and van den Berg, and in our displays seems to explain a significant portion of the variance in observers' memory performance. This provides some evidence that perceptual grouping may occur before observers encode items into memory, allowing observers to encode perceptual chunks rather than individual items per se.

Similar models of perceptual grouping have been proposed to explain why observers are better than expected at empty-cell localization tasks using patterned stimuli much like ours (Hollingworth et al., 2005) and why some displays are remembered more easily than others in same and different tasks (Halberda et al., 2012; Howe & Jung, 1986). However, this previous work did not attempt to formalize of the model of perceptual grouping. This is important because in the current experiments we find that summary-based encoding provides another possible explanation for the benefits observed in patterned displays, and in fact may provide a more general solution, since it helps explain performance in simpler displays better than perceptual grouping. Thus, we believe it is an important open question the extent to which summary-based encoding (as in texture representations) rather than perceptual grouping could explain improved performance for patterned displays in previous experiments (Halberda et al., 2012; Hollingworth et al., 2005; Howe & Jung, 1986).

Fidelity in Visual Working Memory

In line with the previous literature on working memory, the current modeling effort largely treats working memory capacity as a fixed resource in which up to K items may be encoded with little noise. Although expanding on what counts as an "item" (in the chunk-based model) or suggesting a hierarchical encoding strategy (in the summary-based model), nevertheless we do not investigate in detail the fidelity stored in the representations or the extent to which encoding is all-or-none (e.g., slot-like) versus a more continuous resource.

There are several important caveats to the simplistic idea of all-or-none slots that we use throughout the current modeling effort. The first is that for complex objects, observers are able to represent objects with greater detail when they are encoding only a single object or only a few objects than when they are encoding many such objects (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007). In fact, the newest evidence suggests this is true even of memory for color (Zhang & Luck, 2008). For example, Zhang and Luck (2008) find that observers have more noise in their color reports when remembering three colors than when remembering only a single color. It has been proposed that this is due to either a continuous resource constraint with an upper bound on the number of objects it may be split between (Alvarez & Cavanagh, 2004), a continuous resource with no upper bound (Bays et al., 2009; Bays & Husain, 2008), a continuous resource that must be divided up between a fixed number of slots (Awh et al. 2007), or because observers store multiple copies of an object in each of their slots when there are fewer than the maximum number of objects (Zhang & Luck, 2008). In any case, our simplistic model in which several items are perfectly encoded would need to be relaxed to incorporate these data.

Furthermore, in real-world displays that contain many real objects in a scene, observers continually encode more objects from the display the more time they are given (Hollingworth, 2004; Melcher, 2001, 2006). In fact, even on displays with objects that are not in a coherent scene, if those objects are semantically rich real-world objects, observers remember more detailed representations for a larger number of objects as they are given more time to encode the objects (Brady, Konkle, Oliva, & Alvarez, 2009; Melcher, 2001).

Despite these complications, in the current modeling we focus on expanding a basic all-or-none slot model to the case of dealing with higher order regularities and perceptual organization. We use such a model as our basic architecture of working memory because of its inherent simplicity and because it provides a reasonable fit to the kind of change detection task where the items to be remembered are simple and the changes made in the change detection task are large, as in the current studies (e.g., categorical changes in color; Luck & Vogel, 1997). Future work will be required to

explore how perceptual grouping and summary-based encoding interact with memory fidelity.

Conclusion

Memory representations of real-world scenes are complex and structured: Observers encode both scene-based semantic and visual information as well as specific objects, and the objects they encode are chosen based on the scene information. By contrast, formal models of working memory have typically dealt with only simple memory representations that assume items are treated independently and no summary information is encoded.

In the current work, we presented a formal model of change detection that uses Bayesian inference to make predictions about visual working memory architecture and capacity. This model allowed us to take into account the presence of summary information and perceptual organization, while making quantitative predictions about the difficulty of particular working memory displays. We found evidence that observers make use of more structured memory representations not only in displays that explicitly contain patterns, but also in randomly generated displays typically used in working memory experiments. Furthermore, we provided a framework to model these structured representations—Bayesian inference in a model of change detection—and showed that it can allow us to understand how observers make use of both summary information and perceptual grouping.

By treating change detection as inference in a generative model, we make contact with the rich literature on a Bayesian view of low-level vision (Knill & Richards, 1996; Yuille & Kersten, 2006) and higher level cognition (e.g., Griffiths & Tenenbaum, 2006; Tenenbaum et al., 2006). Furthermore, by using probabilistic models, we obtain the ability to use more complex and structured knowledge in our memory encoding model, rather than treating each item as an independent unit (e.g., Kemp & Tenenbaum, 2008; Tenenbaum et al., 2006). Our model is thus extensible in ways that show promise for building a more complete model of visual working memory: Within the same Bayesian framework, it is possible to integrate existing models of low-level visual factors with existing models of higher level conceptual information (e.g., Kemp & Tenenbaum, 2008), both of which will be necessary to ultimately predict performance in working memory tasks with real-world scenes.

References

Ahumada, A. J., Jr. (1996). Perceptual classification images from Vernier acuity masked by noise [Abstract]. *Perception, 25,* 18.

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106,* 20–29. doi:10.1016/j.jecp.2009.11.003

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15,* 122–131. doi:10.1016/j.tics.2011.01.003

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15,* 106–111. doi:10.1111/j.0963-7214.2004.01502006.x

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19,* 392–398. doi:10.1111/j.1467-9280.2008.02098.x

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics: Efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 7345–7350. doi:10.1073/pnas.0808981106

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *Journal of Neuroscience, 31,* 1128–1138. doi:10.1523/JNEUROSCI.4125-10.2011

Ariely, D. (2001), Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162. doi:10.1111/1467-9280.00327

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items, regardless of complexity. *Psychological Science, 18,* 622–628. doi:10.1111/j.1467-9280.2007.01949.x

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423. doi:10.1016/S1364-6613(00)01538-2

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 7. doi:10.1167/9.10.7

Bays, P. M., & Husain, M. (2008, August 8). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician, 24,* 179–195.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika, 64,* 616–618. doi:10.1093/biomet/64.3.616

Brady, T. F. (2011, July). *Trial-by-trial variance in visual working memory capacity estimates as a window into the architecture of working memory*. Poster presented at the 33rd Annual Meeting of the Cognitive Sciences Society, Boston, MA.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22,* 384–392. doi:10.1177/0956797610397956

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual short-term memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138,* 487–502. doi:10.1037/a0016797

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and towards structured representations. *Journal of Vision, 11*(5), 4. doi:10.1167/11.5.4

Brady, T. F., Konkle, T., Oliva, A., & Alvarez, G. A. (2009). Detecting changes in real-world objects: The relationship between visual long-term memory and change blindness. *Communicative & Integrative Biology, 2*(1), 1–3. doi:10.4161/cib.2.1.7297

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13,* 207–230. doi:10.1016/0010-0285(81)90008-6

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81. doi:10.1016/0010-0285(73)90004-2

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. doi:10.1017/S0140525X01003922

Cowan, N. (2005). *Working memory capacity*. Hove, England: Psychology Press. doi:10.4324/9780203342398

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science, 15,* 634–640. doi:10.1111/j.0956-7976.2004.00732.x

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466. doi:10.1016/S0022-5371(80)90312-6

Driver, J., & Baylis, G. C. (1998). Attention and visual object segmentation. In R. Parasuraman (Ed.), *The attentive brain* (pp. 299–325). Cambridge, MA: MIT Press.

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention, and phenomenal visual objects. *Cognition, 80,* 61–95. doi:10.1016/S0010-0277(00)00151-7

Eckstein, M. P., & Ahumada, A. J., Jr. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision, 2*(1), 1. doi:10.1167/2.1.i

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science, 208,* 1181–1182. doi:10.1126/science.7375930

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition, 107,* 1–18. doi:10.1016/j.cognition.2007.07.006

Fine, M. S., & Minnery, B. S. (2009). Visual salience affects performance in a working memory task. *Journal of Neuroscience, 29,* 8016–8021. doi:10.1523/JNEUROSCI.5503-08.2009

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108,* 316–355. doi:10.1037/0096-3445.108.3.316

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17,* 673–679. doi:10.3758/17.5.673

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741. doi:10.1109/TPAMI.1984.4767596

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5,* 236–243. doi:10.1016/S1364-6613(00)01662-4

Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20,* 464–472. doi:10.1111/j.1467-9280.2009.02316.x

Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58,* 137–176. doi:10.1016/j.cogpsych.2008.06.001

Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene property. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1430–1442. doi:10.1037/a0019058

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17,* 767–773. doi:10.1111/j.1467-9280.2006.01780.x

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17,* R751–R753. doi:10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72,* 1825–1838. doi:10.3758/APP.72.7.1825

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *From perception to consciousness: Searching with Anne Treisman* (pp. 339–349). New York, NY: Oxford University Press.

Halberda, J., Simons, D. J., & Whetherhold, J. (2012). *Superfamiliarity affects perceptual grouping but not the capacity of visual working memory*. Manuscript submitted for publication.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science, 17,* 572–576. doi:10.1111/j.1467-9280.2006.01746.x

Hemmer, P., & Steyvers, M. (2009). Integrating episodic and semantic information in memory for natural scenes. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 1557–1562). Austin, TX: Cognitive Science Society.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short- and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 519–537. doi:10.1037/0096-1523.30.3.519

Hollingworth, A. (2006). Scene and position specificity in visual memory for objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 58–69. doi:10.1037/0278-7393.32.1.58

Hollingworth, A. (2007). Object–position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 31–47. doi:10.1037/0096-1523.33.1.31

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition, 7,* 213–235. doi:10.1080/135062800394775

Hollingworth, A., & Henderson, J. M. (2003). Testing a conceptual locus for the inconsistent object change detection advantage in real-world scenes. *Memory & Cognition, 31,* 930–940. doi:10.3758/BF03196446

Hollingworth, A., Hyun, J., & Zhang, W. (2005). The role of visual short-term memory in empty cell localization. *Perception & Psychophysics, 67,* 1332–1343. doi:10.3758/BF03193638

Howe, E., & Jung, K. (1986). Immediate memory span for two-dimensional spatial arrays: Effects of pattern symmetry and goodness. *Acta Psychologica, 61,* 37–51. doi:10.1016/0001-6918(86)90020-X

Kaplan, R. M., & Saccuzzo, D. P. (2008). Psychological testing: Principles, applications, and issues. Belmont, CA: Wadsworth.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America, 10,* 10687–10692. doi:10.1073/pnas.0802631105

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511984037

Koffka, K. (1935). *Principles of gestalt psychology*. New York, NY: Harcourt, Brace.

Kubovy, M., Holcombe, A. O., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology, 35,* 71–98. doi:10.1006/cogp.1997.0673

Kubovy, M., & van den Berg, M. (2008). The whole is equal to the sum of its parts: A probabilistic model of grouping by proximity and similarity in regular patterns. *Psychological Review, 115,* 131–154. doi:10.1037/0033-295X.115.1.131

Lampinen, J. M., Copeland, S., & Neuschatz, J. S. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1211–1222. doi:10.1037/0278-7393.27.5.1211

Li, S. Z. (1995). *Markov random field modeling in computer vision*. Secaucus, NJ: Springer.

Luck, S. J. (2008). Visual short-term memory. In S. J. Luck & A. Hollingworth (Eds.), *Visual memory* (pp. 43–85). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195305487.003.0003

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. doi:10.1038/36846

Mandler, J. M., & Johnson, N. S. (1976). Some of the thousand words a picture is worth. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 529–540. doi:10.1037/0278-7393.2.5.529

Mandler, J. M., & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 38–48. doi:10.1037/0278-7393.2.1.38

Melcher, D. (2001). Persistence of visual memory for scenes. *Nature, 412,* 401. doi:10.1038/35086646

Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision, 6*(1), 2. doi:10.1167/6.1.2

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. doi:10.1037/h0043158

Miller, M. B., & Gazzaniga, M. S. (1998). Creating false memories for visual scenes. *Neuropsychologia, 36,* 513–520. doi:10.1016/S0028-3932(97)00148-6

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139174909

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70,* 772–788. doi:10.3758/PP.70.5.772

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Olson, I. R., & Jiang, Y. (2004). Visual short-term memory is not improved by training. *Memory & Cognition, 32,* 1326–1332. doi:10.3758/BF03206323

Olson, I. R., Jiang, Y., & Moore, K. S. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 889–900. doi:10.1037/0096-1523.31.5.889

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 2745–2750. doi:10.1073/pnas.0708424105

Pashler, H. (1988). Familiarity and the detection of change in visual displays. *Perception & Psychophysics, 44,* 369–378. doi:10.3758/BF03210419

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics, 16,* 283–290. doi:10.3758/BF03203943

Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., & Wattenberg, M. (2009). *An intuitive model of perceptual grouping for HCI design*. Proceedings of the 27th International Conference on Human Factors in Computing Systems (pp. 1331–1340). New York, NY: ACM Press.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80,* 1–46. doi:10.1016/S0010-0277(00)00152-9

Schreiber, E., & Griffiths, T. L. (2007). Subjective randomness and natural scene statistics. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1449–1454). Mahwah, NJ: Erlbaum.

Sebrechts, M. M., & Garner, W. R. (1981). Stimulus-specific processing consequences of pattern goodness. *Memory & Cognition, 9,* 41–49. doi:10.3758/BF03196950

Stirk, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision, 7*(10), 3. doi:10.1167/7.10.3

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10,* 309–318. doi:10.1016/j.tics.2006.05.009

Victor, J. D., & Conte, M. M. (2004). Visual working memory for image statistics. *Vision Research, 44,* 541–556. doi:10.1016/j.visres.2003.11.001

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 92–114. doi:10.1037/0096-1523.27.1.92

Wertheimer, M. (1938). *Laws of organization in perceptual forms*. London, England: Harcourt Brace Jovanovich.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 11. doi:10.1167/4.12.11

Woodman, G. F. Vecera, S. P., & Luck, S. J. (2003). Perceptual organization influences visual working memory. *Psychonomic Bulletin & Review, 10,* 80–87. doi:10.3758/BF03196470

Wright, M. J. (2005). Saliency predicts change detection in pictures of natural scenes. *Spatial Vision, 18,* 413–430. doi:10.1163/1568568054389633

Xu, Y. (2006). Understanding the object benefit in visual short-term memory: The roles of feature proximity and connectedness. *Perception & Psychophysics, 68,* 815–828. doi:10.3758/BF03193704

Xu, Y., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 18766–18771. doi:10.1073/pnas.0705618104

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10,* 301–308. doi:10.1016/j.tics.2006.05.002

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235. doi:10.1038/nature06860

(*Appendies follow*)

Appendix A

Patterned Display Generation

To generate the patterned displays used in Experiments 1A and 1B, we sampled a set of 16 displays from a Markov random field (MRF) smoothness model like those used in our summary-based encoding model and our chunking model. Using an MRF with separate parameters for horizontal and vertical smoothness, we used Gibbs sampling to generate a set of four displays from each of four possible parameter settings. These parameters encompassed a wide range of possible patterns, with horizontal and vertical smoothness set to all combinations of ± 1 [(1, 1), (−1, −1), (1, −1), (−1, 1)]. This gave us 16 displays with noticeable spatial patterns. In addition, we generated eight displays by randomly and independently choosing each dot's color (50%/50%). In Experiment 1A, these 24 displays consisted of red and blue dots. In Experiment 1B they were the same displays, but composed of black and white squares instead. For all of the displays, we chose the change to make in the display by picking a random item and flipping it to the opposite color.

Appendix B

Models as Applied to Experiment 2

To apply the models to the displays from Experiment 2, we use the same model and treat the items that are adjacent in the grid as neighbors. Blank spots on the display are ignored, such that the Markov random field (MRF) is calculated only over pairs of items (cliques, N_v and N_h) that do not contain a blank location. In addition, we expanded the range of parameter values we considered for G to be -5 to 5, rather than -1.5 to 1.5, since the smaller numbers of items in these displays result in more extreme values for the summary parameters.

To do inference in the summary-based encoding model, we can no longer use exact inference, since calculating the partition function $Z(G)$ for these displays is computationally implausible. Instead, to calculate the likelihood of a given display under a particular summary representation, we use the pseudolikelihood, which is the product, for all the items, of the conditional probability of that item given its neighbors (Besag, 1975, 1977; Li, 1995). Thus, $P(D^1|G)$ is calculated as

$$P\left(D^1|\ G\right) = \prod_i \frac{\exp\left(-En\left(D_i^1|G\right)\right)}{\exp\left(-En_i\left(0|G\right)\right)\ +\ \exp\left(-En_i\left(1|G\right)\right)} \tag{B1}$$

$$En_i(D^1|G) = G_v \sum_{j \in N_v(i)} \Psi(D_i^1, D_j^1) + G_h \sum_{j \in N_h(i)} \Psi(D_i^1, D_j^1). \tag{B2}$$

Such an estimate of the likelihood is computationally straightforward, and in MRFs has been shown to be a reasonable approximation to the true underlying likelihood function (Besag, 1977). We can calculate how good an approximation it is for our particular change detection model by examining how closely predictions using the pseudolikelihood approximate the exact likelihood computations in Experiment 1. In that model (with $K = 4$), the change detection estimates (how likely each test display is to be the same as the study display) correlate .98 between the model that uses exact inference and the model that relies on the pseudolikelihood to estimate the likelihood. This suggests that the pseudolikelihood provides a close approximation of the true likelihood in our displays.

In addition, in Experiment 2 there were only eight items present on the displays. Thus, it was computationally feasible to consider all sets of K items for inclusion in S, the set of remembered items, rather than consider only the K most outlier items, so our model considered all items.

Received November 6, 2011
Revision received September 10, 2012
Accepted September 12, 2012 ■